Proceedings of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 499–506
DOI: 10.15439/2025F3908
ISSN 2300-5963 ACSIS, Vol. 43

ELEVATE-AI: Evaluation of Learning Environments Via Assessment Tools Enhanced by AI

Linda Burchiellaro
0009-0008-1098-1398
University of Modena and Reggio Emilia
Department of Physics, Informatics and Mathematics
Via Campi 213, 41125 Modena, Italy
Email: 273342@studenti.unimore.it

Francesco Faenza, Claudia Canali
0000-0002-7258-7192, 0000-0001-8448-7693
University of Modena and Reggio Emilia,
Department of Engineering "Enzo Ferrari",
Via P. Vivarelli 10, 41125 Modena, Italy
Email: {francesco.faenza, claudia.canali}@unimore.it

***Abstract*—The number of STEM camps and extracurricular initiatives has risen considerably in recent years, driven by the increasing emphasis on the workforce shortage in STEM fields. However, despite their growth, these programs often suffer from a lack of structured evaluation practices, a well-known issue that hinders a comprehensive understanding of their effectiveness. This work focuses specifically on outreach initiatives and proposes ELEVATE-AI, a standardized evaluation platform that includes data-cleaning procedures, Exploratory Factor Analysis (EFA), and regression analysis to measure impacts effectively. Furthermore, we discuss the potential integration of AI-based tools to support non-experts in interpreting the results of the proposed analysis flow. The platform aims to lower technical barriers, promote systematic assessment, and encourage the widespread adoption of data-driven practices in evaluating CS and STEM outreach activities. To facilitate adoption and reproducibility, the platform will be made available as an open-source tool.**

I. Introduction

IN RECENT years, particularly since the mid-2010s, there has been growing attention at all levels of society toward increasing the number of graduates in STEM fields, particularly in computer science (CS). Public institutions such as the European Union have highlighted the urgent need for more professionals in digital and technological sectors, citing projected shortages of ICT specialists and the importance of closing the gender gap in these domains [1]. Similarly, the OECD and national education authorities have acknowledged the underrepresentation of women in STEM careers and the need for inclusive educational strategies [2].

Among the most prominent approaches to address this challenge are extracurricular and outreach initiatives, activities not formally associated with standard curricula but aimed at introducing students to STEM topics through informal, hands-on learning. Particularly in the case of girls, experiential and mentorship-based learning environments have shown effectiveness in fostering self-efficacy and interest in STEM subjects [3]. Broader studies confirm that these benefits extend to students of all genders when appropriate pedagogical strategies are used [4].

Despite their increasing prevalence, many outreach initiatives lack systematic evaluation, limiting their scalability and long-term impact [4], [5]. Programs often shift focus to trending topics, such as artificial intelligence, without assessing the impact of each iteration, resulting in design decisions based more on intuition than evidence [6]. One of the reasons for this absence of evaluation is the informal nature of these activities: unlike curricular settings, where learning outcomes can be tested directly, outreach initiatives tend to rely on self-perception measures rooted in Bandura's theory of self-efficacy [7].

Nevertheless, recent research demonstrates that robust evaluation is feasible even in informal contexts [5]. In this paper, we propose an open-source, self-hostable platform that provides a standardized, literature-informed workflow for evaluating extracurricular STEM initiatives. The platform, intended to be accessible to both researchers and educators, automates the main key phases of the evaluation process: survey design and deployment, data collection and validation, Exploratory Factor Analysis (EFA), and regression-based impact evaluation. The platform is designed to promote broader adoption of data-driven evaluation practices and to support community-driven contributions to its development. A valuable additional contribution of this paper is that the platform will be publicly released as an open-source resource, to maximize its impact and facilitate rapid adoption by researchers and practitioners.

The remaining part of the paper is structured as follows. Sec. II discusses the motivational background of the proposal. Sec. III describes the methodology including the three main phases of survey design, data collection and validation, and data analysis and implementation. Sec. IV presents the platform architecture and Sec. V illustrates the user interaction workflow. Finally, Sec. VI provides some concluding remarks and depicts the future research directions.

II. Motivation and Background

The evaluation of STEM and CS outreach activities, particularly in extracurricular settings, presents methodological challenges not typically encountered in formal education. In such informal contexts, learners often engage voluntarily, and activities are less structured, with outcomes being more affective or motivational than knowledge-based.

In contrast, the curricular context benefits from the increasing automation of evaluation processes, which often extends to automatic analysis and feedback generation. Several examples

illustrate this trend, particularly in STEM education. Cipriano et al. [8] present Drop Project, a platform designed to automate test result analysis and feedback for programming assignments. Similarly, Web-CAT[9] offers an automated grading system that compiles, tests, and evaluates student code submissions, providing immediate feedback. Majerník [10] introduces an e-Assessment Management System aimed at comprehensively evaluating medical students' knowledge, while, more generally, Stanescu et al. [11] propose a solution for automatic assessment of narrative answers, further demonstrating how structured digital platforms can support rigorous assessment practices even in highly specialized domains. These platforms are just some examples of how it is not only feasible but also widely applied to automate evaluation systems when grading and formal tests are involved.

Another contribution to formal testing automation in STEM is *LASSO* [12], a tool designed to support instructors in the assessments of their courses. LASSO's primary purpose is to promote the implementation of research-based teaching practices; in fact, it guides the teacher through a rigorous process of pre-test and post-test analysis. The platform consists of a generalized version of the previously cited Web-CAT and Drop Project, providing a platform to facilitate rigorous formal evaluation in all STEM fields.

More broadly, in formal education, structured assessments such as knowledge tests allow for systematic and replicable evaluation. Notably, Hattie's Visible Learning synthesis [13] aggregated thousands of studies to quantify the effect size of nearly every measurable educational intervention. This kind of quantification is feasible mainly because of the controlled nature of formal instruction and the availability of consistent outcome measures.

Nevertheless, the principle behind Visible Learning, making instructional choices based on solid evidence rather than trends or intuition, is just as applicable in informal learning. Even in the absence of formal testing, it is possible to build robust evaluation models using survey-based self-perception data as long as those instruments are rigorously validated and analyzed. By supporting outreach practitioners in collecting and interpreting valid data, we argue that evidence-informed design is not only possible but necessary in extracurricular education.

While standardized questionnaires and evaluation frameworks developed for formal environments cannot be directly transplanted into outreach contexts without adaptation, several initiatives have attempted to improve evaluation in informal contexts, such as the CISE REU toolkit [5] and the survey repository curated by Decker and McGill [14]. Although these resources provide a helpful starting point, they require contextual refinement to be used effectively in pre-college or extracurricular settings.

Knekta et al. [15] indeed underline the issue of contextualization by arguing that validity is not a property of an instrument itself but of the inferences drawn from its use in a specific context. Their study, titled "One Size Doesn't Fit All," emphasizes the importance of context-specific validation using techniques such as Exploratory Factor Analysis (EFA). In their view, even well-designed instruments must be empirically validated with data from the actual population and learning setting in which they are used. This is particularly relevant for outreach initiatives targeting diverse, often underrepresented groups in non-standard learning environments.

In line with the need for contextualization, the *Towards s'more Connected Coding Camps* study [16] outlines the vision of the European OSCAR initiative, which seeks to integrate informal STEM learning experiences, such as coding camps, into students' broader educational and professional trajectories. The initiative focuses on the current fragmentation of the panorama of extracurricular activities. The proposal is to give continuity and align these activities with formal education, developing a coherent learning pathway in which informal learning is no longer treated as an isolated experience but contributes meaningfully to long-term educational outcomes.

TheFragebogen [17], a web-based, open-source questionnaire framework initially developed for Quality of Experience (QoE) research, provides a compelling model for digital instrument delivery. Its architecture is guided by three design principles: responsiveness to research needs (e.g., multimedia support, scalability), extensibility and flexibility (including serverless deployment), and a focus on robustness and reproducibility (e.g., long-term data archiving). While it originates outside the domain of education, TheFragebogen exemplifies how modular, browser-based systems can meet the needs of research-grade data collection. These principles are particularly relevant to the design of evaluation tools for outreach education, where technological robustness must be paired with usability and adaptability.

The evaluation platform proposed in this paper results from several years of iterative refinement grounded in field experience. Specifically, its development began in 2016 within the context of *Digital Girls* [18], [19], a long-running Italian initiative launched in 2013 to engage female students aged 16–18 and introduce them to computer science (CS). The project is recognized in the Observatory for Public Sector Innovation Case Study Library [1] and was cited in the European Commission's 2021 *She Figures* report[2].

Digital Girls evolved from a single-university summer camp into a regionally coordinated program supported by local universities, schools, and stakeholders. It combines site visits, guest lectures, and project-based learning in a female-only setting, addressing confidence gaps and gender stereotypes before university entry. The program spans two weeks and offers approximately 50 hours of immersive activities in university settings.

The iterative enhancement of this program and its evaluation methodology followed a design-based research model [20]. The research adopted an iterative and literature-informed de-

[1] https://oecd-opsi.org/innovations/digital-girls-emilia-romagna/

[2] https://research-and-innovation.ec.europa.eu/knowledge-publications-tools-and-data/publications/all-publications/she-figures-2021_en

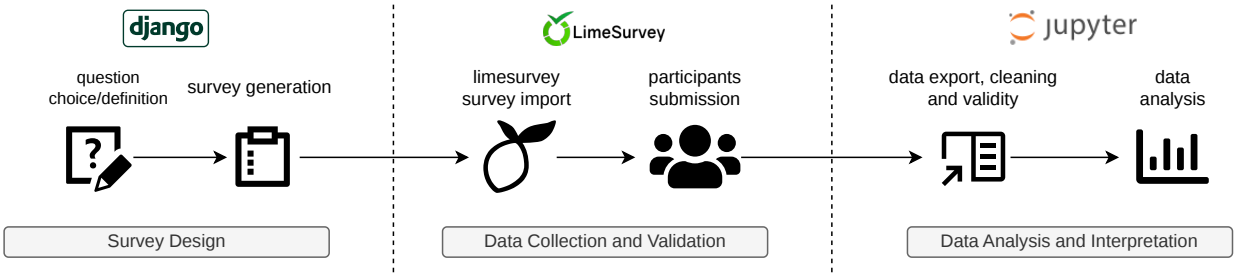

Fig. 1. Overall Platform Flow and Underlying Technologies

sign approach [21] involving cycles of instrument refinement, deployment, and validation. Survey instruments were informed by existing literature references but were adapted through feedback gathered via focus groups with past participants [22].

The resulting methodology includes two core survey phases, pre-camp, and post-camp, designed to capture changes in participants' perceptions, intentions, and confidence regarding CS. Several questions are repeated across both surveys to enable a longitudinal comparison of key indicators. Additional dimensions include background data (school type, video game experience, coding exposure), social influence (parental background), and affective constructs (motivation, stereotype perception), which align with Bandura's theory of self-efficacy [7] and Ajzen's theory of planned behavior [23].

Given the informal learning setting commonly associated with extracurricular initiatives, the evaluation strategy commonly deliberately excludes formal knowledge testing. Instead, it relies on participants' self-assessment to understand whether and how their attitudes have shifted as a result of the experience. A typical example is the measurement of self-efficacy in relation to the main topic of the camp, such as programming or algorithmic thinking. These constructs are typically evaluated through Likert-scale items, which provide a straightforward basis for statistical analysis. In some cases, free-text questions are added as they can provide meaningful additions or subtractions to the Likert items [24].

One of the main goals of STEM extracurricular initiatives is to engage participants and encourage them to consider future academic or professional paths in STEM fields. These programs are designed with the hope that, after the experience, participants will be more inclined to include STEM-related options among their future choices. In this context, Ajzen's theory of planned behavior [23] is particularly relevant, as it provides a framework for analyzing changes in intention. For instance, one of the main objectives of the *Digital Girls* initiative is to investigate whether the camp experience influences participants to reconsider their university plans, specifically, whether they become more likely to pursue a degree in computer science. To this end, the pre- and post-camp surveys are structured to detect shifts in intention, which can be analyzed as dependent variables in regression models to assess the impact of specific features of the camp.

III. Methodology

The multi-year experience derived from the organization of Digital Girls and other initiatives related to STEM disciplines led to the formalization of the platform presented in this paper. The platform is open-source, self-hostable, and modular, being written in Django, a Python framework for web platforms. It is structured, as shown in Figure 1, to assist both researchers and educators in evaluating CS outreach activities through three main phases: Survey Design, Data Collection and Validation, and Data Analysis and Interpretation.

A. Survey Design

One of the most critical components of any evaluation system is the design of the survey instrument, which essentially involves selecting appropriate questions for the context being evaluated.

Designing a proper survey is a complex activity, as it involves several steps that may be difficult to carry out without specific expertise. Some of the initial phases in survey development, such as literature reviews, interviews, or focus groups, require time and methodological competence [25].

The proposed platform supports this process by offering a curated set of questions, collected over time from relevant literature and organized according to the underlying constructs they aim to measure. Nonetheless, users can add new questions, provided with literature references, allowing more experienced users to expand and tailor the instrument to their specific needs.

Questions are grouped by thematic categories, more precisely referred to as *constructs*, defined in educational research as latent variables that represent abstract concepts measured through multiple related items [25]. Examples of such constructs include self-efficacy, motivation, or future academic intentions. Each construct can be measured at different levels of detail. For instance, satisfaction may be assessed through a single global item or through more granular questions addressing specific aspects such as satisfaction with the activity, teamwork, or instructor interaction. Similarly, future career intentions in STEM fields may be captured through a general yes/no item or a more detailed question listing potential academic or career paths.

While the platform provides a validated and structured base, it allows users to control the level of depth and complexity of their evaluation. Users can select only a minimal set of items for a lightweight assessment or opt for a more in-depth investigation depending on their goals and available resources.

This modular structure also supports the automatic generation of tailored analysis notebooks. When a construct is assessed through multiple items, the analysis notebook includes pre-configured aggregation options, such as mean scores, sum scores, or principal component analysis, based on the selected configuration.

B. Data Collection and Validation

The data collection phase occurs outside the platform itself. The platform provides an export functionality that generates a survey in a format compatible with LimeSurvey, an open-source and self-hostable tool for questionnaire administration. LimeSurvey enables the collection of responses in a way that allows pre- and post-survey data to be linked while still preserving respondent anonymity through the use of pseudonymized identifiers.

A crucial aspect of this phase, as emphasized by Knekta et al. [15], is the importance of validating the survey instrument in the specific context in which it is used. Even well-designed instruments must be empirically tested to ensure they measure the intended constructs within the target population. The platform facilitates this validation process by providing pre-configured scripts for performing Exploratory Factor Analysis (EFA) and reliability checks, such as Cronbach's alpha, based on the collected data.

However, the implementation of the validation process, including when and how it is conducted, remains the responsibility of the organizers. Depending on available resources, this may involve administering the generated surveys to a control group before the camp, or performing post-hoc validation using the responses from a single cohort at the end of the activity.

Furthermore, generating the questionnaire in LimeSurvey through the platform enables the seamless export of data that can be directly analyzed using the pre-configured notebooks. By assigning matching labels to the survey questions during the generation process, the platform ensures that the collected data can be automatically linked with the corresponding analysis.

C. Data Analysis and Interpretation

Data analysis and interpretation represent the most critical phase of the evaluation process, enabling organizers and stakeholders to derive meaningful insights about the impact of the activities undertaken. However, data analysis is rarely a linear or fully generalizable process. As emphasized in the literature on educational measurement and social science research (e.g., Maxwell's work on qualitative inquiry [26]), effective analysis depends heavily on the context, goals, and structure of the collected data. For non-expert users, this complexity can pose significant challenges.

To address these difficulties, the proposed platform supports a dual approach. On one hand, it provides a set of ready-to-use Jupyter notebooks tailored to the survey configuration chosen by the user. These notebooks guide the analyst through a basic data analysis pipeline, including data import, cleaning, outlier detection, and the application of standard techniques such as regression and multivariate analysis. While the notebooks are not meant to replace the role of a trained data scientist, they offer a transparent and extensible framework that reduces the technical burden on users.

On the other hand, to further support interpretation, the platform includes an optional notebook equipped with prompts and suggestions designed to work in conjunction with large language models (LLMs), aligning with a growing trend in educational research to adopt LLMs for supporting instructional tasks [27], [28]. This tool assists users, particularly those without a strong background in statistics, in understanding the meaning of key outputs and how they might inform future program design.

IV. Platform Architecture

The platform is built to be released free and open-source. Surveys in fields involving underage students often require strict privacy measures; as a consequence, the ability to self-host the platform allows for more control over the data being handled. The choice of technologies reflects these principles; in particular, the platform is divided into a central application and several integrated external tools, all open-source and self-hostable.

As shown in Figure 2, the central platform is responsible for managing the survey lifecycle, from creation to deployment, as well as integrating with external tools for data collection and analysis. The technology chosen for this part is Django [3], a Python framework widely used in open-source projects. Django was selected for its scalability, security, and extensibility. Its internal architecture is well-suited to handle complex workflows like survey management while also providing an admin interface that simplifies user interaction.

Once the user completes the survey definition, a Python package called Citric[4] is used to interact with the LimeSurvey platform[5], automating the transfer of the survey to LimeSurvey for deployment. LimeSurvey is a well-known, open-source, self-hosted tool for survey creation and data collection, which aligns with the platform's goals of maintaining control over survey data. While many survey platforms are not specifically designed for scientific research [17], LimeSurvey is a widely used and well-maintained solution that reduces both development time and maintenance burden.

The central platform also generates Jupyter Notebooks [6] for data analysis. The choice of Jupyter Notebooks is based on their widespread use in the research community and their ability to integrate code and text in a single, interactive

[3]https://www.djangoproject.com/
[4]https://github.com/edgarrmondragon/citric
[5]https://www.limesurvey.org/
[6]https://jupyter.org/

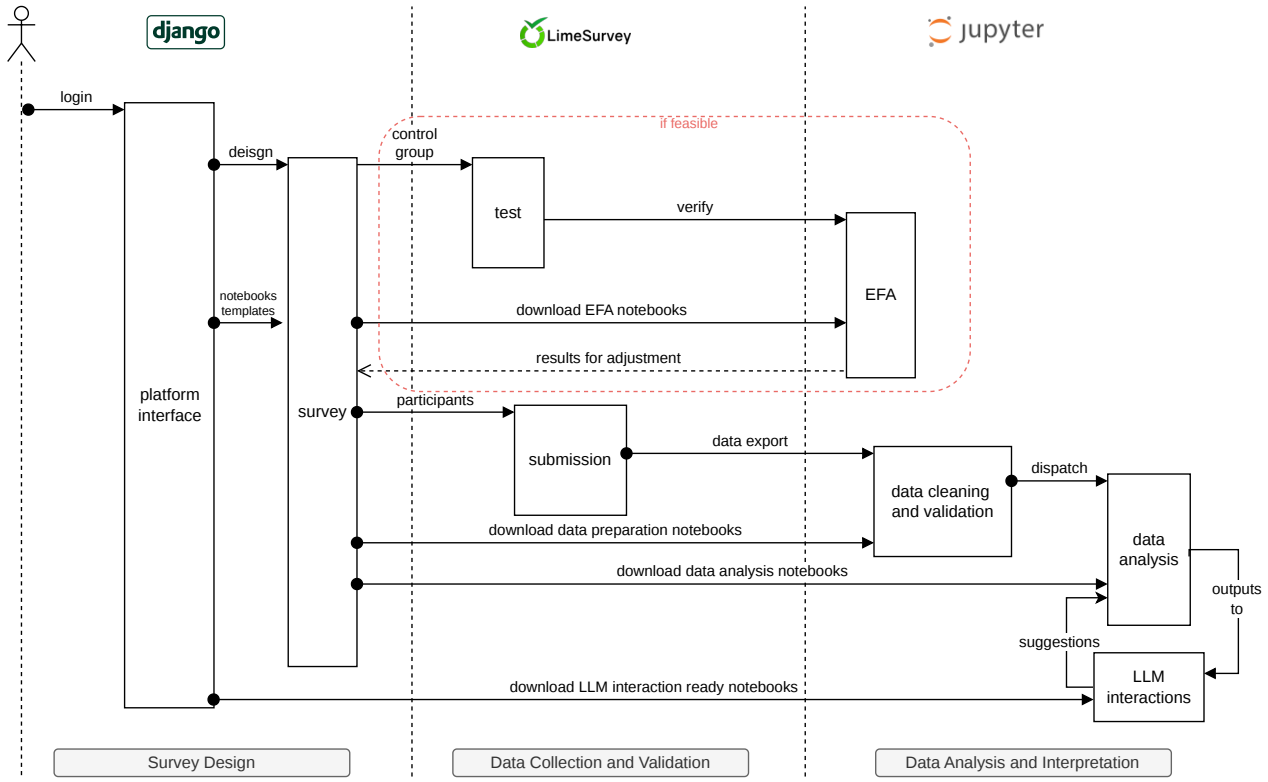

Fig. 2. User interaction flow

environment. This supports reproducibility in research [29], allowing users to document their analysis while performing it. Additionally, the Django template system enabled the customization of notebooks for each survey, allowing the creation of notebooks that fit research questions chosen in the design phase.

The generated notebooks integrate both explanatory text and executable code, allowing users to understand each step of the analysis and modify it as needed. The notebooks are divided into preparation and analysis notebooks. The formers are designed to verify the validity of the survey through Exploratory Factor Analysis (EFA), clean the data by removing any outliers, and assess consistency using Cronbach's alpha. Each step is accompanied by visualizations of the results, such as scree plots and factor loadings, to help users better understand the findings. The analysis notebooks assist users in performing further analysis; in addition to providing descriptive representations of the results, they offer templates for executing regression analysis on key variables and provide appropriate verifications and suggestions for improving the model fit.

Furthermore, the platform offers additional notebooks with the integration of Large Language Models (LLMs) to support users in interpreting their analysis. The main purpose is to provide contextualised suggestions and interpretations for users who do not have a strong background in data analysis. While it is well-known that LLMs cannot substitute for statisticians and that improper use of LLMs can even lead to mistakes that invalidate results, recent research by Zhu et al. [30] demonstrated that certain LLMs can perform reasonably well for relatively simple tasks, such as providing interpretations and suggestions for basic statistical analysis. This research helped guide the selection of an LLM that balances performance with the specific needs of our platform.

The final step in ensuring that the platform is truly self-hostable was the creation of a standardized deployment process that includes all the necessary components. Docker, and specifically Docker Compose, was selected for this task. A Docker Compose configuration[7] is provided, which encapsulates all the services required for the platform to function correctly. This configuration allows end users to launch the entire platform on their premises with minimal setup, requiring only a machine with Docker pre-installed. This approach greatly simplifies deployment, ensures consistency across different environments, and reduces the setup effort for users, allowing them to take full control of their data and platform infrastructure. All source code, configuration files, and usage instructions are openly available in the project repository at

[7]https://www.docker.com/

https://gitlab.com/frfaenza/elevate-ai.

V. User Interaction Workflow

The platform's design prioritizes usability, allowing non-expert users, such as researchers and educators, to effortlessly deploy and analyze surveys for outreach programs. The interaction process is intuitive and modular, guiding users through the entire survey lifecycle, from design to data analysis. The proposed workflow reflects our own experience with outreach initiatives [31].

Following the flow presented in Figure 2, users begin by creating a new evaluation process. They define a title, a description, and optional metadata to help document the assessment process. Additionally, they specify a start and end date, which are used by the platform to automate the creation of surveys in LimeSurvey. The overall sequence of steps and system interactions is also graphically illustrated in Figure 3.

The next step is to select questions to include in the evaluation. To encourage adoption, a set of questions frequently used in the literature is already embedded in the Django migration system, ensuring the database is pre-populated at deployment. This does not prevent users from creating new, custom questions tailored to their specific needs. Each question is linked to a literature reference, making its origin traceable. Questions are also grouped under constructs (e.g., self-efficacy, motivation), a feature implemented in the model schema to facilitate notebook generation for subsequent analysis.

By default, the evaluation is structured as a pre-post comparison. During survey composition, the user can specify whether each question should appear in the pre-survey, the post-survey, or both. It is also possible to mark questions as mandatory or conditionally displayed based on prior responses. While these are just a subset of LimeSurvey's full capabilities, they cover the most commonly used options in educational research.

Once the questions, order, and pre/post assignment have been finalized, the user can invoke the function that uses Citric to generate and upload the surveys to the LimeSurvey instance.

An essential aspect of the process is survey validation via Exploratory Factor Analysis (EFA). As highlighted in Figure 2, this step is optional and marked in red, as its feasibility depends on the specific constraints of the organizers. A robust validation process ideally requires a control group to complete the survey, followed by EFA to verify that the constructs align with the collected data (see Knekta et al. [15]). The platform provides a notebook to perform EFA, allowing practitioners to make informed decisions about modifying or retaining the current questionnaire structure.

If validation with a control group is not feasible, the user can proceed directly to deployment using LimeSurvey's native interface and then validate with participants' data. The platform generates both a public link and a QR code for distribution, though direct interaction with LimeSurvey remains possible for users with advanced needs.

One factor that determines whether further interaction with the LimeSurvey platform is needed is the method used to connect pre- and post-surveys. The default approach allows for multiple options. The first and easiest way is to use participants' email addresses. In this case, the provided notebook will assign a pseudonym and remove the email for privacy reasons. Alternatives could be assigning random codes or letting users generate their own codes through a specific question.

Once data has been collected, the analysis phase begins. The deployment includes a Jupyter server, allowing users to launch pre-generated analysis notebooks directly. Based on the selected survey configuration, the following notebooks are provided:

- **data cleaning notebook**: Processes exported responses, cleans the dataset, and connects pre- and post-survey entries
- **outliers and validity notebook**: Identifies outliers (e.g., through response time [32] _) and performs validity checks using EFA and Cronbach's alpha
- **descriptive analysis notebook**: Provides summary statistics and visualizations based on survey constructs
- **regression notebook**: Offers pre-configured regression models with customization instructions

While the analysis notebooks include embedded explanations, statistical knowledge is still required. To support non-expert users, the platform offers an additional notebook that integrates with a Large Language Model (LLM). The user will be able to carry on the analysis while interacting with LLM for assistance.

Following Zhu et al. [30], who compared LLM performance in statistical reasoning, the platform incorporates LLM-generated prompts to help users interpret outputs and navigate the analysis steps. To preserve the open-source and self-hostable nature of the platform, we support integration with local models via the Hugging Face framework[8]. Although these models can run on modern laptops with longer response times, they may be too resource-intensive for some users. Therefore, we also provide a notebook that connects to the ChatGPT-4 API[9], accompanied by prompt templates similar to those used locally. In both cases, however, users must register on a platform, either Hugging Face or OpenAI (for ChatGPT), to obtain an API key, which must be inserted into the notebook for it to function correctly.

VI. Conclusions and Future Development

In this paper, we presented an open-source, self-hostable platform designed to support the evaluation of STEM and CS outreach initiatives. The platform aims to address the lack of appropriate assessment during extracurricular activities. By automating survey generation, facilitating data validation, and guiding users through standard analysis workflows, including support from integrated Large Language Models (LLMs), our platform lowers the technical barrier to performing robust evaluations in informal settings.

The platform's design is grounded in educational research and informed by practical field experience. It aligns with

[8]https://huggingface.co
[9]https://chatgpt.com/

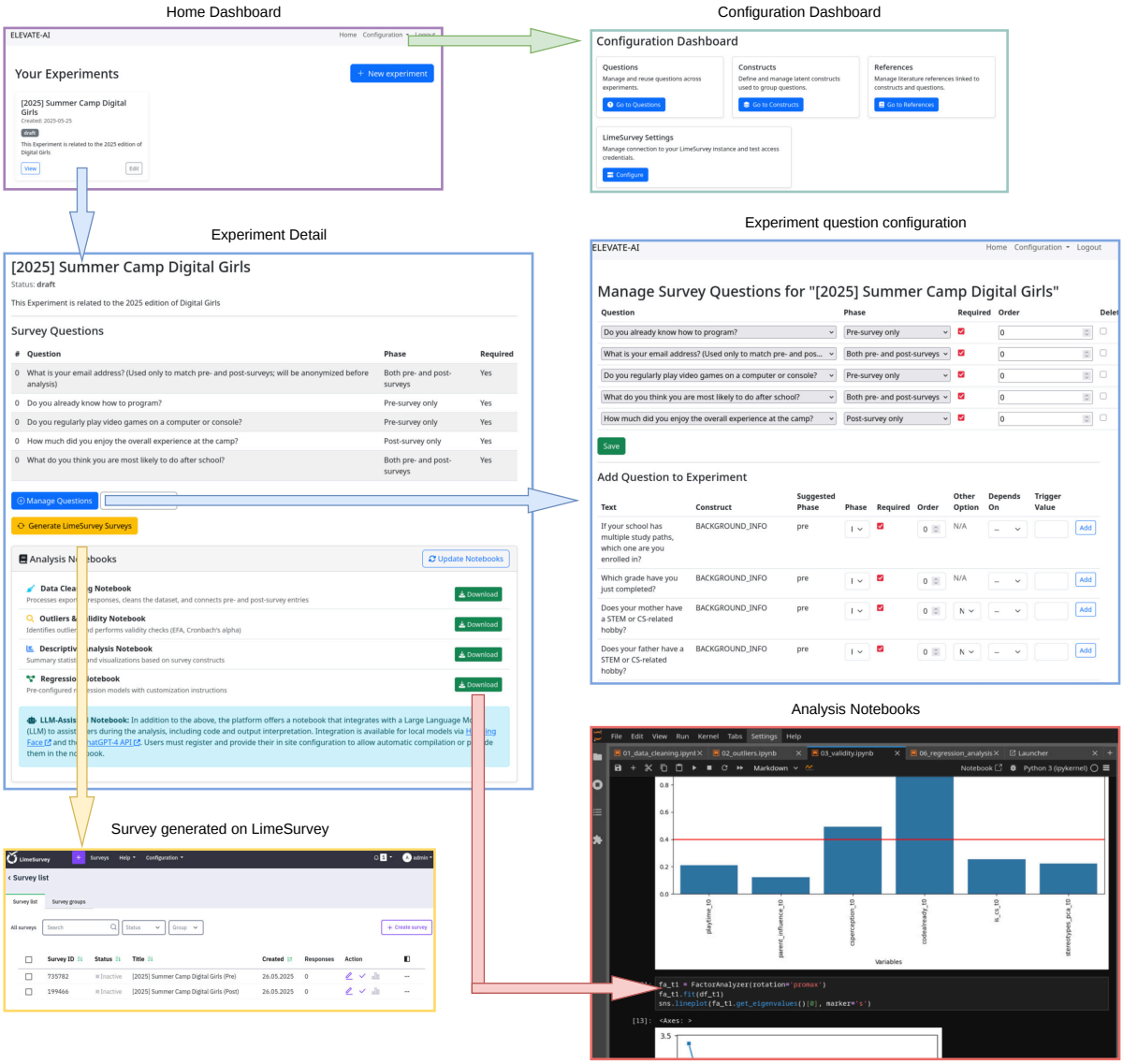

Fig. 3. Use case flow

widely adopted practices in survey validation and quantitative analysis. The tool builds upon several years of implementation in gender-focused STEM outreach, notably the Digital Girls initiative, and has been refined to meet the needs of organizers working in contexts where both methodological rigor and usability are essential.

The project's open-source nature is intended to foster a community of contributors who can collaboratively expand the platform by adding validated questions with corresponding literature references, sharing analysis templates, and suggesting feature improvements. All source code, deployment files, and usage documentation are freely available at https://gitlab.com/frfaenza/elevate-ai.

Looking ahead, several lines of development are planned. First, we aim to expand the LLM-assisted functionalities to support users not only during data analysis but also in survey design and validation. Furthermore, inspired by recent findings from Zhu et al. [30], we are exploring the possibility of integrating a fine-tuned model optimized explicitly for the types of evaluation workflows supported by the platform.

Finally, we plan to further simplify the user experience by streamlining access to the analysis tools. A step in this direction is the integration of Jupyter notebooks directly within the Django interface, removing the need for separate notebook deployment or configuration. Similar approaches have been adopted in prior research focused on improving the accessi-

bility of computational notebooks for non-technical users [33].

REFERENCES

[1] European Commission, "STEM Education Strategic Plan," https://education.ec.europa.eu/sites/default/files/2025-03/STEM_Education_Strategic_Plan_COM_2025_89_1_EN_0.pdf, Mar. 2025, accessed: 2025-07-15.

[2] OECD, "Gender, Education and Skills," https://www.oecd.org/en/publications/gender-education-and-skills_34680dd5-en.html, 2023, accessed: 2025-07-15.

[3] M. M. Msambwa, K. Daniel, C. Lianyu, and F. Antony, "A systematic review using feminist perspectives on the factors affecting girls' participation in stem subjects," *Science & Education*, pp. 1–32, 2024. doi: 10.1007/s11191-024-00524-0

[4] F. Beroíza-Valenzuela and N. Salas-Guzmán, "Stem and gender gap: A systematic review in wos, scopus, and eric databases (2012–2022)," in *Frontiers in Education*, vol. 9. Frontiers Media SA, 2024. doi: 10.3389/feduc.2024.1378640 p. 1378640.

[5] A. S. Rorrer, "An evaluation capacity building toolkit for principal investigators of undergraduate research experiences: A demonstration of transforming theory into practice," *Evaluation and program planning*, vol. 55, pp. 103–111, 2016. doi: 10.1016/j.evalprogplan.2015.12.006

[6] X. Chen, D. Zou, H. Xie, G. Cheng, and C. Liu, "Two decades of artificial intelligence in education," *Educational Technology & Society*, vol. 25, no. 1, pp. 28–47, 2022.

[7] E. A. Locke, "Self-efficacy: The exercise of control," *Personnel psychology*, vol. 50, no. 3, p. 801, 1997.

[8] B. P. Cipriano, N. Fachada, and P. Alves, "Drop project: An automatic assessment tool for programming assignments," *SoftwareX 18 (2022) 101079*, vol. 18, 2022. doi: 10.1016/j.softx.2022.101079

[9] S. H. Edwards and M. A. Perez-Quinones, "Web-cat: automatically grading programming assignments," in *Proceedings of the 13th annual conference on Innovation and technology in computer science education*, 2008. doi: 10.1145/1384271.1384371 pp. 328–328.

[10] J. Majerník, "E-assessment management system for comprehensive assessment of medical students knowledge," in *2018 Federated Conference on Computer Science and Information Systems (FedCSIS)*. IEEE, 2018. doi: 10.15439/2018F138 pp. 795–799.

[11] L. Stanescu and B. Savu, "Automatic assessment of narrative answers using information retrieval techniques," in *2019 Federated Conference on Computer Science and Information Systems (FedCSIS)*. IEEE, 2019. doi: 10.15439/2019F96 pp. 355–358.

[12] B. Van Dusen, "Lasso: A new tool to support instructors and researchers," *arXiv preprint arXiv:1812.02299*, 2018. doi: 10.48550/arXiv.1812.02299

[13] J. Hattie, *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. routledge, 2008.

[14] A. Decker and M. M. McGill, "A topical review of evaluation instruments for computing education," in *Proceedings of the 50th ACM Technical Symposium on Computer Science Education*, 2019. doi: 10.1145/3287324.3287393 pp. 558–564.

[15] E. Knekta, C. Runyon, and S. Eddy, "One size doesn't fit all: Using factor analysis to gather validity evidence when using surveys in your research," *CBE—Life Sciences Education*, vol. 18, no. 1, p. rm1, 2019. doi: 10.1187/cbe.18-04-0064

[16] I. Fronza, P. Ihantola, O.-P. Riikola, G. Iaccarino, T. Mikkonen, L. García Rytman, V. Lappalainen, C. Rebollo Santamaría, I. Remolar Quintana, and V. Rossano, "Towards s'more connected coding camps," in *Proceedings of the 56th ACM Technical Symposium on Computer Science Education V. 1*, 2025. doi: 10.1145/3641554.3701849 pp. 353–359.

[17] D. Guse, H. R. Orefice, G. Reimers, and O. Hohlfeld, "Thefragebogen: A web browser-based questionnaire framework for scientific research," in *2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE, 2019. doi: 10.1109/QoMEX.2019.8743231 pp. 1–3.

[18] F. Faenza, C. Canali, M. Colajanni, and A. Carbonaro, "The digital girls response to pandemic: Impacts of in presence and online extracurricular activities on girls future academic choices," *Education Sciences*, vol. 11, no. 11, p. 715, 2021. doi: 10.3390/educsci11110715

[19] F. Faenza, C. Canali, and A. Carbonaro, "Ict extra-curricular activities: The "digital girls" case study for the development of human capital," in *The International Research & Innovation Forum*. Springer, 2021. doi: 10.1007/978-3-030-84311-3_18 pp. 193–205.

[20] P. Cobb, K. Jackson, and C. Dunlap, "Design research: An analysis and critique," in *Handbook of international research in mathematics education*. Routledge, 2015, pp. 481–503.

[21] L. Markauskaite, P. Freebody, and J. Irwin, *Methodological choice and design: Scholarship, policy and practice in social and educational research*. Springer Science & Business Media, 2010, vol. 9.

[22] C. Canali and F. Faenza, "An evaluation tool for extracurricular activities to reduce the gender gap in computer science," in *ICGR 2023 6th International Conference on Gender Research*. Academic Conferences and publishing limited, 2023. doi: 10.34190/icgr.6.1.1026

[23] I. Ajzen, "Nature and operation of attitudes," *Annual review of psychology*, vol. 52, no. 1, pp. 27–58, 2001. doi: 10.1146/annurev.psych.52.1.27

[24] J. F. McKenzie, M. L. Wood, J. E. Kotecki, J. K. Clark, and R. A. Brey, "Establishing content validity: Using qualitative and quantitative steps." *American Journal of Health Behavior*, vol. 23, no. 4, 1999.

[25] A. R. Artino Jr, J. S. La Rochelle, K. J. Dezee, and H. Gehlbach, "Developing questionnaires for educational research: Amee guide no. 87," *Medical teacher*, vol. 36, no. 6, pp. 463–474, 2014. doi: 10.3109/0142159X.2014.889814

[26] J. A. Maxwell, *Qualitative research design: An interactive approach: An interactive approach*. sage, 2013.

[27] E. Rudolph, H. Seer, C. Mothes, and J. Albrecht, "Automated feedback generation in an intelligent tutoring system for counselor education," in *2024 19th Conference on Computer Science and Intelligence Systems (FedCSIS)*. IEEE, 2024. doi: 10.15439/2024F1649 pp. 501–512.

[28] T. C. Freitas, M. J. V. Pereira, A. C. Neto, and P. R. Henriques, "Goliath, a programming exercises generator supported by ai," in *2024 19th Conference on Computer Science and Intelligence Systems (FedCSIS)*. IEEE, 2024. doi: 10.15439/2024F8479 pp. 331–342.

[29] F. Perez and B. E. Granger, "Project jupyter: Computational narratives as the engine of collaborative data science," *Retrieved September*, vol. 11, no. 207, p. 108, 2015.

[30] Y. Zhu, S. Du, B. Li, Y. Luo, and N. Tang, "Are large language models good statisticians?" *Advances in Neural Information Processing Systems*, vol. 37, pp. 62 697–62 731, 2024.

[31] D. Maniglia, F. Faenza, and C. Canali, "Empowering girls in cs: The impact of digital girls outreach camp," in *Proceedings of The 7th International Conference on Gender Research*. Academic Conferences and publishing limited, 2025. doi: 10.34190/icgr.8.1.3528

[32] G. Szyjewski and L. Fabisiak, "Survey as a source of low quality research data," in *2017 Federated Conference on Computer Science and Information Systems (FedCSIS)*. IEEE, 2017. doi: 10.15439/2017F266 pp. 939–943.

[33] S. Pigozzi, F. Faenza, and C. Canali, "Sophon: An extensible platform for collaborative research," in *Practice and Experience in Advanced Research Computing 2022: Revolutionary: Computing, Connections, You*, 2022. doi: 10.1145/3491418.3535163 pp. 1–4.