What impact does real-time written-visual feedback, through the lens of different design roles, have on novel students' attitudes toward learning programming?

Jesse Strijker
j.j.strijker@student.tue.nl
Technical university of Eindhoven
Eindhoven, Noord-Brabant, Netherlands

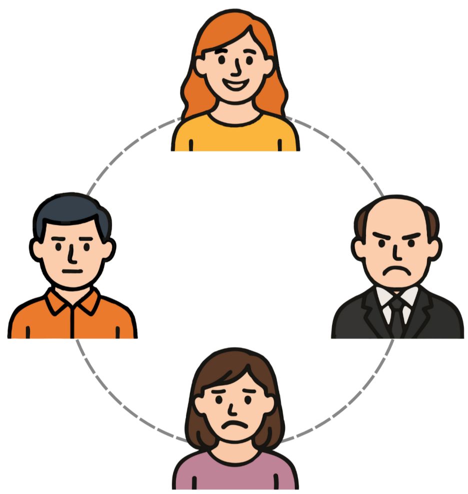

Figure 1: The four avatars used in the experiment

Abstract

This study examines how four digital-twin feedback roles, varying valence and character type, affect novice programmers' self-efficacy and perceived difficulty in a programming task (N=41). Utilizing a pre/post measurement results showed a significant Time x Valence interaction (F(1, 36)=5.18, p=.029), indicating positive-valence feedback provides greater self-efficacy increase, whereas four variants had no effect on perceived difficulty. This suggests that for digital twins in programming education, emotional tone is a primary lever for boosting learner confidence.

Keywords

Feedback, Education, Programming, Design

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

ACM Reference Format:
Jesse Strijker. 2025. What impact does real-time written-visual feedback, through the lens of different design roles, have on novel students' attitudes toward learning programming?. In . ACM, New York, NY, USA, 21 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction V2

Digital twins are interactive virtual counterparts that not only mirror physical systems but also respond, advise, and correct in real time [40]. This makes them more than passive representations, they become interactive agents. As they take action, provide feedback, or highlight anomalies, they inevitably begin to take on a role.

The way a digital twin communicates information, affects how users experience it [43]. This is not a new insight within human-computer interaction, the way a system "speaks" or presents itself influences user perception, trust, and engagement [6, 15, 26, 29, 32]. However, for digital twins, this aspect is underexplored. Design efforts tend to focus on accuracy, not necessarily on the character that emerges from the twin's ongoing behavior [40].

The same monitoring, analysis, and predictive capabilities defining a digital twin could also make it suitable for educational use [23]. In a learning environment, a digital twin can track student behavior, analyze errors or misconceptions, and provide targeted feedback in real time. A digital twin, of not an object, but an interaction. This aligns with programming education, where learners benefit from immediate, context-sensitive responses [3, 14]. However, the way a system watches and responds may shape how a student experiences the learning task [14, 24].

This study investigates what impact different feedback “roles” of a digital twin have on student attitudes in a programming context. A custom Arduino-based learning environment was developed in which a digital twin provided live feedback through written comments and a visual avatar. Four variants were designed, each combining a tone (positive or negative) with a character style (authoritative, brevity, emotion). While the prototype used in this study does not capture the full analytical depth of an industrial digital twin, it mimics its core structure, monitoring user behavior and responding with adaptive feedback in real time. The study examines how these variations affect learners' self-efficacy (SE) and perceived difficulty (PD).

By approaching feedback design through the lens of digital twin characters, this study highlights the importance of delivery style as a deliberate and impactful design choice on student persistence [2, 8, 9].

This paper scopes around solely real-time feedback, but it is relevant to mention that only providing real-time feedback or providing feedback on outcomes can be significantly less efficient [2]. In other words, feedback cannot be seen as isolated moments in the instructional cycle.

2 Background - Related Work

In this paper, feedback can be defined as information given to a student regarding factors of their understanding and performance. Feedback does not have an effect in a vacuum, it should appear in a learning context where a student receives an initial instruction, responds to it, and then receives information on aspects of their performance on their understanding or performance [18]. Generally speaking, people like to receive feedback about how they are performing even if the impact is negligible [1]. But while most feedback has an impact on achievement, the effect size of feedback is strongly dependent on the type of feedback.

Hattie and Timperley [14] define a multitude of factors that are relevant for the effectiveness of feedback: *the focus of feedback, the timing of feedback, positive and negative feedback (valence), source of feedback (direction)*, and in their updated article [3] they expand it with the *medium/channel of feedback and user variables*. While they address a lot of relevant factors it can be augmented with factors identified by Panadero & Lipnevich [24] who reviewed feedback models and theories. They expand the list by adding *context* as a relevant factor, and rename some of the factors names as coined by Hattie & Timperley for clarity. Each of these factors also contain their own subcategories (see Figure 2). To provide a clear overview, table 1 provides an oversight of each factor, subcategory, and type of impact it has.

Type: Hattie and Timperley propose a model of feedback, to identify when feedback has the most impact. They provide an interesting narrative of dividing the focus of feedback into four categories. **Task-level (FT)** feedback provides information on correctness, behavior, tidiness or another element related to completion of a task. It generally provides a high impact on immediate learning outcomes [14]. Task feedback might prompt learners to seek additional, alternative, or more precise knowledge. A drawback however can be that task-level feedback is often specific to a question at hand and cannot be generalized to other questions [39]. **Process-level (FP)** feedback (information on strategies) is often even more effective, fostering deeper understanding and problem-solving [14]. This feedback is often more aimed at directing the student to a strategy or process, to help solve an issue. It has been noted that a combination of FP and FT can be very effective, were FT has the effect of improving confidence and SE, and thereby making acting on FP more appealing [10]. **Self-regulation (FR)** feedback is aimed at the student themselves and their knowledge ("you are already aware of this knowledge piece, check if you have included it"). It has a high impact (meta-analysis found Hedges' $g \approx 0.88$) on learning, strategy use and SE [14]. In contrast, **Self-level (FS)** feedback, or praise/critique of the learner's traits (e.g.: "Good job!"), is low impact. In positive cases possibly improving motivation slightly but doing little for achievement [14]. Often it has no task or process information and is generally considered ineffective (not disliked however). Hattie and Timperley go as far as claiming that self-level feedback combined with task-level, brings down the effectiveness of task-level as opposed to just by itself. They also address 2 meta-analyses giving 0.12 and 0.09 as an effect size, while no praise had a higher impact on the students achievement (0.34).

Timing: Both Immediate feedback and Delayed feedback produce strong learning gains [14]. An argument for immediate feedback is that when relating to task-level feedback, having an accurate memory of an assignment, upon completion is vital [14]. Additionally immediate error correction can result in quicker learning rates [20]. Whereas delayed feedback can be stronger depending on the difficulty of the task at hand [7]. This could be attributed to the fact delayed feedback provides the recipient more time to process a difficult task. With easy tasks this is not needed and potentially even detrimental because of the importance of having an accurate memory [7].

Source: Teacher feedback typically has a high impact (meta-analysis in writing found $g \approx 0.90$) [38], whereas Peer feedback is moderate ($g \approx 0.68$, rising to 0.74 with training) [38]. Self-generated feedback (learner self-assessment) is not directly researcher-controllable and its impact varies with the student's skill [24]. Computer-delivered (automated) feedback returns high effects (medium-high to high) on cognitive outcomes [3].

Medium/channel: Written (textual) comments are highly effective, significantly more so than a numeric grade/score, which has low

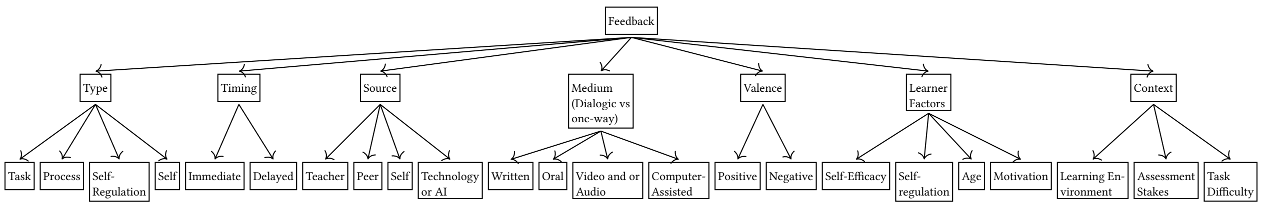

Figure 2: A flowchart showing different types of feedback elements, and their subcategories.

impact [3]. Oral (spoken) feedback and video feedback also show high effectiveness, similar to computer-assisted delivery [3]. But since they all show good effectiveness, between and within group changing of channel proved to be non-significant [3]. This conflicts with previous findings, for example oral feedback is less effective than written feedback [4].

Valence: The effect of positive (praise) feedback is generally low, since praise often adds little information [3]. Likewise, negative (critical) feedback without constructive guidance has low or inconsistent effect [3]. Therefore Hattie and Timperley claim the effect of positive or negative feedback is more dependent "on the level" it is executed (FT, FP, etc.) [14, p. 98], significantly impacting its effectiveness positively. Both positive or negative feedback can be useful at the level of task feedback. Where on the level of self-regulation positive feedback increases motivation and negative decreases motivation. In other words, simply giving positive or negative remarks does little for learning, though it may slightly influence motivation or self-esteem depending on task-level [3].

Learner Factors: Student characteristics strongly moderate feedback impact. High prior knowledge is critical (high impact), students need sufficient base knowledge to benefit [24]. In contrast, student motivation and SE tend to show lower impacts: feedback studies often find large cognitive gains but only modest changes in motivation or confidence [3].

Context: The educational context matters greatly. Feedback is more effective in higher grades/higher education [23]. The study setting matters: controlled lab studies produce larger effects than real classroom settings [23]. Finally, discipline or subject area significantly influences feedback impact, with some subjects showing stronger gains than others [23].

Additional elements: Other factors, such as providing feedback on correct responses being more effective than providing feedback on incorrect responses [14], restate Kulhavy's statement that it is important to address faulty interpretations over a complete lack of understanding [18, p. 220]. It is also beneficial if the feedback includes previous attempts [14]. Lastly, the complexity of feedback is also important, if the complexity is high, students with high confidence are more likely to use it effectively, while students with low confidence will have more trouble [14]. The level of information, is also dictating [3, 14]. Simple forms of reinforcement and punishment have low effects, while high-information feedback is most effective [14]. The confidence of a student on their answer is also relevant, if a student expects an answer to be correct and gets it wrong feedback can have a bigger impact than it would vice-versa [19]. This influence on their receptiveness to feedback, is mainly addressed at feedback upon completion. However, this confidence becomes less relevant for feedback during the process, where it is welcomed as it might provide an easy way out. This addresses the element of instrumental help and executive help (hints vs answers). The effectiveness of either is dependent on the learner themselves [31]. Extrinsic rewards actually have a negative correlation to performance. They are detrimental to intrinsic motivation, especially for tasks perceived as interesting by the recipient [9].

Expansion of Background Literature

As stated in the Introduction, and expanded on in the Methodology, a deep dive into the medium of feedback (written and visual) provides some additional context for this paper.

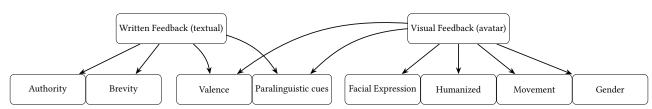

Figure 3: Factor oversight of elements influencing the reception of written and visual feedback

For the purposes of this paper, textual feedback refers specifically to written comments, whereas visual feedback is limited to feedback delivered via avatars. Other visual elements, such as graphs or progress bars, are acknowledged as valid channels for conveying information, but are excluded from our current scope. As illustrated in Figure 3, written and visual feedback share several sub-factors.

In written feedback, a distinction can be made between "authoritarian" or "controlling" feedback from "supportive" or "constructive" feedback. **Authority** can be defined as directive, critical, or judgmental language (e.g. "You failed to…")[29], where non-authoritative (also defined as supportive or constructive) can be defined as empathic, encouraging, or explanatory language (e.g. asking questions, suggesting alternatives, praising effort) [17]. Feedback must balance informativeness with conciseness. The **brevity** (the length of the written test) of feedback should attempt to present a high information density, but also keep it limited to key points to not overwhelm the recipient [26].

Both textual and avatar feedback are influenced by **valence** (the positivity or negativity of a message) and **paralinguistic** features

Factor	Subcategory	Impact	Impacted Variables
Type	Task (FT)	High - Strongly boosts learning on current tasks [14, 23]	Learning outcomes, task mastery [14, 23]
	Process (FP)	High - Typically more effective than task feedback [14, 23]	Learning strategies, problem-solving [14, 23]
	Self-Regulation (FR)	High (e.g. Hedges' $g \approx 0.88$) [14, 23]	Metacognitive skills, strategy use [14, 23]
	Self (FS)	Low - least effective [14]	Motivation/self-esteem (little effect on learning) [14]
Timing	Immediate Feedback	High - significantly improves performance [23]	Boosts immediate task performance and accuracy, accelerates skill acquisition [14, 23]
	Delayed Feedback	High - slightly more effective than immediate [23]	Enhances long-term retention and transfer, improves conceptual understanding and error correction over time [14]
Source	Teacher Feedback	High - large effect ($g \approx 0.90$ in writing) [38]	Learning outcomes (e.g., writing performance) [38]
	Peer Feedback	Medium - $g \approx 0.68$ (up to 0.74 with training) [38]	Learning outcomes (e.g., writing performance) [38]
	Self-Generated Feedback	Variable - depends on learner skill [24]	Self-regulated learning (feedback use) [24]
	Computer-Delivered feedback	High - medium-high to high effect [3]	Achievement (especially cognitive outcomes) [3]
Medium	Written Feedback (Text)	High - effective (written comments > scores) [3]	Learning outcomes [3]
	Grades/Scores	Low - minimal effect [3]	Student motivation (limited) [3]
	Oral (Spoken) Feedback	High - strong effects like video/audio [3]	Learning outcomes [3]
	Video Feedback	High - strong effects like audio [3]	Learning outcomes [3]
Valence	Positive (Praise) Feedback	High - can be effective on task level feedback [3]	Student motivation [3]
	Negative (Critical) Feedback	High - can be effective at self and task level feedback [3]	Learning outcome, can harm motivation [3]
Learner Factors	Prior Knowledge/Ability	High - crucial for using feedback effectively [24]	Learning gains (knowledge construction) [24]
	Motivation	Low - smaller feedback impact on motivation [3]	Engagement/persistence [3]
	Self-Efficacy/Confidence	Low - often only modest improvement [3]	Confidence in learning, strategy use [3]
Context	Educational Level (Age)	High - more effective in older students [23]	Learning gains (higher for older age groups) [23]
	Study Setting (Lab vs School)	High - lab studies show larger effects [23]	Measured learning improvement (higher in lab) [23]
	Discipline/Subject Area	High - significant variation by subject [23]	Learning outcomes (vary with discipline) [23]

Table 1: Detailed Impacts of Feedback-Related Factors on Learning Outcomes and Variables
***Note.* Effect sizes come from a variety of sources and were not all computed on the same metric, they should therefore be interpreted qualitatively rather than compared quantitatively.**

(such as icons or emoticons). In written feedback, valence is conveyed through word choice. Positive language (e.g., "well done," "great progress") versus negative language (e.g., "needs improvement," "unsatisfactory"). Paralinguistic cues often appear as icons or emoticons embedded within written text, although these cues are technically visual elements, they are most commonly incorporated into written feedback to moderate tone and can make comments feel "friendlier," thereby increasing students' commitment to follow advice [25].

Facial expressions, humanized features, movement, or gender in avatar design can also impact how feedback is received. **Humanized** avatars can foster trust [15], persuasiveness, trustworthiness and consistency [32]. Additionally, 'strong' and 'sudden' movements can convey anger, where 'light' and 'free' movements express happiness [34]. Even vertically upward and downward motions can signal emotions (excitement vs melancholy) [21]. Although the cited research focuses on non-humanized robots, the same principles can apply to avatar feedback, helping users interpret emotional content [21]. **Gender** can also influence the perception, male avatars are often perceived as more authoritative, whereas female avatars may be interpreted as more supportive [6]. Lastly, **facial expressions** are of the most information rich channels of conveying emotions. 'Positive' versus 'negative' emotional tones from avatars are influenced by upturned versus down turned lips [42] and authority through lowered furrowed brows with a firm, straight or slightly down turned mouth [33].

2.1 Conclusion of Background Literature

Studies have shown that individual feedback dimensions, such as type, timing, source, medium, valence, learner characteristics, and context, each play a distinct role in shaping learning outcomes. Likewise, research on written versus avatar-based feedback has identified factors like authority, brevity, emotional tone, and visual cues as influencing how students respond. However, most work considers these factors in isolation rather than examining their joint effects. While these studies examine individual feedback factors, there's limited research on how combinations of feedback factors, embodied in a particular "role," affect student outcomes. A gap exists in understanding how multiple feedback elements interact to influence outcomes. The combination of some of these factors in for example an anthropomorphic technology such as digital twins provides an interesting angle to look into.

3 Method

This study explored the impact of different roles in the design of real-time feedback, with different visual elements and written elements. Using a within-subjects (pre/post SE and PD) x between-groups (4) design, N=41 novices attempted a programming task with real-time feedback. To perceive this impact, a custom environment around programming was built including these real-time feedback roles.

3.1 Participants

Since a fear of programming is often more prevalent in those more unfamiliar with it, a target group of novice programmers was defined. These programmers were cluster sampled from a bachelors course on electronics, and convenience and snowball sampled from the department of Industrial Design (N=41).

To gather demographic information on the programming levels, self-reported estimated programming levels as proposed by Feigenspan et al. [13] combined with part B of Smith et al. [36] were used. Overlapping questions were omitted, and minor phrasing adapted to increase understandability (e.g. "college" into "university")(Appendix A). A self-reporting tool was chosen over a specialized test (e.g., a programming exam) because a test could affect students' SE and PD beyond the experiment.

3.2 Procedure

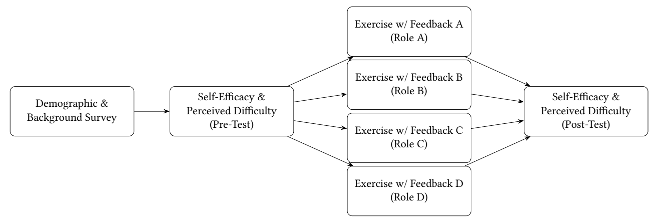

Figure 4: User Experiment flow

For assessing SE, the MSLQ [30] (Motivated Strategies for Learning Questionnaire) was used. This is a self-report instrument. To assess PD a similar approach as study 1 and 2 from Rodgers et al. [41] and the assessment of Brehm et al. [5] were used. Similarly to the MSLQ they provide Likert scale grading of questions on how difficult problems appear to be for the participant. The questions are aimed specifically at the topic of the study. (Appendix B).

4 The design

The design of the experiment consists of 3 main elements: construction of the programming assignment, design rationale behind the four feedback variants and technical implementation.

4.1 Programming Assignment

As mentioned in related works, it is important feedback addresses faulty interpretations and not a complete lack of understanding [18, p. 220]. To align the assignment with the target participants, the course instructor for the participants from the bachelor course provided the topic and programming language, so it aligned with the participants' curriculum, and could actually be beneficial for students at novice level.

Students modified Arduino's Blink example to use non-blocking timing (millis()) instead of delay(). The solution to this requires some logical thinking: saving the last timestamp a led toggled, checking if the timer minus this timestamp is larger than an interval, to toggle again. (figure 5)

4.2 Feedback Variant design

The four variants differ in design on the factors discussed in table 1 from the literature. Some of these variables were kept consistent, to ensure clarity when attempting to analyze the four variants. The rationale for not complete designerly freedom when designing the roles, comes down to if high-impact variables (table 1) (e.g., the amount of praise vs. amount of task info) vary between roles, they become the primary determining factor of observed differences.

```
void loop(){
   static unsigned long lastBlueLedToggle = 0;
   static boolean blueLedStatus = LOW;

   if(millis() - lastBlueLedToggle > 1000){
     lastBlueLedToggle = millis();
     blueLedStatus = !blueLedStatus;
     digitalWrite(11, blueLedStatus);
   }
}
```

```
void loop() {
 digitalWrite(LED_BUILTIN, HIGH);
 delay(1000);
 digitalWrite(LED_BUILTIN, LOW);
 delay(1000);
}
```

Figure 5: Left: Intended end result, Right: Original blink code

Table 2 explains **why** specific variables were locked, or why they variate between the 4 variants. The 'context' and 'learner' vary inherently between participants.

Feedback Category	Variates by design	Chosen Subcategory(s)	Why
Type	No	Combination FT and FP	Deemed best for real-time feedback according to background literature
Timing	No	Immediate	As inherently set by the scope and goal of this research
Source	No	Computer Delivered	Will be digitized, but not programmatic, since it will be prompted from variables in a design role by an LLM
Medium	Yes	Written, Visual	Approach anthropologic nature, achievable within scope of this project
Valence	Yes	Positive, Negative	
Learner Factors	Yes	Not controllable by researcher	X
Context	Yes	Not controllable by researcher	X

Table 2: Explanation behind High Impact choice

To not limit the study to just low impact variables, in which it might be hard to measure actual effects, a 2x2 structure of combinatory low impact and a high impact variables was defined (Table 3). An advantage of this structure is that it can confirm known effects first: By testing for valence differences, it can validate that the experiment works as expected, since the results should return inline with pre-existent literature.

	Authoritarian, brief, non-emotional (critical)	Non-authoritarian, extensive, emotional (supportive)
Positive Valence	X	X
Negative Valence	X	X

Table 3: Types of designed feedback variants

For written feedback: the content of feedback remains consistent throughout the variants on feedback type (e.g. feedback on the process or task), but varies in the identified elements of written feedback, such as authoritative tone, extensiveness and emotionality. Paralinguistic cues were avoided as a variance since it appeared technically challenging to reliably instruct an LLM to use specific paralinguistic cues. The prompt for the LLM is instructed to behave according to predefined rules (see appendix E). The visual elements provide a need for a visual avatar. This avatar was designed at the hand of the literature on gender, eyebrow and mouth positions inline with the written prompts(e.g. males being seen as more authoritative).

A pilot test was carried out to mitigate researcher bias with the open ended choices in the avatar design. The pilot test evaluated whether each avatars design was associated with its corresponding written prompt. Two independent pilot testers were both presented with all avatars and prompts in randomized order and asked to match them. Both raters accurately matched the avatar to its intended prompt, suggesting the visual and textual elements were perceived as intended.

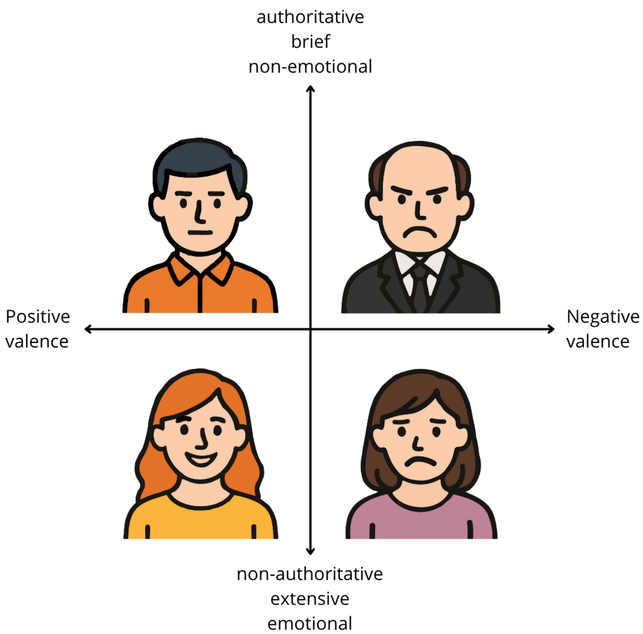

Figure 6: The four variants of avatar design. The left versus the right side shows the valence, whereas the bottom versus the top side shows character type

4.3 Technical Implementation

A web-based environment (SurveyJS [37] + AVR8js [35], Express APIs) provides a programming interface for Arduino code, and stores survey data on DataFoundry [11]. It queries GPT-4.1-nano [27] from the backend for with participant code and assigned variant to return feedback based on role. (Full architecture + prompts in Appendix C,D,E). In the backend the user code and their assigned variant are modified into an adaptive prompt (Appendix E) which is consequently sent to gpt-4.1-nano [27, 28]. Which was chosen for its rapid response rate and low token count while maintaining relatively high quality. *Note: readme on GitHub provides more elaboration (appendix D).*

A pitfall warned for by Timperley and Parr on interface design, is when feedback is not focused on the completion of a goal, that such feedback is ineffective in creating better understanding [16, 22]. So for the experiment, it was made sure the goal was accessible at all time.

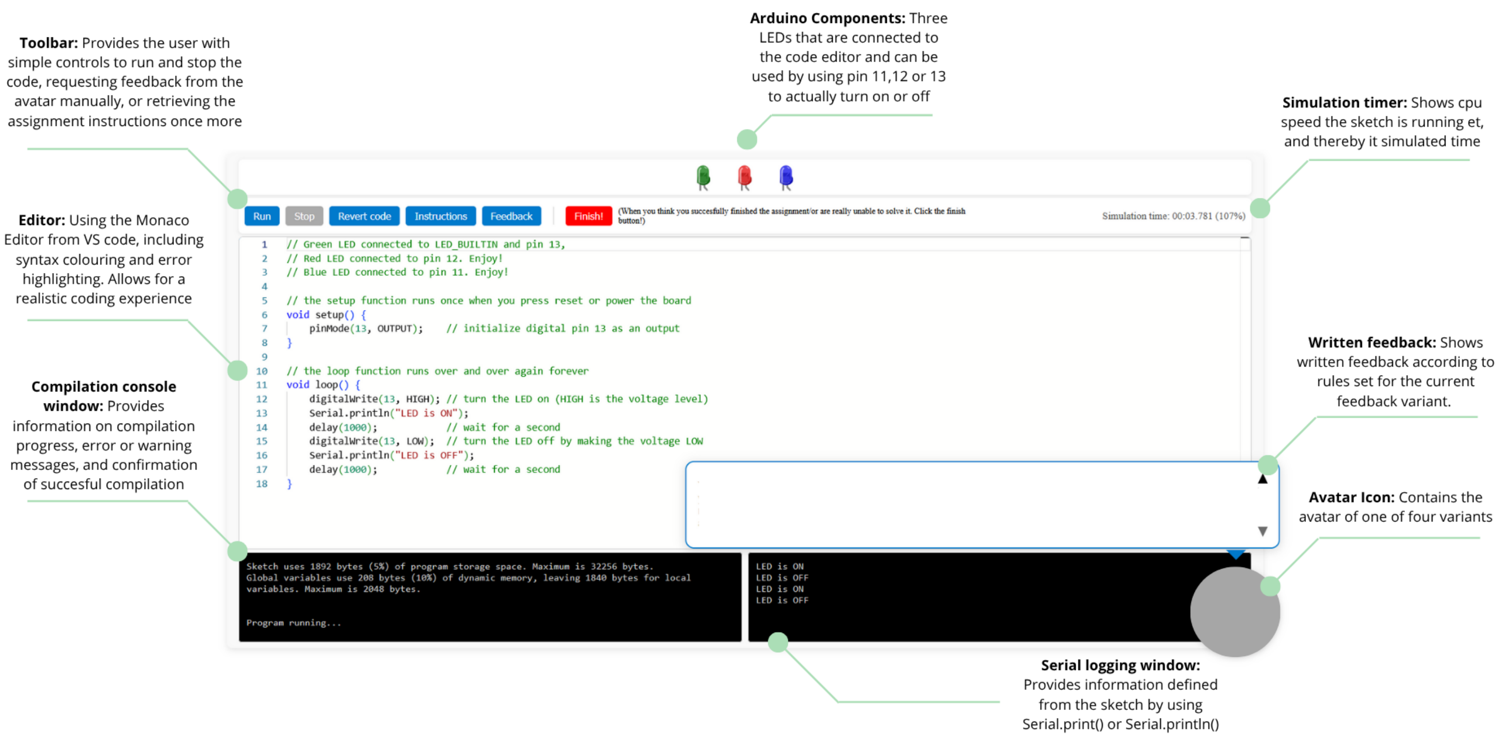

Figure 7: Interface design shown to participants

5 Results

From the experiment (N=41), data on the *demographics*, pre/post *SE* and *PD*, *time spent*, *attention to feedback* and *qualitative data* were collected. With the explorative nature of the study, results can be indicative of trends, but definitive conclusions should be made with caution.

The results consist of three main elements: the coding of the collected data, a quantitative analysis and a short qualitative analysis on researcher notes and participant remarks.

5.1 Data coding

To be able to use all the data, some of the data had to be recoded.

Self-efficacy
The MSLQ [30] is a validated scale. The tool instructs to take the average of all answered questions. This provides a numerical score for each participants' SE.

Perceived difficulty
Question 2 (see Appendix B.2) was recoded (inverted). To ensure comparability across the different scales the three questions were z-scored (i.e., standardized to have a mean of 0 and standard deviation of 1). To then also create a numerical score for PD, each participant's standardized scores were averaged.

Demographic data
Extensive demographic data was collected to assess programming level, but this approach included a small error: the source studies indicated only four items reliably predicted programming level. Consequently, only Questions 2, 3, 8, and 11 were kept to measure programming experience accurately [13], [36].

To create a numerical value of the participants' general programming level, the four aforementioned questions were also z-scored to ensure comparability across scales. Each participant's standardized scores were then averaged to compute a General Programming Score. Categorical responses (e.g., amount of self-learning time (q2 and q3)) were numerically coded based on increasing intensity.

Data exclusion/filtering
One data entry had to be excluded, since P26 only spent 5 seconds on the experiment.

5.2 Quantitative Data Analysis

Because the four feedback variants decompose into two dimensions (Valence & Character Type), the 4-group comparisons are first examined and then the underlying valence and character effects using a 2 x 2 approach are unpacked. Data on feedback variants provide insight into the role as a whole, where looking at valence and character type, provides insight on their respective effect regardless of variant.

The four feedback variants will be identified as V1, V2, V3 and V4:

- V1: positive valence with authoritarian, brief, non-emotional (or even critical) character type
- V2: negative valence with authoritarian, brief, non-emotional (or even critical) character type
- V3: positive valence with non-authoritarian, extensive, emotional (or supportive) character type
- V4: negative valence with non-authoritarian, extensive, emotional (or supportive) character type

5.2.1 *The impact of variants, valence and character type on self efficacy.* **4-Group Mixed-Design ANOVA**

A mixed-design ANOVA with repeated measures on Time (Pre vs. Post) and a between-subjects factor of Feedback Variants revealed no significant main effect of Time, Feedback Variant or Time x Feedback Variant interaction (Table 4). It can however be observed there is a visual trend differing the 4 variants on mean increase and decrease (figure 8).

Table 4: Mixed-Design ANOVA for Self-Efficacy by Feedback Variant

Effect	df	F	p	η_p^2
Time (Pre vs. Post)	1, 36	2.84	.100	.073
Feedback Variant (4-level)	3, 36	0.40	.751	.033
Time × Variant	3, 36	2.17	.108	.153

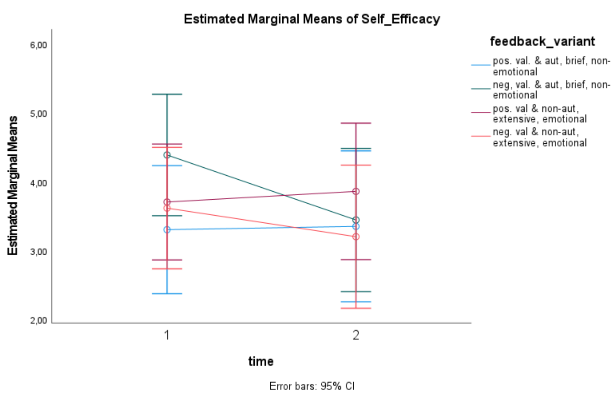

Figure 8: Pre and Post means on Self-Efficacy per variant

2x2x2 Mixed-Design ANOVA

A mixed-design ANOVA with repeated measures on Time and between-subjects factors of Valence and Character Type showed a significant Time x Valence interaction, $p = .029$ (Table 5), indicating positive feedback produced greater self-efficacy gains over time than negative feedback. The main effect of Time, Time x Character Type interaction, or the three-way interaction, were not significant. These results suggest that emotional valence, rather than delivery

Table 5: Mixed-Design ANOVA for Self-Efficacy by Valence and Character Type

Effect	df	F	p	η_p^2
Time	1, 36	2.84	.100	.073
Time × Valence	1, 36	5.18	.029	.126
Time × Character Type	1, 36	0.86	.360	.023
Time × Valence × Character Type	1, 36	0.38	.542	.010

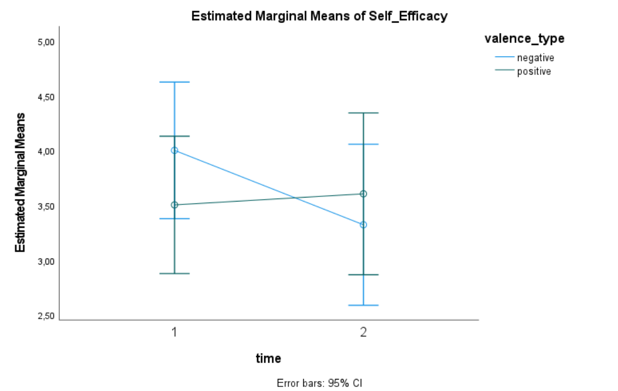

Figure 9: Pre and Post means on Self-Efficacy per valence type

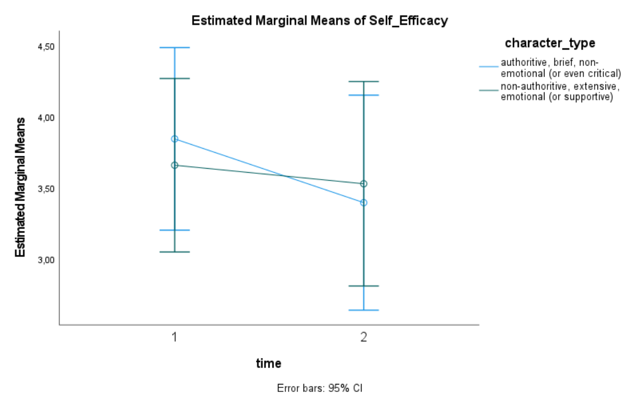

Figure 10: Pre and Post means on Self-Efficacy per character type

style, drives changes in self-efficacy.

In combination with figure 8, the figures 9, 10 can also visually confirm valence is a driver in changing SE.

5.2.2 *The impact of variants, valence and character type on perceived difficulty.* **Perceived Difficulty: 4-Group Mixed-Design ANOVA**

A mixed-design ANOVA with repeated measures on Time (Pre vs. Post) and a between-subjects factor of Feedback Variants revealed no significant main effect of Time, Feedback Variant or Time x Feedback Variant interaction on PD. Indicating neither the passage of time nor the specific feedback condition influenced participants' difficulty ratings. It can however be observed there is a visual trend differing the 4 variants on mean increase and decrease (figure 11).

Table 6: Mixed-Design ANOVA for Perceived Difficulty by Feedback Variant

Effect	df	F	p	η_p^2
Time	1, 36	0.005	.945	<.001
Feedback Variant	3, 36	0.75	.529	.059
Time × Variant	3, 36	0.77	.518	.060

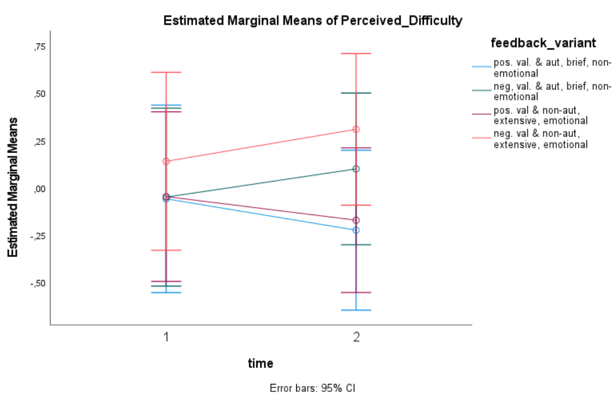

Figure 11: Pre and Post means on Perceived difficulty per variant

Perceived Difficulty: 2x2x2 Mixed-Design ANOVA
A mixed-design ANOVA with repeated measures on Time and between-subjects factors of Valence and Character Type likewise showed no significant main effect of Time, Time x Valence interaction, Time x Character Type interaction or a three-way interaction (Table 7). These results indicate neither the emotional tone (valence) nor the delivery style (character type) of feedback altered PD from pre- to post-feedback.

Even though for PD, valence does not show a *significant* effect, figure 12 compared to figure 13 shows a trend once more, that valence is a more deciding factor than character type.

*5.2.3 **The impact of variants, valence and character type on time spent**.* **Time on Task: 4-Group Mixed-Design ANOVA**
A one-way ANOVA comparing mean time spent on the assignment across the four feedback variants revealed no significant differences (Table 8). Mean times (in minutes) can be found in table 9. Tukey

Table 7: Mixed-Design ANOVA for Perceived Difficulty by Valence and Character Type

Effect	df	F	p	η_p^2
Time	1, 36	0.005	.945	<.001
Time × Valence	1, 36	2.30	.138	.060
Time × Character Type	1, 36	0.02	.880	.001
Time × Valence × Character Type	1, 36	0.003	.958	<.001

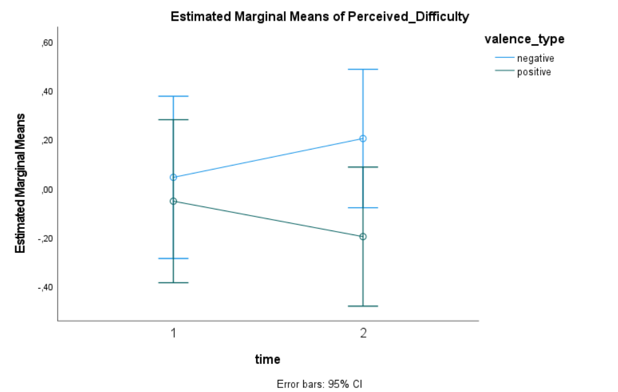

Figure 12: Pre and Post means on Perceived difficulty per valence type

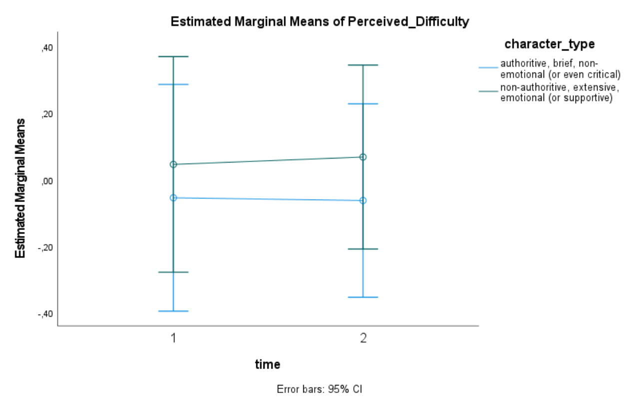

Figure 13: Pre and Post means on Perceived difficulty per character type

HSD post-hoc comparisons confirmed none of the pairwise contrasts reached significance (all $p > .69$), indicating that feedback variants did not affect how long students worked on the assignment.

Time on Task: 2x2 Between-Subjects ANOVA
A 2 (Valence) x 2 (Character Type) between-subjects ANOVA on time spent likewise gave no significant main or interaction effects. There was no main effect of Valence, Character Type or Valence

Table 8: One-Way ANOVA for Time on Task by Feedback Variant

Effect	df	F	p	η_p^2
Feedback Variant	3, 36	0.464	.709	.037

Table 9: Time on Task Means (minutes) by Feedback Variant

Variant	Mean Time (min)
V1	398.6
V2	704.7
V3	703.5
V4	773.7

x Character Type interaction (Table 10). These findings indicate that neither the emotional tone nor the delivery style of feedback influenced the amount of time students invested in the task.

Table 10: Between-Subjects ANOVA for Time on Task by Valence and Character Type

Effect	df	F	p	η_p^2
Valence	1, 35	0.245	.624	.007
Character Type	1, 35	1.099	.302	.030
Valence × Character Type	1, 35	0.142	.709	.004

The amount of time spent on the assignment can most likely be attributed to other factors, such as perhaps programming level. To examine this a Pearson correlation was computed. It indicated a small, non-significant positive relationship between the two, $r(38) = .20$, $p = .217$.

5.2.4 ***The impact of variants on attention paid to feedback***. From the literature it might be expected that when feedback tone conveys respect, support, and fairness, recipients feel empowered to act on it. Conversely, authoritarian tone is consistently linked to negative affect and reduced uptake of feedback. But a one-way ANOVA comparing mean attention paid to feedback across the four variants revealed no significant effect of Feedback Variant (Table 11. Descriptively, mean attention ratings (1-7 scale) can be found in table 12. Tukey HSD post-hoc comparisons showed no pairwise differences (all $p > .93$), indicating that feedback variants did not influence how much attention participants reported paying.

A 2 (Valence) x 2 (Character Type) between-subjects ANOVA on attention paid to feedback was conducted. There was no main effect of Valence, Character Type or Valence x Character interaction with significance (Table 13). Descriptive means (on a 1-7 scale) ranged

Table 11: One-Way ANOVA for Attention Paid by Feedback Variant

Effect	df	F	p
Feedback Variant	3, 36	0.14	.937

Table 12: Attention Paid Means (1–7) by Feedback Variant

Variant	Mean (SD)
V1	4.11 (2.20)
V2	4.70 (1.89)
V3	4.27 (1.74)
V4	4.30 (2.54)

from 4.11 to 4.70 across the four cells (Table 12), indicating that neither valence nor character type influenced how much attention participants reported paying.

Table 13: Between-Subjects ANOVA for Attention Paid by Valence and Character Type

Effect	df	F	p	η_p^2
Valence	1, 36	0.21	.647	.006
Character Type	1, 36	0.03	.859	.001
Valence × Character Type	1, 36	0.18	.676	.005

5.3 Qualitative Data Analysis

Participant comments and researcher observations also reveal several themes. Several participants identified nuanced ways in which feedback variants influenced their SE and PD. Participants in negative-valence conditions (V2,V4), such as P7, felt overwhelmed by complexity, reporting the suggestions were "too complex for me". Likewise, P9 and P14 noted extensive phrasing and repetition raised their cognitive load ("too much text to read," "repeated itself in different words"). Pointing to a tension between providing rich guidance and maintaining simplicity, particularly for novices who may depend on clear, bite-sized feedback [14].

P31, in the authoritative variant (V2), described feeling "frustration" at the directness, also stating that "this did not motivate me to continue the assignment", implying negative delivery may erode motivation. Participants that were exposed to a variant that was more extensive (V3, V4) also noted "the avatar uses a lot of extra words" (P23). Here P32 and P36 suggested they'd prefer it if text was presented in smaller steps.

Lastly, ending remarks were very indicative of which participants got a negative variant. Some simply disliking it, but one participant becoming defensive. P7 stated, "Soo angry, whyy" and P14 noted "the unhappy face made me feel a bit stressful." Both of these suggest the tone, and even the avatar's visual affect, amplified their irritation rather than supporting them. In contrast, participants with positive valence conditions rarely ended on negative statements.

Additionally, the assignment was perceived as a difficult assignment. P6 uttered he found it a hard exercise, which was resonated by a lot of the other participants with verbal remarks during the exercise.

Overall, the researcher notes and participant remarks seem to align with the quantitative data, while providing some additional context to the quantitative data.

6 Discussion

This study provided insight into how valence and other character elements influence SE and PD.

6.0.1 *Impact of Assignment Difficulty on Confidence.*

A key consideration in the design of this experiment is the impact of the assignment itself. Literature indicates that if the assignment does not align with the students knowledge level, instructional support tends to be more effective than feedback. For the experiment, the programming assignment was based on the assumption the teacher correctly assessed the student's skill level and provided an assignment accordingly. This study did not consider how assignment difficulty itself might affect participants' confidence. When an assignment's difficulty isn't well matched to a participant's skill level, the differences between feedback variants become less pronounced. This is reflected in the researchers notes, the assignment itself proved to be significantly more challenging than intended, which might have impacted the strength of the effect the feedback could have had.

6.0.2 *Prototype Fidelity and Implications for Digital Twins.*

The functionality of the prototype is a critical factor for generalizing the found information to digital twins as a whole. The current prototype is an approximation of a digital twin, it does not perfectly encompass all elements a complete digital twin would have. It is thereby important to nuance this experiment by stating that this is a current approximation of a digital twin, future more advanced technology, could do a better approximation (e.g. taking in more data) after which the effect of the measured variables might change.

6.0.3 *Participant Bias and Perceived Researcher Intent.*

Additionally, a consideration arises from a question posed by one of the participants: "Should I fill it in as if this avatar would always be there, when coding?" (P34), referring to the post experiment survey. Rather than solely reflecting on their immediate feelings while responding to the survey, the participant began to contemplate how they might feel in a hypothetical scenario beyond the the experimental setting, 'as if the avatar would always be there when coding'. This might suggest that the participant may have begun to discern the underlying aim of the research. It is therefore plausible that some participants, either consciously or unconsciously, adjusted their responses in alignment with what they perceived to be the researcher's intent. For the self-reporting tools such as the MSLQ, this is a sensitive element, impacting the results.

Here the design of the interface also returns as an element, since the feedback avatar is not an obligatory part of the interface, and people can ignore it, the data could be controlled for removal of participants that scored themselves low on 'how much attention' they paid for the feedback (scored<3). The statistical analysis was ran with this data excluded, but due to no differences, with minor mean changes on the results, and therefor no changed takeaways, it isn't reported.

6.0.4 *Technical Constraints and Future System Improvements.*

An additional technical limitation arises in the rate limits of the system, since the prototype was designed to handle a concurrent 60 users, the quality of the feedback depended on the max requests per minute and token count per minute the researcher was allowed to request from the openAI API. By having 60 concurrent users, the prototype was limited to sending along the current code without context of previous code and responses from the model. If this technical limitation was less prevalent the feedback might have been qualitatively higher and bear closer resemblance to the analytical and predictive elements of a digital twin. Additionally financial constraints also limit the use of the best possible model, and thereby a closer approximation of a digital twin. The usage of simply an LLM could be expanded upon with other additional technologies to better support the experience of a digital twin in an educational feedback providing setting.

6.0.5 *Methodological Considerations and Future Research Directions.*

The setup of the study, with measurements right before and after the experiment only measured short term effects and does not control for a novelty effect. A longer term study might reveal more insightful patterns or different results. For example whether sustained exposure to the real-time avatar alters confidence differently. Additionally, the context specific nature of the research, on a programming learning environment with a digital twin, limits the generalization of results to a broader spectrum on character design in digital twins.

Although the researcher adopted established instruments to establish participants' programming proficiency (demographics), these scales varied in response format (some 4-point, some 5-point). To make them directly comparable, they were standardized via z-scoring. However, this approach may obscure finer nuances: by compressing each scale to a common metric, subtle differences between categories could be lost.

Overall the choice of this explorative experiment to collect quantitative data allowed for less depth in participants responses, while for this topic a quantitative analysis might have provided more useful insights. Future research could benefit from considering mixed-methods approaches that can capture more detailed qualitative insights alongside quantitative data.

7 Conclusion

This paper investigated the impact of real-time written-visual feedback, delivered through anthropomorphic digital twin roles, on novel students' attitudes toward learning programming. The study specifically explored how different design roles, characterized by valence (positive vs. negative) and character type (authoritative, brevity, emotion), influenced SE and PD.

The findings confirm existing trends regarding positive and negative feedback on SE. A significant Time x Valence interaction (F(1,36)=5.18, p=.029) indicated positive valence feedback led to greater increases in SE over time compared to negative feedback. This suggests that for digital twins in programming education, the valence of feedback is a primary factor in learner confidence, while the specific character type had no significant impact on SE and PD changes.

However, the study provided limited insights into how the anthropomorphic roles of digital twin technology, as a whole, might influence PD. No significant effects were found for feedback variant, valence, or character type on PD. This suggests that while emotional tone can boost confidence, it may not be sufficient on its own to alleviate the PD of programming tasks for novice learners.

Future designs for digital twins in educational settings could integrate positive affect with instructional support to not only enhance SE but also address PD. The current prototype, while an approximation of a digital twin, faced technical constraints which restricted the richness and context of the feedback. This highlights the potential for more advanced technologies that can incorporate a broader range of real-time data and contextual understanding to provide even more nuanced and effective feedback. Longer-term studies are also necessary to observe sustained effects and control for novelty, as this research primarily focused on short-term impacts. Additionally, exploring mixed-methods approaches in future research could capture more detailed qualitative insights, complementing quantitative data for a more comprehensive understanding of the complex interplay between feedback character design, learner attitudes, and programming education.

8 Acknowledgments

Acknowledgments

The author wants to show his appreciation for his fellow students, with whom the design research process was a lot more manageable and enjoyable. But also his coach for providing the relevant insights when needed.

The author acknowledges the use of ChatGPT in the development of this project. ChatGPT was utilized to assist in writing portions of the code, debugging certain sections, providing valuable insights and suggestions on writing and during the development process. Its contributions helped streamline coding tasks and improve overall efficiency.

Gpt contributions:

- explanation of techniques and implementation used for high concurrent api usage (rate limiting, bottleneck)
- rewriting of ts files to js files, to use avr8js in my project (arduino compilation in browser)
- providing insight into the coherence of the paper, and suggesting improvements to improve readability flow
- Turning elements as tables into latex syntax (all contents were manually checked to ensure no data loss or hallucinations)
- Organizing statistical results into latex tables (The data was manually checked to ensure no hallucinations took place in the process)

References

- [1] Susan J. Ashford and L.L. Cummings. 1983. Feedback as an individual resource: Personal strategies of creating information. *Organizational Behavior and Human Performance* 32, 3 (Dec. 1983), 370–398. doi:10.1016/0030-5073(83)90156-3
- [2] William K. Balzer, Michael E. Doherty, and Raymond O'Connor. 1989. Effects of cognitive feedback on performance. *Psychological Bulletin* 106, 3 (Nov. 1989), 410–433. doi:10.1037/0033-2909.106.3.410
- [3] Klaus Zierer Benedikt Wisniewski1 and John Hattie. 2020. The Power of Feedback Revisited: A Meta-Analysis of Educational Feedback Research. *Frontiers in Psychology* 10 (Jan. 2020). doi:10.3389/fpsyg.2019.03087
- [4] Douglas Biber, Tatiana Nekrasova, and Brad Horn. 2011. THE EFFECTIVENESS OF FEEDBACK FOR L1-ENGLISH AND L2-WRITING DEVELOPMENT: a META-ANALYSIS. *ETS Research Report Series* 2011, 1 (June 2011). doi:10.1002/j.2333-8504.2011.tb02241.x
- [5] Jack W Brehm, Rex A Wright, Sheldon Solomon, Linda Silka, and Jeff Greenberg. 1983. Perceived difficulty, energization, and the magnitude of goal valence. *Journal of Experimental Social Psychology* 19, 1 (1983), 21–48. doi:10.1016/0022-1031(83)90003-3
- [6] Yuan-Chia Chang, Daniel J. Rea, and Takayuki Kanda. 2024. Investigating the Impact of Gender Stereotypes in Authority on Avatar Robots. *HRI '24: Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction* (March 2024), 106–115. doi:10.1145/3610977.3634985
- [7] Roy B. Clariana, Daren Wagner, and Lucia C. Roher Murphy. 2000. Applying a connectionist description of feedback timing. *Educational Technology Research and Development* 48, 3 (Sept. 2000), 5–22. doi:10.1007/bf02319855
- [8] Cornelia Connolly, Eamonn Murphy, and Sarah Moore. 2008. Programming Anxiety Amongst Computing Students—A key in the retention debate? *IEEE Transactions on Education* 52, 1 (Aug. 2008), 52–56. doi:10.1109/te.2008.917193
- [9] Edward L. Deci, Richard Koestner, and Richard M. Ryan. 1999. A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. *Psychological Bulletin* 125, 6 (Jan. 1999), 627–668. doi:10.1037/0033-2909.125.6.627
- [10] P. C. Earley, G. B. Northcraft, C. Lee, and T. R. Lituchy. 1990. IMPACT OF PROCESS AND OUTCOME FEEDBACK ON THE RELATION OF GOAL SETTING TO TASK PERFORMANCE. *Academy of Management Journal* 33, 1 (March 1990), 87–105. doi:10.2307/256353
- [11] Eindhoven University of Technology. 2025. Data Foundry: Platform for structured data collection and processing. https://data.id.tue.nl/. Accessed: 2025-06-09.
- [12] ElmoNeedsArson. 2024. M12_Entire_Structure: README. https://github.com/ElmoNeedsArson/M12_Entire_Structure/blob/main/README.md. Accessed: 2025-06-09.
- [13] Janet Feigenspan, Christian Kästner, Jörg Liebig, Sven Apel, and Stefan Hanenberg. 2012. Measuring programming experience. In *2012 20th IEEE International Conference on Program Comprehension (ICPC)*. 73–82. doi:10.1109/ICPC.2012.6240511
- [14] John Hattie and Helen Timperley. 2007. The power of feedback. *Review of Educational Research* 77, 1 (March 2007), 81–112. doi:10.3102/003465430298487
- [15] Milena Head, Khaled Hassanein, Edward Cho, and Michael Degroote. 2003. Establishing etrust through humanized website design. *Proceedings of 16th Bled ECommerce Conference* (01 2003).
- [16] P. Reddish R. Jesson R. Adams J. Parr, H. Timperley. 2007. *Literacy Professional Development Project: Identifying Effective Teaching and Professional Development Practices for Enhanced Student Learning.* https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=8b098c465ebf6533bd1febe488f0d8aa22c6be8b
- [17] Panadero E. Jonsson, A. 2018. *Facilitating students' active engagement with feedback.* Cambridge University Press.
- [18] Raymond W. Kulhavy. 1977. Feedback in written instruction. *Review of Educational Research* 47, 2 (Jan. 1977), 211. doi:10.2307/1170128
- [19] Raymond W. Kulhavy and William A. Stock. 1989. Feedback in written instruction: The place of response certitude. *Educational Psychology Review* 1, 4 (Dec. 1989), 279–308. doi:10.1007/bf01320096
- [20] James A. Kulik and Chen-Lin C. Kulik. 1988. Timing of feedback and verbal learning. *Review of Educational Research* 58, 1 (March 1988), 79–97. doi:10.3102/00346543058001079
- [21] Gabriella Lakatos, Márta Gácsi, Veronika Konok, Ildikó Brúder, Boróka Bereczky, Péter Korondi, and Ádám Miklósi. 2014. Emotion attribution to a Non-Humanoid robot in different social situations. *PLoS ONE* 9, 12 (Dec. 2014). doi:10.1371/journal.pone.0114207
- [22] Gary P. Latham and Edwin A. Locke. 1979. Goal setting—A motivational technique that works. *Organizational Dynamics* 8, 2 (Sept. 1979), 68–80. doi:10.1016/0090-2616(79)90032-9
- [23] Jolien M. Mouw Leonie Brummer, Hester de Boer and Jan-Willem Strijbos. 2024. A meta-analysis of the effects of context, content, and task factors of digitally delivered instructional feedback on learning performance. *Learning Environments Research* (June 2024). doi:10.1007/s10984-024-09501-4
- [24] Anastasiya A. Lipnevich and Ernesto Panadero. 2021. A review of feedback Models and Theories: Descriptions, definitions, and conclusions. *Frontiers in Education* 6 (Dec. 2021). doi:10.3389/feduc.2021.720195
- [25] Robyn L. Moffitt, Christine Padgett, and Rachel Grieve. 2019. Accessibility and emotionality of online assessment feedback: Using emoticons to enhance student perceptions of marker competence and warmth. *Computers Education* 143 (Aug. 2019), 103654. doi:10.1016/j.compedu.2019.103654
- [26] David Nicol and Debra Macfarlane. 2004. Rethinking Formative Assessment in HE: a theoretical model and seven principles of good feedback practice. *IEEE Personal Communications - IEEE Pers. Commun.* 31 (01 2004). https://www1.villanova.edu/dam/villanova/vital/vitalinks/7principlesofgoodfeedback04.pdf
- [27] OpenAI. 2025. GPT-4.1-nano Model Documentation. https://platform.openai.com/docs/models/gpt-4.1-nano. Accessed: 2025-06-09.
- [28] OpenAI. 2025. OpenAI API. https://openai.com/api/. Accessed: 2025-06-09.
- [29] Kerr P. 2020. *Giving feedback to language learners.* https://www.cambridge.org/gb/files/4415/8594/0876/Giving_Feedback_minipaper_ONLINE.pdf
- [30] Paul R. Pintrich. 1991. *A Manual for the Use of the Motivated Strategies for Learning Questionnaire (MSLQ).* http://files.eric.ed.gov/fulltext/ED338122.pdf
- [31] Allison M. Ryan and Paul R. Pintrich. 1997. "Should I ask for help?" The role of motivation and attitudes in adolescents' help seeking in math class. *Journal of Educational Psychology* 89, 2 (June 1997), 329–341. doi:10.1037/0022-0663.89.2.329
- [32] Michelle Scott, Lucas Pereira, and Ian Oakley. 2014. Show me or Tell Me: Designing avatars for feedback. *Interacting With Computers* 27, 4 (March 2014), 458–469. doi:10.1093/iwc/iwu008
- [33] C. Senior, James Barnes, R. Jenkins, Sabine Landau, M.L. Phillips, and Anthony David. 1999. Attribution of social dominance and maleness to schematic faces. *Social Behavior and Personality: an international journal* 27 (01 1999), 331–337. doi:10.2224/sbp.1999.27.4.331
- [34] Tal Shafir, Rachelle P. Tsachor, and Kathleen B. Welch. 2016. Emotion Regulation through Movement: Unique Sets of Movement Characteristics are Associated with and Enhance Basic Emotions. *Frontiers in Psychology* 6 (Jan. 2016). doi:10.3389/fpsyg.2015.02030
- [35] Uri Shaked. 2025. avr8js: AVR simulator in JavaScript. https://github.com/wokwi/avr8js. Accessed: 2025-06-09.
- [36] David H. Smith, Qiang Hao, Filip Jagodzinski, Yan Liu, and Vishal Gupta. 2019. Quantifying the Effects of Prior Knowledge in Entry-Level Programming Courses *(CompEd '19)*. Association for Computing Machinery, New York, NY, USA, 30–36. doi:10.1145/3300115.3309503
- [37] SurveyJS. 2025. SurveyJS: The UI Component Library for Surveys and Forms. https://surveyjs.io/. Accessed: 2025-06-09.
- [38] S. Thirakunkovit and Pisarn Chamcharatsri. 2019. A meta-analysis of effectiveness of teacher and peer feedback: Implications for writing instructions and research. *Asian EFL Journal* 21 (01 2019), 140–170.
- [39] W B Thompson. 1998. Metamemory accuracy: effects of feedback and the stability of individual differences. *The American Journal of Psychology* 111, 1 (Jan. 1998), 33. doi:10.2307/1423535
- [40] Fanny Vainionpää, Marianne Kinnula, Atte Kinnula, Kari Kuutti, and Simo Hosio. 2022. HCI and Digital Twins – A Critical Look: A Literature Review. In *Proceedings of the 25th International Academic Mindtrek Conference* (Tampere, Finland) *(Academic Mindtrek '22)*. Association for Computing Machinery, New York, NY, USA, 75–88. doi:10.1145/3569219.3569376
- [41] M. Conner W. M. Rodgers and T. C. Murray. 2008. Distinguishing among perceived control, perceived difficulty, and self-efficacy as determinants of intentions and behaviours. *British Journal of Social Psychology* 47, 4 (01 2008), 607–630. doi:10.1348/014466607x248903
- [42] Ze Wang, Huifang Mao, Yexin Jessica Li, and Fan Liu. 2016. Smile big or not? Effects of smile intensity on perceptions of warmth and competence. *Journal of Consumer Research* (Oct. 2016). doi:10.1093/jcr/ucw062
- [43] Liang Zeng. 2009. Designing the User Interface: Strategies for Effective Human-Computer Interaction (5th Edition) by B. Shneiderman and C. Plaisant. *International Journal of Human-Computer Interaction* 25, 7 (Sept. 2009), 707–708. doi:10.1080/10447310903187949

9 Appendices

A Demographics Questionnaire

(1) Did you take computer science related courses in university? [36]
- I took one CS related course in university
- I took more than one CS related courses in university
- Never

(2) Did you take computer science courses in high school? [36]
- I took one CS course at high school
- I took more than one CS courses at high school
- Never

(3) How much time did you spend on self-learning & teaching of programming outside of school? [36]
- A lot of time
- Sometime
- Little time
- No time

(4) How would you like to describe yourself in terms of programming? [36]
- Have a lot of experience in programming
- Have some experience in programming
- Have very limited experience in programming
- Have no experience in programming

(5) How confident are you in your ability of programming? [36]
- Very confident
- Somewhat confident
- Not confident at all

(6) How do you estimate your programming experience compared to experts with 20 years of practical experience? [13] (1 = very inexperienced, 5 = very experienced)
(7) How do you estimate your programming experience compared to your classmates? [13] (1 = very inexperienced, 5 = very experienced)
(8) How experienced are you with the following language: Java [13] (1 = very inexperienced, 5 = very experienced)
(9) How many additional languages do you know (medium experience or better)? [13]
(10) How experienced are you with the following programming paradigms: object-oriented programming? [13] (1 = very inexperienced, 5 = very experienced)
(11) How experienced are you with the following programming paradigms: functional programming? [13] (1 = very inexperienced, 5 = very experienced)
(12) For how many years have you been programming? [13]
(13) If applicable, for how many years have you been programming for larger software projects, e.g., in a company? (fill in 0 if not applicable) [13]

B Self-efficacy and Perceived difficulty Questionnaire

B.1 Self-efficacy

[30]

(1) I believe I will receive an excellent grade in this course. (1 = Not at all true of me, 7 = Very true of me)
(2) I'm certain I can understand the most difficult coding material presented in the survey exercise I am about to do. (1 = Not at all true of me, 7 = Very true of me)
(3) I'm confident I can understand the basic concepts taught in this course. (1 = Not at all true of me, 7 = Very true of me)
(4) I'm confident I can understand the most complex coding material presented in the survey exercise I am about to do. (1 = Not at all true of me, 7 = Very true of me)
(5) I'm confident I can do an excellent job on the programming exercise in this survey. (1 = Not at all true of me, 7 = Very true of me)
(6) I expect to do well in this course. (1 = Not at all true of me, 7 = Very true of me)
(7) Considering the difficulty of this course, the teacher, and my skills, I think I will do well in the survey's programming exercise. (1 = Not at all true of me, 7 = Very true of me)

B.2 Perceived Difficulty

(1) How difficult do programming problems appear to be for you in general? (1 = not at all difficult, 5 = very difficult) [5]
(2) How much would completing a programming assignment on arduino delays() and millis() be easy to do for you? (1 = very little control/extremely difficult/not at all confident, 9 = complete control/extremely easy/completely confident) [41]
(3) How much effort do you think it will take for you to successfully complete the problem on this topic? (1 = very little effort, 5 = more effort than 1 could possibly give) [5]

Appendix C. Prototype Description Diagram

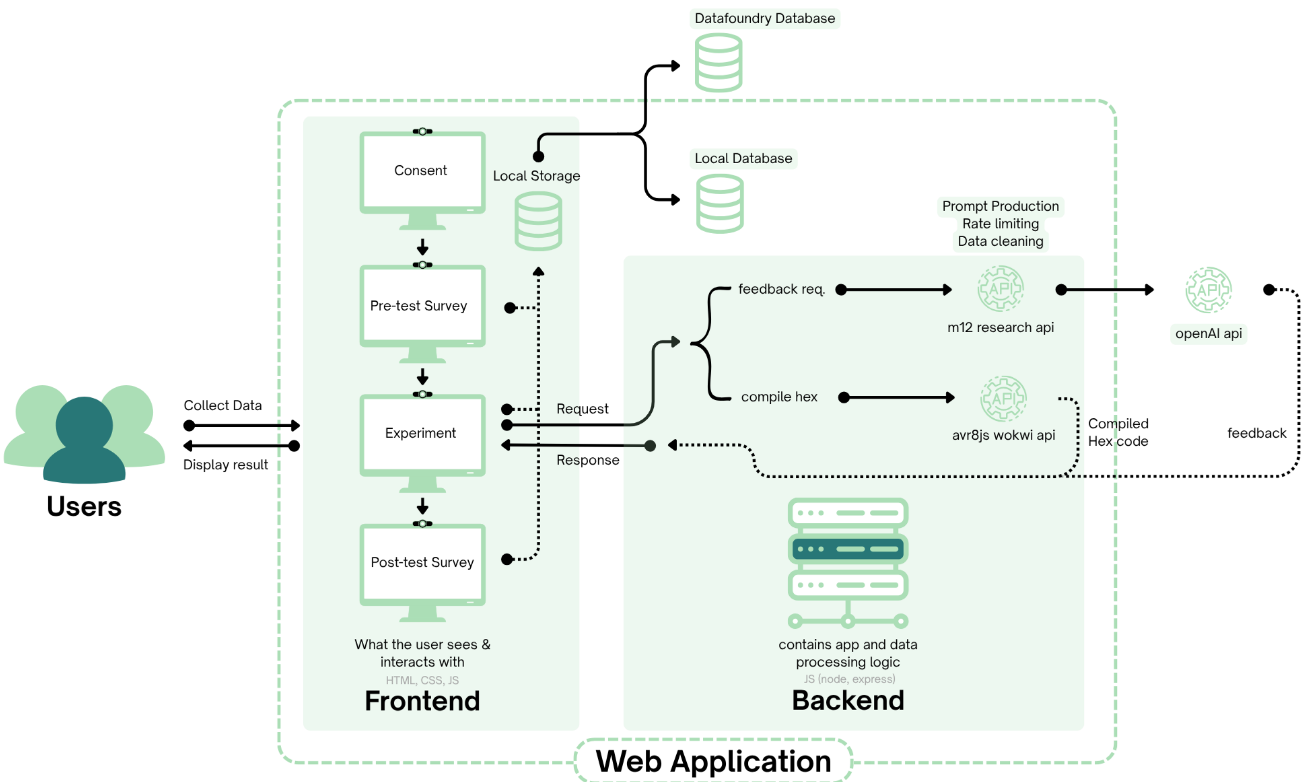

Figure 14: Extremely high level oversight of prototype structure

Appendix D. Code Repository

The entire code repository can be found on GitHub. This repository has been stripped of submitted data, API keys and internal domains called to access the relevant API's. The readme file contains some information on how the platform was hosted using which structure.

For the design of the frontend interface the surveyJS library [37] was used to facilitate the design of the surveys. For the programming interface the avr8js library [35] was used, this JavaScript library supports AVR 8-bit architecture, which is relevant when wanting to run/compile Arduino code in the browser.

Link: github.com/EntireStructureM12

Appendix E. Prompt

```
System:
You are one of four different feedback variants. A big variable is the valence of the
feedback. You are in this case ${valence} valence.
The other variable ties into your character. You are in this case a ${character}
character who guides first-year students step by step, especially around millis() vs
delay().
These character traits and valence go above all other instructions you are given, you
must remain in character.
Based on these properties, come up with a character, and play this role convincingly.

The Student's assignment:
Rewrite the Arduino blink example so LEDs on pins 13, 12, and 11 blink at 700 ms, 900
ms, and 1300 ms intervals using millis().
Each LED must have its own control_PIN_NR_led() function called from void loop(). With a
bool as such "static bool ledStateNR = LOW;"

Example end solution:
void controlBlueLed(){
   static unsigned long lastBlueLedToggle = 0;
   static boolean blueLedStatus = LOW;

   if(millis() - lastBlueLedToggle > 700){
     lastBlueLedToggle = millis();
     blueLedStatus = !blueLedStatus;
     digitalWrite(11, blueLedStatus);
   }
}

The student has currently written the following code:
\`\`\`cpp
${cleanStudentCode}
\`\`\`

Constraints:
- Favor simplicity over efficiency — no arrays/objects.
- The students are beginner level; use only basic functions.

Identify which of the task's micro-steps (removing delay(), three pin setup, timing
logic, state toggle) the student hasn't nailed yet,
call out any errors you see there, and then offer exactly one next action to get them
unstuck.

Feedback style:
- Maximum 1-2 sentences, or 3 brief sentences. (Your character may overrule this if
needed)
- Point out exactly what concept to apply (Removing, adding or altering code).
- You must use examples or pseudocode to illustrate your point.
- You must avoid answering the full code in your reply.
- End with "Your next step: …"

Once again make sure your answer is in character. You must adress the user, by saying
'you'. Lastly make it clear when the assignment is roughly passed, and seize giving
feedback.`
```

```
model: "gpt-4.1-nano"
messages: [{ role: "user", content: prompt }],
```

Figure 15: Prompt used in prototype

F Reflection: Did I do it Right? (Still not sure)

Supervised by Regina Bernhaupt, I undertook a design research project this semester. I developed a tool, using research-through-design to explore how avatar design in digital twins affects programmers' confidence and perceived difficulty.

In my reflection of M11, realized how much I sought affirmation and reassurance. I see myself as an independent person, but I noticed that in unfamiliar environments, I tend to ask many questions to the coach, not just to confirm I'm doing *enough*, but also doing *good*. Knowing coach involvement affects grading, tends to inhibit my ability to make choices without always confirming that it is 'right'. This semester, a main goal for me was to address this realization. Regina's approach on guiding students supported this: she mainly gives feedback if you ask for it, assuming you're doing fine otherwise. This gave me control over how much confirmation I sought. With some exceptions, like at the demo day when I told her I missed some confirmation/critique on my project, I think I mostly managed to hold back from seeking too much confirmation.

But what did this actually bring me? I recognize that I made mistakes, or choices I might have made differently had I requested confirmation continuously. In other projects, my work still contained mistakes, but somehow felt less significant because I wasn't the only one responsible for the direction. Now, those missteps felt more personal, making me more disappointed in myself when things didn't work out as intended. Working with less confirmation, I noticed that I became more aware of areas where I lack experience or perspective (e.g. study design). I recognize there is a healthy balance between asking for support and making decisions independently. For this project, I might have leaned too far into self-reliance. Going forward, I want to find a better balance, seeking input not out of insecurity, but out of curiosity and a desire to learn. One way to do this in future projects is by reflecting in the moment ***why*** I'm reaching out, am I trying to reduce doubt, or genuinely explore a new angle?

F.1 Challenging my comfort zone

Another goal this semester was to broaden the way I collect data. In previous design and research projects, I mostly used qualitative methods or a small-scale mixed approach, typically max 16 participants. While suitable for in-depth qualitative work, I noticed I often leaned into it for convenience sampling. Not because the analysis is easier, qualitative data is still a lot of work, but because the process of recruiting participants felt less daunting or risky in those cases. So when my coach suggested 15 participants for each of my variants (n=60), it seemed like a good connection to my goal, and challenge of my comfort zone. I addressed this goal by contacting a teacher of a bachelor course with my target group and designing around that. During this process, I realized that quantitative data collection was less scary than I expected.

However, this thought didn't last. The data collection didn't work out, besides the teacher misjudging class size, many students started but didn't finish. This highlights one of the biggest challenges in online quantitative data collection, the participants have no real reason to do your research. Even with external rewards and relevance to their course, students didn't care. While fair, I entirely underestimated this element.

I eventually fell back on purposive and snowball sampling within the ID department, but this still feels more like a convenience solution rather than a good execution of a quantitative study. I find it hard to quantify what exactly I learned from this goal. I am happy I undertook it, because familiarizing your self with a new situation is always of value. It wasn't the perfect situation I had hoped for, but it did make me realize some obvious challenges in quantitative data collection to be aware of in future projects. This gives me as a designer more knowledge in conducting such research in future research-through-design processes.

On a more technical note, this also addressed one of my goals for the semester: becoming more comfortable with backend systems, APIs, and basic security principles. This tied directly into my quantitative data collection, as it enabled me to build a secure and publicly accessible data collection environment. While less visible than the recruitment and analysis process, this technical work was essential in making the research possible. I see this as an important step in expanding my design practice, not just conceptually, but also practically. The ability to build digital tools that are functional, secure, and user-ready is useful for me as a designer, especially if I take on more research-through-design projects that involve real user data.

F.2 Additional Statements

This challenge on quantitative data collection highlighted another reflection point. In my goals during this semester I limited the amount of goals I had on academic research itself. Having done research before, I underestimated how much I could still learn. One major challenge, that I underestimated, was narrowing down a topic within digital twins, framing a research question, and doing so within time constraints. In my previous project I had prepared and thought about my chosen topic and research question for a while before I pursued it, this different process this semester was an entirely different experience, and was admittedly a bit of a struggle to define a concise but still interesting research question.

The topic of this research, digital twins, helped me articulate something I'd been feeling for a while. Tying into my professional identity, and perhaps explains a struggle with creating my vision. During this project I identified for myself that there's always a choice: do you design for the near future or the far future? Digital twins are a topic that really drive this question forward for me, this made me remember a conversation I had with Steven Houben when I was deciding who to choose as a coach, and he highlighted a question in designing for improving the current educational system, or designing a new educational system. That conversation, in combination with doing this project made it clear: I prefer to work within existing systems.

I think this originates from a need that I find it important for designers to actually practically engage with their designs. Not for it to stay hypothetical. That you learn the options and limitations are of what you are working with. Designing for the near future means collaborating with people and building on what already exists, which makes it easier to have meaningful, grounded impact. I do however want to critique myself here, I should be mindful that my preference for practicality does not limit imagination.

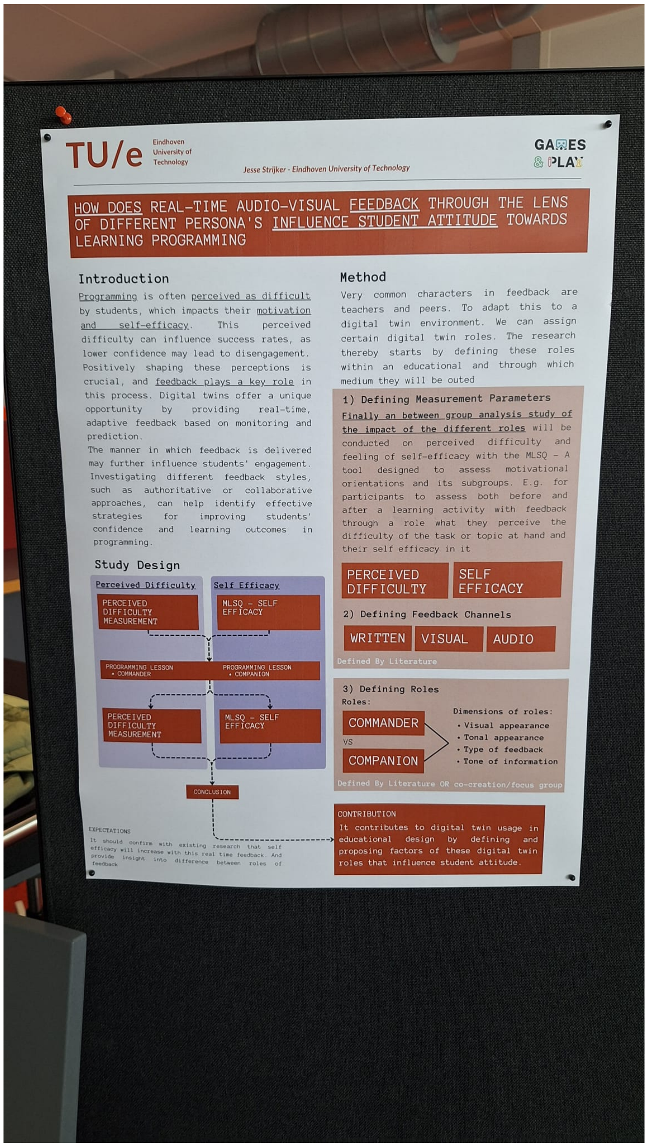

Figure 17: Picture during midterm of midterm poster

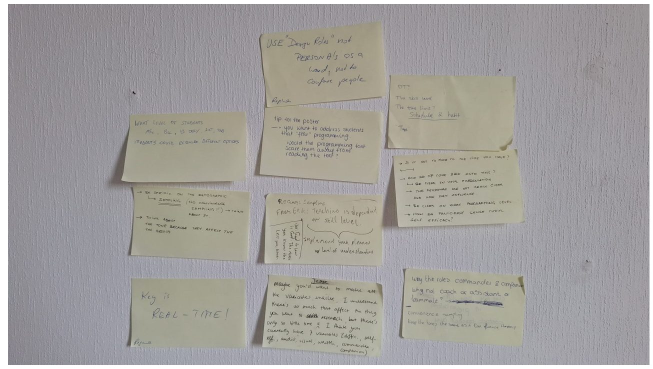

Figure 18: Received feedback on midterm

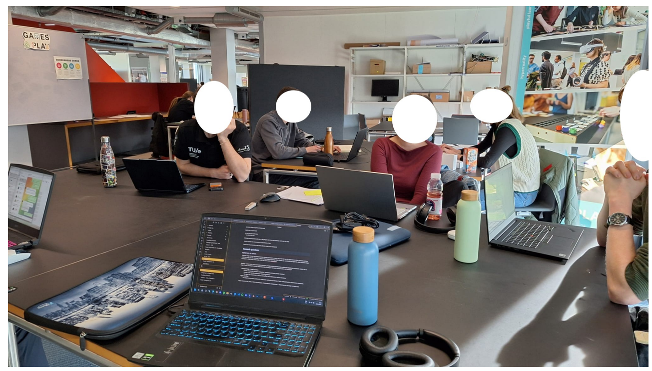

Figure 16: Colloquiums we organized with all M12 students to talk about issues we ran into together and help each other out