Your Text Is Hard to Read: Facilitating Readability Awareness to Support Writing Proficiency in Text Production

Jakob Karolus
jakob.karolus@dfki.de
German Research Center for Artificial Intelligence (DFKI)
Kaiserslautern, Germany
RPTU Kaiserslautern-Landau
Kaiserslautern, Germany

Sebastian S. Feger
sebastian.feger@ifi.lmu.de
LMU Munich
Munich, Germany

Albrecht Schmidt
albrecht.schmidt@ifi.lmu.de
LMU Munich
Munich, Germany

Paweł W. Woźniak
pawel.wozniak@chalmers.se
Chalmers University of Technology
Sweden

ABSTRACT

Allowing users of interactive systems to reflect on their task proficiency is often incidental. This is unfortunate, as communicating meaningful task-related proficiency feedback could improve users' awareness of their abilities and their willingness to improve. To highlight the feasibility of this concept, we evaluated how different methods of readability feedback impacted users during a text production task. In general, our results showed that having access to readability feedback allowed participants to reflect on their task solving approach, facilitating the users' understanding of their proficiency. Revision-based methods are less distracting for the user than continuous feedback methods, while still offering high efficacy. Further, feedback should be paired with a subtle form of gamification elements. We envision this reflection-oriented design to user proficiency to be applicable to a variety of interactive systems, allowing for an improved and engaging user experience.

CCS CONCEPTS

• **Human-centered computing** → **Interactive systems and tools**.

KEYWORDS

writing proficiency, text production, feedback

ACM Reference Format:
Jakob Karolus, Sebastian S. Feger, Albrecht Schmidt, and Paweł W. Woźniak. 2023. Your Text Is Hard to Read: Facilitating Readability Awareness to Support Writing Proficiency in Text Production. In *Designing Interactive Systems Conference (DIS '23), July 10–14, 2023, Pittsburgh, PA, USA.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3563657.3596052

DIS '23, July 10–14, 2023, Pittsburgh, PA, USA

ACM ISBN 978-1-4503-9893-0/23/07.
https://doi.org/10.1145/3563657.3596052

1 INTRODUCTION

Computers have transformed the way we work, study, and interact with information every day. Interactive systems are designed to support our tasks and goals, be it at work, on the move, or at home. Digital writing assistants are a prime example of such support tools. They correct our spelling and grammar, and even tailor essays to desired audiences. All available at our fingertips, for stationary and mobile input devices, whenever we want. Yet, users rarely have the opportunity to reflect on the corrected mistakes, inhibiting situational awareness and thus preventing them from exploring their own performance [59]. Usually, users just accept proposed changes, or do not even note auto-corrected interventions, taking tool-mediated support for granted. Consequently, users may miss the opportunity to improve their writing proficiency. We argue that relying on the *intelligence* of such systems can potentially prove detrimental to a user's proficiency development through deskilling. Allowing users to strive for excellence—if desired—should be a key design goal to create engaging experiences [51] with interactive systems.

It remains a challenge to understand how we can design systems that encourage users to advance their own understanding of a language and their proficiency in writing. As most users will not be able to determine whether improving their proficiency is a worthwhile investment of cognitive effort [36], they are unlikely to willingly spend resources on trying to learn a more efficient way of accomplishing a task. Hence, the goal for such systems should be to facilitate the user's understanding of their own proficiency, supporting them in recognizing and reflecting on their performance. In doing so, such systems can effectively help users reflect on their work, which is a widely recognized design goal in HCI [6]. This is line with a long tradition of reflective practice being a key element of effective learning [59]. Challenging the user to improve their proficiency is a crucial aspect of these systems. As such, it challenges the notion that assistance in assistive systems should be minimally noticeable [33].

Malacria et al. [40] found that feedback on one's performance allows users to rethink their current task solving performance and switch to a different, faster approach. In this paper, we focus on text production as a prominent example of everyday computer

work and showcase how systems can communicate proficiency aspects, focusing on how and when to provide feedback. Contrary to *Skillometers* [40], the main focus of our investigation is not on enhancing the overall performance of users for a given task, but rather facilitating the users' understanding of task-related skills, e.g., being able to tailor one's writing for a specific audience. In other words, we study the impact of underlying skills that contribute to the final performance. This separation is crucial, as it allows self-contained systems to focus on individual skills contributing to a user's proficiency and conceptualizes how we can effectively raise the users' awareness for individual skills [33].

In our concrete example, we investigate how to communicate writing proficiency to users. More specifically, how users can tailor their text readability to a specific audience, an important skill in text production. Consequently, users can focus on the writing task at hand, while the system simultaneously provides them with the means to be aware of the text difficulty of their writing. To evaluate whether users are encouraged to alter their writing approach, we conducted two online studies in which participants were tasked with producing text of a fixed length while ensuring a certain readability rating, making it accessible to the given audience. In Study I, we first explore if feedback on proficiency in the form of estimated reading difficulty can elicit self-reflection and improve performance. We compare different methods of communicating this feedback, continuous vs. revision-based. Having confirmed that real-time proficiency feedback offered benefits, we investigate how to design better feedback in Study II. We study how to limit the negative impact of interruptions through feedback by means of gamification.

Our results showed that task-related proficiency feedback enabled users to complete the writing task requirements more precisely. Importantly, we found that users consciously adjusted their approach in solving the task based on the feedback presented, which suggests that they were aware of their proficiency. While continuous assessment allowed for precise task fulfillment, users also reported that this type of feedback interrupted and distracted them during the text production task. A revision-based feedback approach was perceived as less distracting while still achieving better-than-baseline task fulfillment. Adding gamification elements only marginally influenced user experience. The results suggest that subtle interface elements such as progress bars are most suited to allow users to reflect on their own task proficiency.

In this paper, we contribute an investigation on designing task-related proficiency feedback in a text production task in the form of two online studies. Based on our results, we conclude that text readability feedback elicits self-reflection on one's current task solving approach, facilitating an understanding of one's own writing proficiency. Our work concludes with insights on designing proficiency feedback for future interactive systems, allowing for an engaging user experience.

2 BACKGROUND AND RELATED WORK

The HCI field has an established tradition of designing and studying adaptive systems. Yet, adapting (system changes the interface) or customizing (the user changes the interface themselves) user interfaces may not always be beneficial for the user. Predictability and accuracy of the adaptation play a vital role in user acceptance [23]. Additionally, users tend to be reluctant to undergo customization, as it requires time and cognitive effort while immediate benefits are not visible [37, 39]: the "paradox of the active user" [10].

It still remains an open research area how to create *effective* and *efficient* adaptation on a holistic level. While most often, achieving high performance and good user experience go hand in hand, there exist cases where achieving one is contradictory to the other ("Ubiquitous human computing" [68]). Consequently, the right choice of adapting to, i.e. supporting, the user is crucial for success. We argue that to reach this goal, adaptations towards the user could possibly be withheld and delayed; contrarily to the performance-orientated credo of classical adaptive systems.

An active research area looks at novices to expert transition [14] with regard to alternative expert interfacing options, such as hotkey usage. *Blur* [58] is a system that realizes this switch using subtle notifications, hence first making users aware of an alternative interfacing method. A less subtle method is presented by Bateman et al. [4] with their *search dashboard* comparing users to archetypal profiles to allow users to reflect on their own skills and alter their search behavior. Hence, it is vital to understand what aspects drives users to customize [2] or wanting to adapt the interface [20] during task solving. This implies that we need to investigate *when* the user chooses to adapt their current strategy and *what* makes them rethink their current approach. For our work, we consequently employ different types of proficiency feedback that vary in temporality and noticeability (see Section 3) and evaluate their effects on the users' willingness to change their task solving approach. For the design, we draw from related work on feedback, its impact on interrupting the user and its potential for self-reflection.

2.1 Feedback, Interruptions, and Reflection

Reflection has been a recurring theme in HCI. Schön's [59] work on the nature of reflection has been particularly influential [6]. He distinguishes between reflection-in-action and reflection-on-action. Reflection-in-action happens when performing a task and noticing unexpected outcomes. Reflection-on-action is retrospective. Past research determined that revising data generated by oneself is an effective strategy for fostering reflection [7]. While most systems focus on reflection in a holistic interpretation, focused on daily life patterns [56], wellbeing [1] or crowdsourced tasks [18], our work uses a more atomic approach to reflection. We explore how systems can actively support reflection-in-action through making the consequences of the users' actions easily visible, i.e., allowing them to reflect on their actions while performing the task. Specifically, we investigate how to support this kind of reflection through interface elements during a text production task. As mentioned earlier, varying temporality of the shown feedback (*Revision* vs *Continuous* in Section 3) allows us to determine the potential of reflection-in-action while compromising on the level of interruptiveness.

Balancing feedback and interruption is a known dilemma in interface design. Frequent [57] interruptions are decremental [30], yet research argues that without interruptions and associated focus shifts [8], there can be no opening for learning and no opportunity to improve proficiency. Malacria et al. [40] have highlighted a similar situation in their work *Skillometers*, where feedback needed

to be visible to catch the user's eye while simultaneously being subtle to minimize disruption. These past approaches are rooted in a wider discourse on reflection based learning, outlined in Schön's seminal work [59]. Research in the learning sciences discussed how a reflection-based approach to learning was rooted in classical philosophical thinking and could be used by teachers to stimulate students to think critically [16]. Chang advocated developing new instruction methods specifically for supporting reflection [12]. Within HCI, reflection has been recognised as a key learning strategy [66]. Roldan et al. [52] highlighted the need to develop new kinds of prompts which can stimulate students to reflect. Our work is inspired by and expands on these concepts by following their solution (manipulating the locus of control) and providing proficiency feedback via an ambient display element. Additionally, we purposely interrupt the user (revision-based method) after finishing their task to reflect on their current performance. Thus, our work aims to structure the understanding of the design constraints involved in building proficiency-aware systems for text production.

2.2 Writing Assistance

Since the occurrence of the first automated spell checkers [49], writing assistance systems have steadily improved. Nowadays digital writing assistants[1] are prevalent on nearly any computing device, continuously reviewing our spelling, grammar, and punctuation. Even fully automated text generation (GPT-3 [9]) and programming via natural language input[2] is possible.

However, writing remains a complex tasks with lots of factors influencing the final quality of the produced text, such as ease of understanding, thematic coverage, or creative elements. To aid the process of writing, research has investigated new ways of providing tailored feedback covering relevant subtasks, such as topic identification [55], improving writing styles [38], and on-the-fly text assessment for sensitive commentary [48]. Machine-in-the-loop approaches towards creative writing [13] do not necessarily provide better results, but are effective as a supporting tool [24]. Interestingly, feedback that motivates and engages users can spark meaningful effort investments [34]. In our work, we draw from this idea of creating effort-provoking feedback. We want to allow users to engage with their writing, to reflect on it. We put emphasis on how computing system can support the process, allowing for co-creation where the system monitors a specific aspect of writing (in this work: text difficulty), allowing for subtle feedback (see Section 2.1) if necessary.

2.3 Gamification: Requirements and Opportunities

Gamification, "the use of game design elements in non-game contexts" [17], has proven to motivate and support learning in formal [3, 64] and informal settings [45]. In HCI, gamification is one of the dominant approaches to fostering motivation and fostering a positive user experience in providing feedback using a variety of forms and strategies [46]. Badges, points, and leaderboards are the most common game design elements [28, 60]. However, recent work emphasized the need to review a wider set of game elements and their contexts of use [47, 65].

Our work commits to this exploration, investigating whether or not, and in which form, gamification can support effective proficiency feedback in text production. Here, our design choices were informed by several theories. The first one is *Flow* [43, 44]. A person finding themselves in a *flow* state is fully immersed in an activity which they consider enjoyable and fulfilling. Up to nine dimensions are commonly described which contribute to a flow experience [27, 43]. Of those dimensions, challenge-skill-balance, clear goals, and feedback are directly compatible with proficiency feedback. Second, Self-Determination Theory (SDT) [53] is concerned with the interplay between extrinsic motivation (e.g. rewards, fear of punishment) and intrinsic motivation, i.e. motivation created and sustained within the self by curiosity, interest, or identified values. The Basic Psychological Needs Theory (BPNT), one of the six SDT mini theories, focuses on three basic needs that promote wellbeing: *competence*, *autonomy*, and *relatedness*. By providing feedback that enables users to assess and improve performance, we expect to stimulate competence and autonomy directly. Finally, the Organismic Integration Theory (OIT) is concerned with various forms of extrinsic motivation [54]. OIT refers to a spectrum of internalization of values represented by a task or environment. These theories helped us select game design elements and evaluation metrics used in our work. This paper investigates how these theories can be applied to understand the experience of users completing text production tasks. Further, we study how gamification can inform the design of proficiency feedback in text production.

2.4 Summary and Research Questions

We hypothesize that such proficiency-aware systems should confer the benefits of undergoing a change of task approach and modality, allowing users to understand that the increased effort will pay off. In this work, we envision self-reflection on proficiency (user skill in writing text at a given difficulty rating) as a means to facilitate a user's understanding of their own writing proficiency.

Informed by related work, we investigate the impact of task-related proficiency feedback on the user's performance and experience. In particular, we evaluate whether feedback on one specific aspect of one's writing (estimated readability of a written proposal) can elicit self-reflection in participants, allowing them to better understand their own writing proficiency. Here, we draw from existing works on the balance between interruptive and reflective feedback. We further refine the design of proficiency feedback through gamification elements, aiming to alleviate its adverse effects. To operationalize this investigation, we formulated three research questions. In short, these questions ask if, when and how feedback on one's writing proficiency can be effective:

RQ1: *Can task-related proficiency feedback in text production facilitate an understanding of one's own proficiency?* Task-related proficiency feedback should assist the user in improving their skill level by facilitating an understanding for their current (lack of) proficiency. It would not directly address task performance, but rather assess the underlying skill set necessary to complete this task. We address this research question by analyzing whether users

[1] E.g.: Grammarly (https://www.grammarly.com/)
[2] https://openai.com/blog/gpt-3-apps/

change their task solving strategy, i.d., altering their text proposal, after being presented with feedback.

Research has postulated that the willingness to customize or adapt one's strategy during task solving might be linked to task proficiency. Power users adapt more willingly because of their understanding of the benefits of adapting strategies [41]. Consequently, we specifically selected an everyday task of expressing an argument in a submitted proposal to limit the impact of prior proficiency levels.

RQ2: *When should task-related proficiency feedback be presented to users?* We implemented a continuous feedback option that is updated whenever the user changes the text. While this increases the user's awareness of the functionality (see Section 2.1), it might distract them from the primary task, yet conversely, breakdowns and associated focus shifts "can be openings for learning" [8]. To strike a balance, we additionally implemented a revision-based method that presented the user with a proficiency evaluation upon completing their initial proposal and let them reflect on their current performance. This distinction allows us to study the timing of the feedback as a design consideration for proficiency feedback.

RQ3: *Can gamification improve task-related proficiency feedback?* Feedback can distract or even annoy users. This contrasts with the feed for feedback being perceived as supportive and useful. Given the potential of gamification (see Section 2.3) to communicate information, status, and achievements in an enjoyable and motivating manner, we explored game design elements for proficiency feedback: progress bars, social comparison, and emojis. In our work we study not only how feedback affects performance in a text production task, but also the impact of feedback forms on motivation and perceived distraction.

3 METHOD

We based our investigation on two studies as listed in Table 1. The full 3 *x* 3 design looks at the independent variables *Feedback Type* and *Gamification Type*, both of which have three levels. Study I first looks at the impact of *Feedback Type* in isolation (**RQ1, RQ2**), while Study II focuses additionally on the gamification part (**RQ3**).

Table 1: Study design and respective conditions for the full 3 *x* 3 design including the independent variables *Feedback Type* (rows) and *Gamification Type* (columns).

Feedback Type \ Gamification Type	*No Gamification*	*Progress Bar*	*Emoji*
No Feedback	Study I	N/A	N/A
Revision	Study I and II	Study II	Study II
Continuous	Study I and II	Study II	Study II

Participants were asked to express an argument in a proposal using a submission form on a web page. Note that web views for all conditions of Study I and II as well as the source code for the web form are available in the supplementary materials. Please refer to the "web_form" and "web_form_conditions" folders. Filling out forms on the web is a mundane task and users are familiar with the environment. Further, producing text to express our opinion happens daily, be it in emails, essays or articles. Improving the user experience and secondary benefits (increasing proficiency) for these tasks can thus be contributory on a large scale. Ultimately, broad adoption can foster data-driven algorithms enabling sophisticated recommender systems [42]. For a definition and metadiscourse on proficiency, we refer the readers to work by Karolus and Wozniak [33]: "**Proficiency** is the aggregated construct of any skills, knowledge, competence, or experience of a person relevant to the interaction between the person and a system (the task domain)."

We want to highlight that task-related proficiency feedback **is not direct feedback on task performance** — writing a good proposal — but rather **assesses and communicates the necessary skill** — here: writing in plain English — of the user to complete the primary task in the first place (cf. [33]). For this purpose, we implemented a system that assesses a user's submitted text proposal in term of readability. Proficiency feedback is provided as an additional display element next to the form (see Figure 1 for an example). We leverage the Flesch reading-ease score (FRES) [22] which indicates how difficult it is to understand a given text. It is based on average-sentence length (ASL) and average number of syllables per word (ASW). The resulting Flesch reading-ease score can be computed as follows:

$$FRES(ASL, ASW) = 206.835 - 1.015 * (ASL) - 84.6 * (ASW) \quad (1)$$

A score between 60 and 70 is interpreted as plain English. Texts with higher scores are easier to read but can be too simplistic. A lower score indicates a more difficult text. The score is widely used for evaluating a text's readability, e.g., the state of Florida requires insurance policies to have a FRES of at least 45 [3].

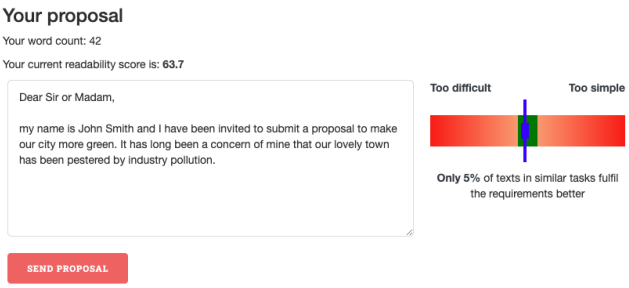

Figure 1: Screenshot of our proficiency-aware web form, showing the *Continuous x Progress Bar* condition, i.d., continuously displaying the current readability score as well as providing a gamification element in the form of a progress bar which includes a social comparison aspect (see Section 5). Views for all conditions of Study I and II as well as the source code for the web form are available in the supplementary materials. Please refer to the "web_form" and "web_form_conditions" folders.

[3] http://www.leg.state.fl.us/statutes/index.cfm?App_mode=Display_Statute&Search_String=&URL=0600-0699/0627/Sections/0627.4145.html

Participants were tasked with writing a proposal to make the neighborhood more green in plain English (defined by a FRES between 60 and 70), which will be presented to the city council; a task taken from a preparation course for the Cambridge English Advanced exam [26]. Participants will be judged by their ability to write a concise proposal in plain English. The complete task description is as follows:

> *You see this announcement on the notice board of your local council.*
> *The Environmental Planning Committee is organizing a campaign to make our town more 'green'. You are invited to submit a proposal related to your neighborhood, which will be presented to the city council. Present some factual information about the area, pointing out any relevant environmental issues, and suggest practical measures which individuals and families could take to make the neighborhood more green.*
> ***Write your proposal in 220-260 words in plain English. Your proposal should be readable by a broad audience.***

For both studies, task, participant recruitment and procedure remained consistent apart from adapted questionnaires. Since our participant recruitment options were limited due to the pandemic situation at the time of writing, we opted to recruit participants via the Amazon Mechanical Turk Service (MTurk). While this limited our possibilities in obtaining more information about the participants' experiences, e.g. through qualitative post-task assessment, it allowed us to collect a large number of samples. To ensure high data quality, we carefully screened all submitted results for completion and proposal quality (see Section 4 and Section 5).

4 STUDY I - TEMPORAL GRANULARITY OF PROFICIENCY FEEDBACK

In this first evaluation, we addressed the temporal granularity of the provided proficiency feedback (see Table 1). We specifically looked at *Revision*-based and *Continuous* feedback methods. Together with a *No Feedback* condition, this one-factorial design resulted in three levels.

Revision-Based. After submitting the proposal once, the user was informed about their last readability score. Upon resubmission, the user was able to make changes to achieve a better result if they deemed this necessary. During this revision process, the last score was always visible. Participants were instructed accordingly:

> *Your proposal will be automatically scored in terms of reading difficulty after submission. You will have the option to revise your proposal afterwards once. You should aim for a score between 60 and 70. Higher scores indicate a too simplistic proposal, while lower scores indicate low readability.*

Continuous. In this condition, participants' readability scores were calculated at fixed time intervals (2 seconds) and displayed. The score was always visible above the text entry field. Figure 1 shows the *Continuous* feedback for a gamified condition. Again, participants were instructed accordingly:

> *Your proposal will be continuously scored in terms of reading difficulty. You should aim for a score between 60 and 70. Higher scores indicate a too simplistic proposal, while lower scores indicate low readability.*

No Feedback. As a baseline we added a *No Feedback* condition, where users were not given any feedback on their readability score throughout the whole writing process. Consequently, no additions to the task description were made.

4.1 Hypotheses

With regard to **RQ1** and **RQ2**, we formulated three hypotheses:

H1a: *Continuous task-related proficiency feedback leads to more accurate task fulfillment.* We hypothesize that the more users are confronted with feedback about their text readability, the better they can solve the given task. In our case, task fulfillment was measured two-fold: (1) being able to submit an adequate proposal and (2) getting one's readability score close to the target zone. We analyzed final readability scores of valid proposals and looked at the change of readability over time.

H2a: *Revision-based task-related proficiency feedback is less distracting.* We know from related work that constant feedback can be decremental (see Section 2). Hence, we hypothesize that more subtle and less constant feedback is less distracting for users. We measured this with tailored questions.

H3a: *Task-related proficiency feedback alters the task approach of users.* Proficiency feedback should incentivize users to improve their inherent skill set. It is thus important to investigate if users alter their task solving approach to adhere to task constraints. We analyzed this aspect with tailored questions as well as investigating changes in writing behavior (readability over time).

4.2 Participants

After screening for uncompleted or inadequate proposals, we accepted a total of $N = 70$ submitted proposals from participants. Additionally, three researchers independently graded each accepted proposal to assess their quality. The criteria for assessment were ease of understanding and relevance to the task. The average grade was a B with a standard deviation of one grade[4]. A high inter-rater agreement ($r_{WG(J)} > .99$ [31]) of the three researchers confirmed high consensus. This additional grading process not only ensured high quality data for later analysis but also confirmed the suitability of our initial screening procedure. Participants were reimbursed with $2 and offered an additional a $1 bonus for an acceptable proposal text. This rate was approved by the institution of the first author (equivalent of $12/hour). Out of these participants, 19 resided in the European Economic Area, 18 in Canada and 33 in the USA. All participants were informed that study participation was voluntary, that the study could be aborted at any point and that the data would be collected in anonymized form. It took approximately 15 minutes to complete the survey, including writing the proposal text. The average age of the participants was $M = 35.4\,y$ ($SD = 8.9\,y$)

[4]US grading system: A (best) to F (worst).

with 29% identifying as female, 71% as male. Additionally we asked all participants to provide their self-assessed writing skills [15] (see supplementary material). After the final submission of the proposal text, we again asked them to evaluate their writing quality based on an adapted scale [32] (see supplementary material).

The writing assessment tests helped us to assess whether participants exhibited the necessary skills to complete the writing task (see Section 3). We found that participants rated themselves highly proficient in this writing task: $M = 13.4$ ($SD = 1.8$) for the writing self assessment (max score: 15) and $M = 65.4$ ($SD = 11.0$) for the writing quality (max score: 80) of their proposal. This confirms that all participants were sufficiently proficient to execute the given writing task. Additionally, the experimenters evaluated each proposal for correctness and adequacy.

4.3 Procedure

After providing informed consent, participants were asked to provide demographics and assess their writing skills. They were then randomly assigned to one of the three conditions and were given the task to write a proposal for the city council in which they explained their ideas to make the city greener. Depending on condition, they were provided with *No Feedback*, *Revision*-based or *Continuous* feedback of their current readability score. Participants were made aware of these conditions, by telling them that their proposal would be scored and how often this would be the case.

After completing their proposal, participants were asked to fill a raw NASA-TLX [29], assess their writing quality and asked custom questions tailored at their perception of the scoring system, including accuracy and perceived disruptiveness (see Table 2). Note that the custom questions were only present for conditions *Revision* and *Continuous*.

Table 2: Additional questions for conditions *Revision* and *Continuous*; from *strongly disagree* to *strongly agree*; all visual analog scale (0 to 100).

Perception of the scoring system	
Q1a	I felt that my performance was accurately assessed.
Q2a	I felt pressured by the scoring system.
Q3a	I performed better using the scoring system.
Q4a	The system interrupted me during the task.
Q5a	I could have done the task without the scoring system.

4.4 Results

For the final dataset, conditions were distributed as follows: *No Feedback*: 24, *Revision*: 24 and *Continuous*: 22 entries. We report our analysis on the following metrics as collected in Study I: task completion time (TCT), NASA-TLX, the final Flesch reading-ease score (FRES) and its deviation from the target zone (60 to 70). Additionally, we take a look at the FRES over time for each condition and analyze in particular whether the *Revision*-based method has prompted participants to alter their proposal. Further, we analyzed the grades given for the proposals (see Section 4.2). Lastly, our custom questions gave insights into the disruptiveness of each feedback method. If not stated otherwise, we conducted one-way ANOVAs to analyze the data. If normality was violated, we first aligned rank transformed [67] the data. All tests (if necessary) were adjusted for multiple comparison using the Tukey method. Effect sizes are given using η^2 (Partial Eta Squared): small ($> .01$), medium ($> .06$), large ($> .14$). Analysis scripts and raw data are available in the supplementary material. Please refer to the "scripts" folder.

4.4.1 Task Completion Time, NASA-TLX, Proposal Grades and Final Readability Score. We did not observe any significant difference between the conditions (*No Feedback*, *Revision*, *Continuous*) for task completion time, nor for the NASA-TLX scores or the final FRES. Deviation from the target zone (measured in absolute deviation from 65) was also not significant. Descriptive statistics are available in the supplementary material. Further, we could confirm that the feedback conditions had no impact on proposal quality, as graded by the researchers.

4.4.2 Influence of Readability Feedback on Task-Solving Approach. To further evaluate the influence of the feedback methods on the participants' readability scores, we compared the temporal course of each condition. Since task completion times varied across participants, we rescaled all trials to the median task completion time of 758 *s*. This allowed us to visually compare the different trials at once and draw conclusions based on the mean and the standard deviation of the readability score over time. An overview of this analysis is shown in Figure 2. The red line marks the average (at any given point in time) over all valid[5] trials by participants, while the red corridor marks the standard deviation. Additionally, for *Revision*, the thick black vertical line show the mean point in time when the participants started the revision of their text. Note that the FRES is highly volatile for short texts during creation. Hence, we omitted the first 100 seconds for these plots.

Figure 2 illustrates that participants in the *Continuous* condition narrowed down faster on the target zone. The *No Feedback* condition was worst in this regard. The final deviation over all participants per condition was lowest for *Continuous*. *Revision* and *No Feedback* conditions exhibited the highest variances for the final readability score.

We additionally evaluated whether the revision prompt in the condition *Revision* had an impact on participants' writing behavior. To do so, we fitted a linear model for the collected FRES data after the revision prompt. We fitted the model with the averaged data after the mean revision prompt time (thick black line in Figure 2). We then tested against a null model that simulated no change in writing behavior, in other words: no incline. We found a significant difference ($F(1, 41.7) = 346.0, p < .001, \eta^2 = .52$) between the models with a large effect size. This indicates that participants did indeed try to improve their readability score after the first revision.

4.4.3 Custom Questions on System Accuracy and Interruptiveness. The analysis of our custom questions (see Table 2) revealed a significant difference (large effect) for Q4a: "The system interrupted me during the task.": $F(1, 44) = 9.92$, $p < .01$, $\eta^2 = .18$. All other

[5] A researcher cross-referenced each submitted proposal with the respective time series for the readability score. Trials were omitted if it was evident that the submitted text was just copied into the form.

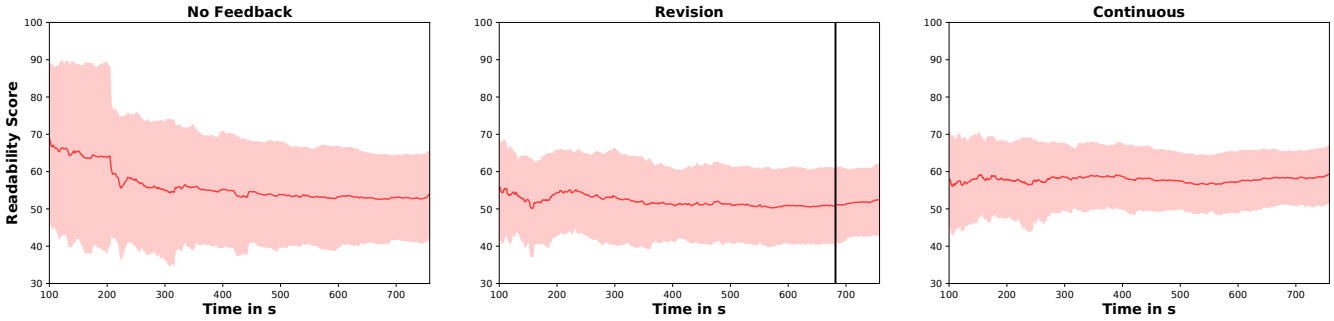

Figure 2: Deviation of readability score over time. All trials have been rescaled to the median answer time and start at $t = 100\,s$. The mean over all trials is displayed by the red line, the red-shaded corridor shows the standard deviation. Condition *Revision* (middle) additionally shows the timing for the mean revision prompt (thick black vertical line). It can be observed that both *Revision* and *Continuous* depict less overall variance (than *No Feedback*). Additionally, *Revision* shows a clear incline after the revision prompt (see Section 4.4.2). High quality vector graphics are available in the supplementary material, additionally depicting individual trials by participants. Please refer to the "tables_graphs" folder.

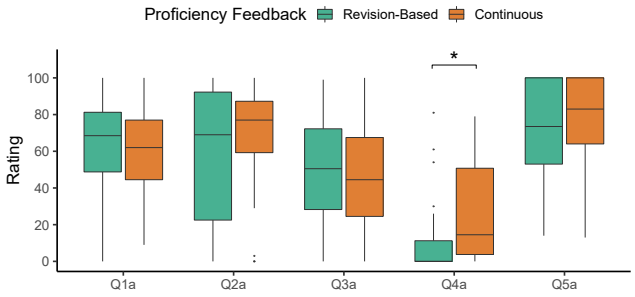

Figure 3: Ratings for questions Q1a-Q5a (see Table 2). Q4a ("The system interrupted me during the task.") shows a significant difference between *Revision* and *Continuous* (marked with *).

questions showed no significant differences. Note that for this analysis only the conditions *Revision* and *Continuous* are present. An overview is provided in Figure 3, additionally showing slightly higher ratings for Q5a and average ratings for Q3a, indicating that participants believed that they had performed adequately even without the system. Q2a shows that users felt more pressured in the *Continuous* condition, while both conditions were rated as sufficiently accurate (Q1a).

4.5 Summary

Our initial evaluation of different proficiency feedback methods has raised some noteworthy facets regarding the design qualities of task-related proficiency feedback. The following summary discusses these first results and highlights aspects that informed the design of Study II.

When users were confronted with a system that continuously assesses their proficiency, it allowed them to reach a target with higher precision, partly confirming **H1a**. This effect was not only present for the final readability score, but also allowed them to narrow down on the target zone more quickly compared to other assessment types (**H3a**). This advantage in performance did not impair task completion times, as there was no significant difference over the different assessment types. Interestingly though, participants in the *Revision* condition tended to narrow down on the target zone as quickly without having had access to their readability score yet. Thus, it remains to be investigated why users in the *No Feedback* condition converged less quickly and were less accurate. One possible effect could be increased extrinsic motivation by participants in *Revision*. Here, participants were informed that they were being assessed by a scoring system. Conversely, participants in *No Feedback* were unaware of this fact and were not put under pressure. Once participants started to revise in *Revision*, a definitive effort by participants could be observed to close in on the target zone, confirming **H3a**.

While participants reported no differences in perceived workload, it was evident from our custom questions that the *Continuous* feedback method caused significantly more disruptions (Q1a) than the *Revision*-based method, confirming **H2a**, though both conditions pressured users to a fair degree (Q2a). The *Revision*-based feedback seemed to cause split reactions from participants in this regard.

Ultimately, *Continuous* feedback offered the best performance in terms of hitting a target readability score at the cost of pressuring users into the need to perform (**H1a**). Here, a revision-based system can produce relief if precision is not essential (**H2a**).

Recognizing that proficiency feedback increases perceived stress and interruption, we aimed to further explore feedback modalities that might remedy those issues. In Study II, we investigated whether gamification elements could make an impact through enjoyable game design components.

5 STUDY II - DESIGNING BETTER PROFICIENCY FEEDBACK THROUGH GAMIFICATION

Based on the findings of Study I, we further investigated effects and perceptions around different forms of continuous and revision-based task-related proficiency feedback. In Study II, we explored

gamification as a means to mitigate perceived pressure and interruption reported by participants in Study I. In particular, we designed and evaluated contrasting gamification components for both continuous and revision-based feedback.

The taxonomy of gamification elements, described by Robinson and Bellotti [50], provided a good reference for our additional designs and conditions in Study II. They presented 42 gamification elements coded according to their minimum level (i.e. low, medium, high, and variable) of engagement required. The authors further categorized those elements into six top level categories. We chose to focus on two categories. First, *"Feedback and Status Information"* provided a highly relevant overview of game elements that directly support the purpose of our study. To contrast those rather rational and subtle game design elements, we further chose to explore the design elements which support *"Intrinsic Motivation"*, as we wanted to understand the impact of joyful and interesting gamification elements on users' perceptions of additional feedback information. In the end, we implemented two gamified views, shown in Figure 4: *Emoji* and *Progress Bar*. Together with a *No Gamification* condition as baseline, this resulted in a two-factorial design including the factors *Feedback Type* (*Revision*, *Continuous*) and *Gamification Type*, yielding a total of six conditions (see Section 3). We note that those feedback modalities do not represent a full systematic exploration of applicable gameful components. Rather, they are designed to contribute an early exploration of how different metaphors impact performance.

Emojis. The left side of Figure 4 shows the *Emoji* implementation that indicates compliance with the text production task through five unicode emojis, ranging from sad to happy. This design and implementation is reflected in the *Entertainment* element ("Simple elements can work very quickly in low commitment settings"), as described in the *Intrinsic Motivation* group by Robinson and Bellotti [50].

Progress Bar. The second gamification implementation, shown on right side of Figure 4, relates to the element *Graphical Indicators* ("Easy to design, and in fact critical for all gamification."), as described by Robinson and Bellotti in the *Feedback and Status Information* group [50]. We chose to implement a progress bar with a center target range (green zone), as it represents an easy-to-understand visualization of the tasks and provides a form of *progress feedback*, which Tondello et al. [65] referred to as a gameful design element. Having a sense of progress is important in stimulating the basic psychological need *competence*. In addition, we added a statement that places the users' current proposal quality into context with proposals written "in similar tasks". This statement is simulated and based on the current readability score. For example, at a readability score of 49, the statement indicates that "**26%** of texts in similar tasks fulfill the requirements better because **your text is too difficult to read.**", while at a score of 63, the message is more positive: "**Only 5%** of texts in similar tasks fulfill the requirements better". The purpose of this ranking, or text-based leaderboard, is to create a form of perceived competition, also referred to as gameful design element "social comparison" by Tondello et al. [65], nudging participants into better performance, which should work even in those situations where they already hit the inner edges of the green target range on the progress bar. For consistency and simplicity, we refer to this gamified view as *Progress Bar* view.

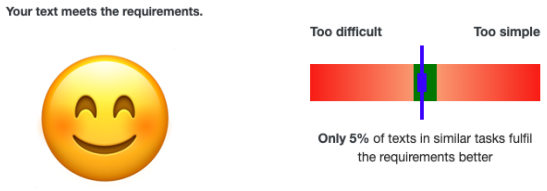

Figure 4: Gamification elements of our proficiency-aware web form: *Emoji* (left) and *Progress Bar* (right) including the social comparison statement (percentages are simulated). These display elements are added on the right side of the original web form (see Figure 1).

5.1 Hypotheses

We reused our hypotheses from Study I (**H1b, H2b, H3b**) and explicitly describe our hypothesis regarding the use of gamification elements in task-related proficiency feedback in **H4**.

H4: *Gamification elements alleviate distractions induced by task-related proficiency feedback.* We know from Study I that proficiency feedback can be perceived as pressuring and distracting. We hypothesize that the joyful nature of gamified applications can turn proficiency feedback into an enjoyable experience that is perceived as an enrichment, rather than a distraction. To this end, we implemented two gamified views for both the *Revision*-based and *Continuous* feedback conditions as well as adapted our custom questions accordingly. We further employed several subscales of the Situational Motivation Scale (SIMS) [25] and the Intrinsic Motivation Inventory (IMI) [11], to assess impact on intrinsic motivation and different forms of extrinsic regulation.

5.2 Participants

Participant recruitment was analog to Study I over the Amazon Mechanical Turk Service (MTurk), additionally aiming for a higher number of participants. Ensuring the same selection criteria for accepted proposals, a total of $N = 147$ data records were used for analysis. As we already confirmed the suitability of our initial screening process, we decided not to grade the proposals for Study II. Reimbursement and information provided to prospective participants was identical to Study I. Out of these participants, 40 resided in the European Economic Area, 42 in Canada, 3 in Australia, 1 in Asia and 61 in the USA. It took approximately 15 minutes to complete the survey, including writing the proposal text. The average age of participants was $M = 35.9\,y$ ($SD = 9.85\,y$) with 46% identifying as female, 54% as male. We asked all participants to provide their self-assessed writing skills [15], but decided to omit the self-assessed writing quality of their proposal and the NASA-TLX in Study II, to streamline the procedure.

Similarly to Study I, we found that participants rated themselves highly proficient in this writing task: $M = 13.5$ ($SD = 1.6$) for the writing self assessment[6], again confirming that all participants were

[6] Maximum score: 15.

sufficiently proficient to execute the given writing task. Additionally, the experimenters evaluated each proposal for correctness and adequacy.

5.3 Procedure

Following the procedure of Study I, participants were asked to provide their demographics and to assess their writing skills after providing informed consent. They were then randomly assigned to one of the six conditions and were again given the same writing task as in Study I (see Section 4.3) and likewise informed about their condition.

After completing the writing task, participants completed the SIMS [25] and IMI scales [11]. We included all subscales of SIMS: Instrinsic motivation, Identified regulation, External regulation and Amotivation. For IMI we included three subscales[7]: Perceived competence, Effort/Importance and Pressure/Tension. Each subscale was scored on a 7-item Likert scale and included four to six items that were averaged. We again concluded with a set of custom questions. Although similar to those in the first study, we focused more on the aspects of interruptions and distractions by the scoring system. We also explicitly tailored to a change in task approach. Table 3 shows the final set of questions.

Table 3: Additional questions for all conditions; from *strongly disagree* to *strongly agree*; all visual analog scale (0 to 100).

	Perception of the scoring system
Q1b	I felt pressured by the system.
Q2b	I felt that my performance was accurately assessed by the system.
Q3b	I adapted my approach in solving the task due to the system.
Q4b	The system enabled me to complete the task accurately.
Q5b	The system interrupted me during the task.
Q6b	The system helped me to see how well I was doing.
Q7b	The system distracted me during the task.

5.4 Results

We report our analysis on the following metrics as collected in Study II. Inferential statistics are conducted analogously to Study I. The distribution of data entries over conditions is given in Table 4.

Table 4: Distribution of collected data entries over all conditions in Study II.

	No Gamification	*Progress Bar*	*Emoji*
Revision	25	20	22
Continuous	23	31	26

5.4.1 Task Completion Time and Final Readability Score. We did not observe any significant difference between the conditions for task completion time (TCT) in Study II. TCT was highest for *Continuous x No Gamification* at $M = 1280\,s$ ($SD = 787\,s$) and lowest for *Revision x No Gamification* at $M = 810\,s$ ($SD = 496\,s$). For the final readability score, we found significant differences with a large effect size for the factor *Feedback Type* ($F(1, 141) = 30.0, p < .001, \eta^2 = .17$). Figure 5 shows the final scores for the two factors.

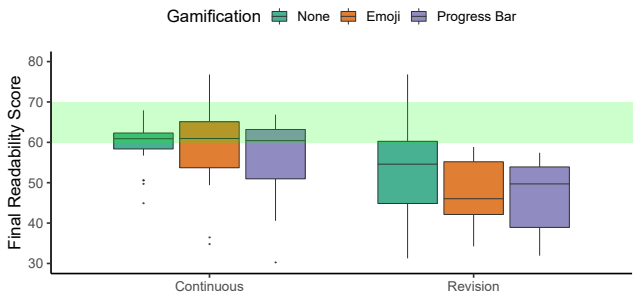

Figure 5: Final readability scores given both factors: *Feedback Type* and *Gamification Type*. Scores for *Feedback Type* (*Revision - Continuous*) are significantly different. The shaded green area marks the target zone for the readability score (60 to 70).

5.4.2 IMI and SIMS Scales. For IMI [11] and SIMS [25] scales we found significant differences (medium effect) for intrinsic motivation ($F(2, 141) = 5.36, p < .01, \eta^2 = .07$) for *Gamification Type*. Post-hoc pairwise comparison showed a significantly lower score for intrinsic motivation for *Emoji* compared to *No Gamification* and *Progress Bar*. Further, we found significant differences (small effect) for identified regulation ($F(2, 141) = 4.05, p < .05, \eta^2 = .05$) for *Gamification Type*. Again, post-hoc tests showed a significantly lower score for identified regulation for *Emoji* compared to *No Gamification*. We did not find any significant differences for external regulation and amotivation as well as for any of the IMI subscales (Perceived competence, Effort/Importance, Pressure/Tension) that we employed. Please refer to the supplementary material for a graphical representation.

5.4.3 Influence of Readability Feedback on Task-Solving Approach. Similar to Study I, we analyzed the readability scores over time. We applied the same data transformation (see Section 4.4.2), adapting the median task completion time to $927\,s$. An overview of this analysis is shown in Figure 6. It illustrates again that participants in the *Continuous* conditions narrowed down faster on the target zone and also exhibited less variance during this process and at the end (see Figure 5). Compared to Study I, the difference to *Revision*-based conditions is more pronounced. Participants in the *Progress Bar* conditions also narrowed down on the target zone in a more linear fashion. The other conditions exhibited a more ad-hoc adaption.

We could additionally confirm the same effect for *Revision*-based feedback conditions as in Study I. Participants did again try to improve their readability score after the first revision. We performed the same statistical analysis for each revision-based condition and could confirm statistical significance. Further, we fitted a model with aggregated data over all *Revision*-based conditions. Tests against the null model confirmed a significant difference with a large effect size ($F(1, 74.7) = 8298.9, p < .001, \eta^2 = .97$).

[7]The other subscale were redundant with SIMS or not of interest for this study.

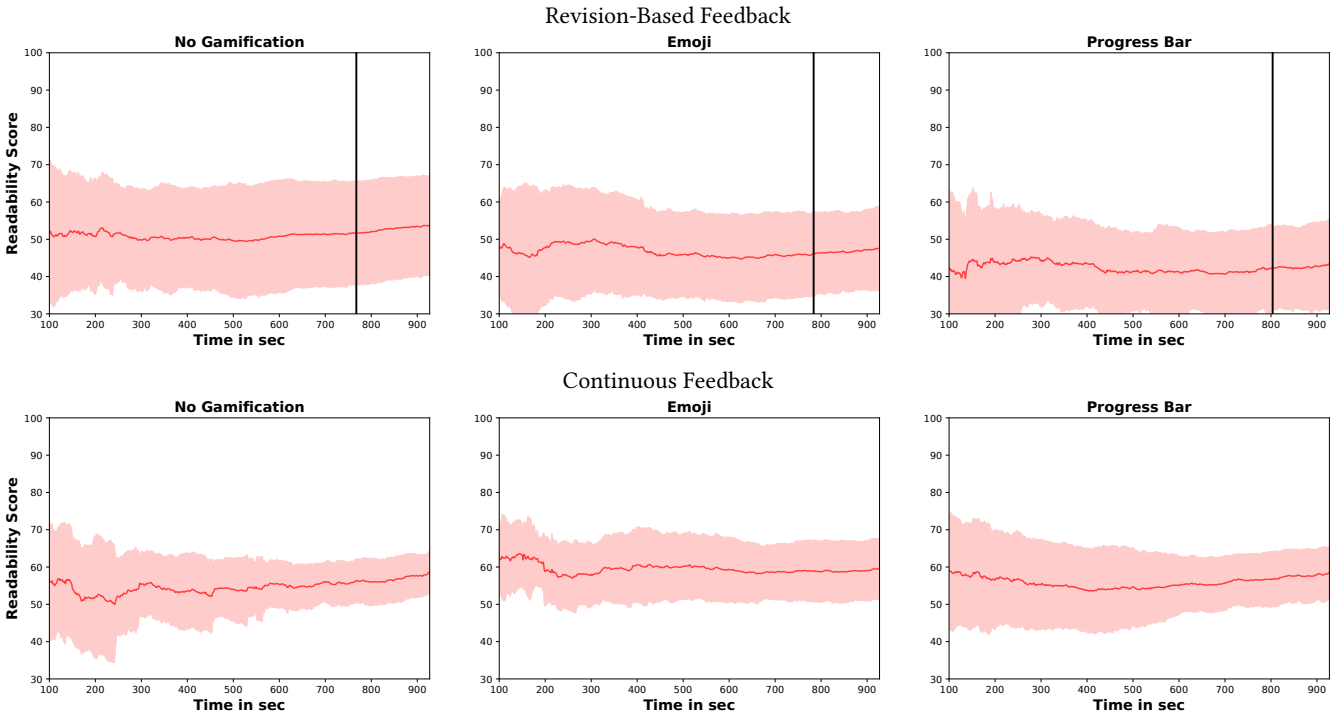

Figure 6: Deviation of readability score over time. All trials have been rescaled to the median answer time and start at $t = 100\,s$. The mean over all trials is displayed by the red line, the red-shaded corridor shows the standard deviation. Conditions based on *Revision* feedback (top row) additionally show timings for the mean revision prompt (thick black vertical lines). It can be observed that all *Continuous* conditions depict considerably less overall variance (than *Revision*), being closer to the target zone (see Section 5.4.1). Additionally, all *Revision* conditions show a clear incline after the revision prompt (see Section 5.4.3). High quality vector graphics are available in the supplementary material, additionally depicting individual trials by participants. Please refer to the "tables_graphs" folder.

5.4.4 Additional Questions on System Accuracy, Interruptiveness and Changes in Task Solving. Our analysis revealed significant differences for *Feedback Type* for Q3b: "I adapted my approach in solving the task due to the system." ($F(1, 141) = 5.24, p < .05, \eta^2 = .04$), Q5b: "The system interrupted me during the task." ($F(1, 141) = 13.38, p < .001, \eta^2 = .09$), Q6b: "The system helped me to see how well I was doing." ($F(1, 141) = 10.74, p < .01, \eta^2 = .07$), Q7b: "The system distracted me during the task." ($F(1, 141) = 26.75, p < .001, \eta^2 = .16$) with varying effect sizes. For Q7b we additionally found a significant effect of *Gamification Type*, though no post-hoc significances between levels were present. All other questions showed no significant differences. An overview is provided in Figure 7 showing high rating (more prominent for *Continuous*) for Q3b, Q4b and Q6b polling the interaction between the participants' task approach and the assistance offered by the system. Scores on Q1b, Q5b and Q7b (pressure and interruptiveness) are more split between the conditions, while Q2b scores just above the midpoint range, polling accuracy of the scoring system.

5.5 Summary

Similarly to the results of Study I, we observed that *Continuous* feedback allowed participants to reach the target zone with higher precision. The bigger sample size of Study II confirmed that this effect was significant compared to *Revision*-based methods, fully confirming **H1b**. Additionally, Study II provides evidence that *Continuous* methods allowed users to narrow down on the target zone more quickly (**H3b**), showcasing that users adapted their task solving approach in the presence of the scoring system. For *Revision*-based methods, we once again confirmed that once participants started their revision, a definitive effort could be observed to close in on the target zone, confirming **H3b**.

Our updated custom questions allowed us to take a closer look at the disruptiveness of the system and how users perceived its assistance. Here, we could confirm that *Continuous* methods were significantly more distracting and interrupted the user (**H2b**). While both feedback method prompted users to adapt their task solving approach, the impact from the *Revision*-based method was significantly lower (**H3b**).

Our results confirmed, in the *Continuous* feedback conditions, that gamification can alleviate feedback distraction (**H4**). While the participants felt strongly distracted in the *Continuous x No Gamification* condition, they reported less distraction in the *Continuous x Emoji* condition, and even less in *Continuous x Progress Bar*. This shows a valuable benefit of gamified feedback modalities. However,

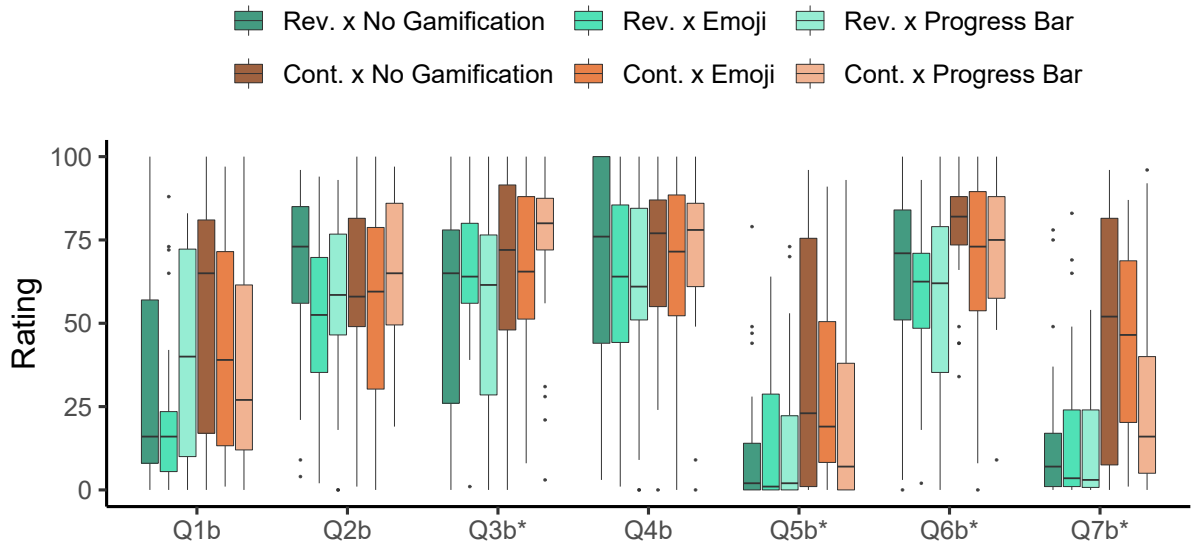

Figure 7: Ratings for questions Q1b-Q7b (see Table 2). Q3b ("I adapted my approach in solving the task due to the system."), Q5b ("The system interrupted me during the task."), Q6b ("The system helped me to see how well I was doing.") and Q7b ("The system distracted me during the task.") show significant differences for *Feedback Type*: *Revision* (shaded green) vs. *Continuous* (shaded orange). Q7b is additionally significant for *Gamification Type*. Significant questions are marked with *.

our results related to SIMS and IMI subscale responses indicated that the choice of game design elements significantly impacted users' motivation. Asked about their intrinsic motivation, participants indicated a significantly lower intrinsic motivation for the *Emoji* views, as compared to the *Progress Bar* views and to *No Gamification*. Reasons might include that users felt personally attacked by sad or disappointed emojis and might have seen little interest or value in feedback that was based on metaphors that they are commonly subjected to in social interactions. An additional indication of this was the significant difference in identified regulation between *No Gamification* and the *Emoji* views.

6 DISCUSSION

Here we reflect on the results of our inquiry and discuss how different forms of task-related proficiency feedback affect performance and user experience, taking a particular interest in its potential for self-reflection and facilitating an understanding for the user's own proficiency.

6.1 Task-Related Proficiency Feedback Provides Tangible Benefits for Task Fulfillment

Our investigation showed that users who were given proficiency feedback were able to fulfill task requirements (write in plain English) to a higher degree than those who did not receive any feedback, suggesting it as an effective method in helping users achieve a desired task outcome (**RQ1**). Interestingly, *Revision*-based methods were almost as accurate as *Continuous* feedback methods in Study I. We hypothesize that this can be attributed to the participants' awareness that they would be rated by a scoring system [18]. Study II confirmed that *Continuous* was superior in terms of task fulfillment (**RQ2**). Hence, **feedback types moderate task fulfillment and awareness of proficiency** [21]. Further, providing proficiency feedback has not significantly slowed down task completion, nor has it impacted its quality or increased the perceived workload of users. These results show that **task-related proficiency feedback can lead to tangible benefits in task fulfillment without negatively affecting effort, quality or speed** (**RQ2**).

6.2 Revision-Based Proficiency Feedback Exhibits Low Interruption Cost

While both feedback methods were able to support users in their task, *Revision*-based methods were perceived by the users as less disruptive. For *Continuous* methods, users reported that they felt interrupted and distracted by the scoring system to a higher degree. Thus, it is important for future designers to recognize the possible negative implications of providing proficiency feedback. **Providing feedback at sparser intervals is likely to reduce disruption** (**RQ2**). Consequently, the designer is faced with the challenge of finding the optimal balance for a given task. Frequent or continuous feedback will increase quality, but may frustrate users [30].

6.3 Proficiency Feedback Facilitates an Understanding of One's Own Proficiency

Our results suggest that users not only managed to perform more accurately when using proficiency feedback, but they also gained an understanding of their proficiency. We observed that users changed their behavior in reaction to being informed about their proficiency [8], evident from our slope analysis on *Revision*-based

methods and the accurate (and less variant) FRES scores over time for *Continuous* methods. Additionally, users stated that they had consciously adapted their task solving strategy and that the system helped them in doing so (**RQ1**), a clear indication of reflection-in-action [59]. Users were aware of their own approach to complete the task and could recognize opportunities to improve their strategy. This fact presents a design opportunity for future interactive systems. If proficiency (or aspects thereof) can be effectively recognized, **systems can guide users to selecting task completion strategies which are optimized for a given user's proficiency level**. At the same time, this quality may be misused for malicious intent as discussed further in Section 6.5 and needs to be moderated. Within the broader context of learning, our results suggest that tailoring automatic proficiency feedback and, potentially, combining it with peer feedback is a promising strategy for fostering learning quality [19].

6.4 Gamified Feedback Can Be Effective, but It Is Subject to Design Constraints

While gamified feedback was not more beneficial in terms of conveying proficiency to users, both gamification elements helped reduce the perceived interruption of the scoring system (**RQ3**). Compared to *Emoji*, *Progress Bar* provided a less distracting experience, which was almost on the level of *Revision*-based methods. We attribute this to the much more ambient appearance of the *Progress Bar*. The additional information regarding distance to the target zone and overall progress provided a tangible way to keep track of one's progress towards the target zone. This shows that gameful elements can help mitigate the disruptive effect of continuous proficiency assessment. Consequently, **future systems that want to avoid disruptiveness without jeopardizing performance, can explore gamified feedback instead of reducing feedback frequency**. At the same time, meaningful social comparison elements that appeal to the users are likely to stimulate extra effort that goes beyond merely reaching the target zone. A transparent communication of such information across co-workers and peers is likely to impact one's feeling of relatedness; one of the basic psychological needs [62].

Further, using the *Emoji* resulted in a significantly lower intrinsic motivation than the *Progress Bar* and the *No Gamification* condition. In conjunction with the lower score for identified regulation as well as perceived competence, we hypothesize that participants associated the *Emoji* with a form of childish, frivolous feedback. This was an impression that may have contrasted with the serious writing task [63]. Consequently, **if gamified elements are to be used for feedback, it is a key design consideration to align the feedback form with the content of the task at hand** (**RQ3**).

6.5 Opportunities and Challenges of Proficiency-Aware Systems

Our work contributes to exploring the design space of proficiency-aware systems, i.e., interactive systems that make use of task-related proficiency feedback. Related work by Malacria et al. [40] already found that proficiency feedback could increase productivity, through directly measuring and displaying the user's task performance. Whereas our work focuses on necessary skills for task fulfillment. By providing users with the means to reflect on specific aspects of their proficiency, we can **enable adaptive feedback to foster self-reflection and skill development**. Our findings provide insight into how to increase productivity through proficiency feedback and contributing ways to do so in a user-friendly manner. It remains a challenge for HCI to explore how proficiency-aware systems can be used in a wider array of tasks and contexts.

As our inquiry explores the means to aid users in performing tasks more effectively, it is also our responsibility to consider the ethical implications of the concepts proposed here on working conditions. Future systems must embody human values central to an ethical workplace [61]. Despite the fact that the feedback elements in our study were not graphically prominent during the experiment, some users did feel pressured by the system. More aggressive forms of feedback can potentially intimidate the user, coercing them into exhibiting certain behaviors. This way, users can lose their autonomy to what Zitrrain [68] dubbed "harvesters of human mindpower". This is even more dangerous if considered in the context of computer-based tasks having a tendency to deprive users from making judgements about the morality of their work [68]. To counteract possible negative implications, we recommend using ethics-oriented design methods, such as adversary design fictions [5], early in the design process to assure that ethical pitfalls are avoided. Most importantly, **users must always be given the opportunity to opt out from proficiency assessment** (see [68]). This provides them with the opportunity to make their own moral judgements of the assessment provided by the system.

6.6 Limitations and Future Work

As discussed in the previous section, having participants scored by a computing system can potentially be harmful and additionally changed their attitude towards the task at hand. Being told that one's submission would be scored might have influenced how participants in feedback conditions addressed the task, as compared to the *No Feedback* condition, where no scoring was present. This might explain the higher variance for FRES scores in this condition. Possible adjustments for future work include placebo-controlled study designs for AI systems [35] if working with conditions where no feedback is given.

Our work has investigated writing support for the English language[8] due to the availability of simple scoring mechanisms (FRES score). We highlight that the interpretation of and the openness towards our proficiency feedback might be differently for other languages and cultures. Future work should investigate which design factors can be generalized and which aspects should be individualized across cultures and writing styles, taking ethical implications into account as mentioned earlier.

7 CONCLUSION

In this paper, we evaluated the design qualities of task-related proficiency feedback in a text production task. Our investigation focused on different feedback types to communicate text readability to the user. We found that being aware of one's proficiency benefits task

[8]We included both native and L2 speakers in our study sample.

fulfillment and facilitates an understanding of their own writing expertise for the user. Consequently, users can adjust their approach towards the task and increase their proficiency.

While communicating proficiency is beneficial, we also found that feedback needs to be balanced to moderate potential interference with the task at hand. We suggest a revision-based approach if precise performance is not essential, as it is less distracting for the user. Similarly, subtle gamification elements like progress bars and associated social comparison complement this method by lowering perceived disruption.

Our work contributes to the understanding of how people reflect on their own skill assessments and, more importantly, how this self-reflection can be used to encourage users to improve their proficiency further. We envision this investigation as an initial exploration towards a more reflection-oriented design to user proficiency in interactive systems, allowing for an improved and engaging user experience.

ACKNOWLEDGMENTS

This research is supported by the European Union's Horizon 2020 Programme under ERCEA grant no. 683008 AMPLIFY and 952026 HumanE-AI-Net.

REFERENCES

[1] Amid Ayobi, Paul Marshall, and Anna L Cox. 2020. Trackly : A Customisable and Pictorial Self-Tracking App to Support Agency in Multiple Sclerosis Self-Care. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2020), 1–15.

[2] Nikola Banovic, Fanny Chevalier, Tovi Grossman, and George Fitzmaurice. 2012. Triggering Triggers and Burying Barriers to Customizing Software. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '12)*. Association for Computing Machinery, New York, NY, USA, 2717–2726. https://doi.org/10.1145/2207676.2208666

[3] Gabriel Barata, Sandra Gama, Joaquim Jorge, and Daniel Gonçalves. 2013. Improving Participation and Learning with Gamification. In *Proceedings of the First International Conference on Gameful Design, Research, and Applications* (Toronto, Ontario, Canada) *(Gamification '13)*. Association for Computing Machinery, New York, NY, USA, 10–17. https://doi.org/10.1145/2583008.2583010

[4] Scott Bateman, Jaime Teevan, and Ryen W. White. 2012. The Search Dashboard: How Reflection and Comparison Impact Search Behavior. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '12)*. Association for Computing Machinery, New York, NY, USA, 1785–1794. https://doi.org/10.1145/2207676.2208311

[5] Eric P.S. Baumer, Timothy Berrill, Sarah C. Botwinick, Jonathan L. Gonzales, Kevin Ho, Allison Kundrik, Luke Kwon, Tim LaRowe, Chanh P. Nguyen, Fredy Ramirez, Peter Schaedler, William Ulrich, Amber Wallace, Yuchen Wan, and Benjamin Weinfeld. 2018. What Would *You* Do? Design Fiction and Ethics. In *Proceedings of the 2018 ACM Conference on Supporting Groupwork (GROUP '18)*. Association for Computing Machinery, New York, NY, USA, 244–256. https://doi.org/10.1145/3148330.3149405

[6] Eric P.S. Baumer, Vera Khovanskaya, Mark Matthews, Lindsay Reynolds, Victoria Schwanda Sosik, and Geri Gay. 2014. Reviewing Reflection: On the Use of Reflection in Interactive System Design. In *Proceedings of the 2014 Conference on Designing Interactive Systems (DIS '14)*. Association for Computing Machinery, New York, NY, USA, 93–102. https://doi.org/10.1145/2598510.2598598

[7] Marit Bentvelzen, Jasmin Niess, and Paweł W. Woźniak. 2021. The Technology-Mediated Reflection Model: Barriers and Assistance in Data-Driven Reflection. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 246, 12 pages. https://doi.org/10.1145/3411764.3445505

[8] Susanne Bødker. 1995. Applying Activity Theory to Video Analysis: How to Make Sense of Video Data in Human-Computer Interaction. In *Context and Consciousness: Activity Theory and Human-Computer Interaction*. Massachusetts Institute of Technology, USA, 147–174.

[9] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models Are Few-Shot Learners. *arXiv:2005.14165 [cs]* (July 2020). arXiv:2005.14165 [cs] http://arxiv.org/abs/2005.14165

[10] J. Carroll and M. Rosson. 1987. Paradox of the Active User. In *Interfacing Thought: Cognitive Aspects of Human-Computer Interaction*. The MIT Press.

[11] Center for Self-Determination Theory. [n. d.]. Intrinsic Motivation Inventory. http://selfdeterminationtheory.org/intrinsic-motivation-inventory/

[12] Bo Chang. 2019. Reflection in learning. *Online Learning* 23, 1 (2019), 95–110.

[13] Elizabeth Clark, Anne Spencer Ross, Chenhao Tan, Yangfeng Ji, and Noah A. Smith. 2018. Creative Writing with a Machine in the Loop: Case Studies on Slogans and Stories. In *23rd International Conference on Intelligent User Interfaces (IUI '18)*. Association for Computing Machinery, New York, NY, USA, 329–340. https://doi.org/10.1145/3172944.3172983

[14] Andy Cockburn, Carl Gutwin, Joey Scarr, and Sylvain Malacria. 2014. Supporting Novice to Expert Transitions in User Interfaces. *ACM Comput. Surv.* 47, 2 (Nov. 2014), 31:1–31:36. https://doi.org/10.1145/2659796

[15] Richard T. Cole, Larry A. Hembroff, and Andrew D. Corner. 2009. National Assessment of the Perceived Writing Skills of Entry-Level PR Practitioners. *Journalism & Mass Communication Educator* 64, 1 (March 2009), 9–26. https://doi.org/10.1177/107769580906400102

[16] David Denton. 2011. Reflection and learning: Characteristics, obstacles, and implications. *Educational Philosophy and Theory* 43, 8 (2011), 838–852.

[17] Sebastian Deterding, Dan Dixon, Rilla Khaled, and Lennart Nacke. 2011. From Game Design Elements to Gamefulness: Defining "Gamification". In *Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments* (Tampere, Finland) *(MindTrek '11)*. ACM, New York, NY, USA, 9–15. https://doi.org/10.1145/2181037.2181040

[18] Steven Dow, Anand Kulkarni, Scott Klemmer, and Björn Hartmann. 2012. Shepherding the Crowd Yields Better Work. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work (CSCW '12)*. Association for Computing Machinery, New York, NY, USA, 1013–1022. https://doi.org/10.1145/2145204.2145355

[19] Peggy A Ertmer, Jennifer C Richardson, Brian Belland, Denise Camin, Patrick Connolly, Glen Coulthard, Kimfong Lei, and Christopher Mong. 2007. Using peer feedback to enhance the quality of student online postings: An exploratory study. *Journal of Computer-Mediated Communication* 12, 2 (2007), 412–433.

[20] Leah Findlater and Joanna McGrenere. 2004. A Comparison of Static, Adaptive, and Adaptable Menus. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '04)*. Association for Computing Machinery, New York, NY, USA, 89–96. https://doi.org/10.1145/985692.985704

[21] Leah Findlater and Joanna McGrenere. 2010. Beyond Performance: Feature Awareness in Personalized Interfaces. *International Journal of Human-Computer Studies* 68, 3 (March 2010), 121–137. https://doi.org/10.1016/j.ijhcs.2009.10.002

[22] Rudolph Flesch. 1948. A New Readability Yardstick. *Journal of Applied Psychology* 32, 3 (1948), 221–233. https://doi.org/10.1037/h0057532

[23] Krzysztof Z. Gajos, Katherine Everitt, Desney S. Tan, Mary Czerwinski, and Daniel S. Weld. 2008. Predictability and Accuracy in Adaptive User Interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '08)*. Association for Computing Machinery, New York, NY, USA, 1271–1274. https://doi.org/10.1145/1357054.1357252

[24] Katy Ilonka Gero and Lydia B. Chilton. 2019. Metaphoria: An Algorithmic Companion for Metaphor Creation. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300526

[25] Frédéric Guay, Robert J. Vallerand, and Céline Blanchard. 2000. On the Assessment of Situational Intrinsic and Extrinsic Motivation: The Situational Motivation Scale (SIMS). *Motivation and Emotion* 24, 3 (Sept. 2000), 175–213. https://doi.org/10.1023/A:1005614228250

[26] Guy Brook-Hart and Simon Haines. 2014. *Complete Advanced* (second edition ed.). Cambridge University Press.

[27] Juho Hamari and Jonna Koivisto. 2014. Measuring flow in gamification: Dispositional flow scale-2. *Computers in Human Behavior* 40 (2014), 133–143. https://doi.org/10.1016/j.chb.2014.07.048

[28] Juho Hamari, Jonna Koivisto, Harri Sarsa, et al. 2014. Does Gamification Work?-A Literature Review of Empirical Studies on Gamification.. In *HICSS*, Vol. 14. 3025–3034.

[29] Sandra G. Hart. 2006. Nasa-Task Load Index (NASA-TLX); 20 Years Later. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 50, 9 (2006), 904–908. https://doi.org/10.1177/154193120605000909 arXiv:https://doi.org/10.1177/154193120605000909

[30] Shamsi T. Iqbal and Eric Horvitz. 2010. Notifications and Awareness: A Field Study of Alert Usage and Preferences. In *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work (CSCW '10)*. Association for Computing Machinery, New York, NY, USA, 27–30. https://doi.org/10.1145/1718918.1718926

[31] Lawrence R. James, Robert G. Demaree, and Gerrit Wolf. 1984. Estimating Within-Group Interrater Reliability with and without Response Bias. *Journal of Applied*

Psychology 69, 1 (1984), 85–98. https://doi.org/10.1037/0021-9010.69.1.85

[32] Adam C. Johnson, Joshua Wilson, and Rod D. Roscoe. 2017. College Student Perceptions of Writing Errors, Text Quality, and Author Characteristics. *Assessing Writing* 34 (Oct. 2017), 72–87. https://doi.org/10.1016/j.asw.2017.10.002

[33] Jakob Karolus and Pawel W. Wozniak. 2021. Proficiency-Aware Systems: Designing for User Reflection in Context-Aware Systems. *it Inf. Technol.* 63, 3 (2021), 167–175. https://doi.org/10.1515/itit-2020-0039

[34] Ryan Kelly, Daniel Gooch, and Leon Watts. 2018. 'It's More Like a Letter': An Exploration of Mediated Conversational Effort in Message Builder. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW (Nov. 2018), 87:1–87:23. https://doi.org/10.1145/3274356

[35] Thomas Kosch, Robin Welsch, Lewis Chuang, and Albrecht Schmidt. 2023. The Placebo Effect of Artificial Intelligence in Human–Computer Interaction. *ACM Trans. Comput.-Hum. Interact.* 29, 6 (Jan. 2023), 56:1–56:32. https://doi.org/10.1145/3529225

[36] Justin Kruger and David Dunning. 1999. Unskilled and Unaware of It: How Difficulties in Recognizing One's Own Incompetence Lead to Inflated Self-Assessments. *Journal of Personality and Social Psychology* 77 (1999), 1121–1134.

[37] David M. Lane, H. Albert Napier, S. Camille Peres, and Aniko Sandor. 2005. Hidden Costs of Graphical User Interfaces: Failure to Make the Transition from Menus and Icon Toolbars to Keyboard Shortcuts. *International Journal of Human-Computer Interaction* 18, 2 (May 2005), 133–144. https://doi.org/10.1207/s15327590ijhc1802_1

[38] Xiaotong Liu, Anbang Xu, Zhe Liu, Yufan Guo, and Rama Akkiraju. 2019. Cognitive Learning: How to Become William Shakespeare. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems (CHI EA '19)*. Association for Computing Machinery, New York, NY, USA, 1–6. https://doi.org/10.1145/3290607.3312844

[39] Wendy E. Mackay. 1991. Triggers and Barriers to Customizing Software. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '91)*. Association for Computing Machinery, New York, NY, USA, 153–160. https://doi.org/10.1145/108844.108867

[40] Sylvain Malacria, Joey Scarr, Andy Cockburn, Carl Gutwin, and Tovi Grossman. 2013. Skillometers: Reflective Widgets That Motivate and Help Users to Improve Performance. In *Proceedings of the 26th Annual ACM Symposium on User Interface Software and Technology (UIST '13)*. Association for Computing Machinery, St. Andrews, Scotland, United Kingdom, 321–330. https://doi.org/10.1145/2501988.2501996

[41] Sampada Marathe and S. Shyam Sundar. 2011. What Drives Customization? Control or Identity?. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11)*. Association for Computing Machinery, New York, NY, USA, 781–790. https://doi.org/10.1145/1978942.1979056

[42] Justin Matejka, Wei Li, Tovi Grossman, and George Fitzmaurice. 2009. CommunityCommands: Command Recommendations for Software Applications. In *Proceedings of the 22nd Annual ACM Symposium on User Interface Software and Technology (UIST '09)*. Association for Computing Machinery, New York, NY, USA, 193–202. https://doi.org/10.1145/1622176.1622214

[43] Csikszentmihalyi Mihaly. 1990. Flow: The Psychology of Optimal Performance. (1990).

[44] Jeanne Nakamura and Mihaly Csikszentmihalyi. 2009. Flow Theory and Research. *The Oxford Handbook of Positive Psychology* (2009), 195–206. https://doi.org/10.1093/oxfordhb/9780195187243.013.0018

[45] Eslam Nofal, Georgia Panagiotidou, Rabee M. Reffat, Hendrik Hameeuw, Vanessa Boschloos, and Andrew Vande Moere. 2020. Situated Tangible Gamification of Heritage for Supporting Collaborative Learning of Young Museum Visitors. *J. Comput. Cult. Herit.* 13, 1, Article 3 (Feb. 2020), 24 pages. https://doi.org/10.1145/3350427

[46] Rita Orji, Gustavo F. Tondello, and Lennart E. Nacke. 2018. *Personalizing Persuasive Strategies in Gameful Systems to Gamification User Types*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3173574.3174009

[47] Rita Orji, Gustavo F Tondello, and Lennart E Nacke. 2018. Personalizing persuasive strategies in gameful systems to gamification user types. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–14. https://doi.org/10.1145/3173574.3174009

[48] Zhenhui Peng, Qingyu Guo, Ka Wing Tsang, and Xiaojuan Ma. 2020. Exploring the Effects of Technological Writing Assistance for Support Providers in Online Mental Health Community. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–15. https://doi.org/10.1145/3313831.3376695

[49] James L. Peterson. 1980. Computer Programs for Detecting and Correcting Spelling Errors. *Commun. ACM* 23, 12 (Dec. 1980), 676–687. https://doi.org/10.1145/359038.359041

[50] David Robinson and Victoria Bellotti. 2013. A preliminary taxonomy of gamification elements for varying anticipated commitment. In *Proc. ACM CHI 2013 Workshop on Designing Gamification: Creating Gameful and Playful Experiences*.

[51] Yvonne Rogers. 2006. Moving on from Weiser's Vision of Calm Computing: Engaging UbiComp Experiences. In *UbiComp 2006: Ubiquitous Computing (Lecture Notes in Computer Science)*, Paul Dourish and Adrian Friday (Eds.). Springer, Berlin, Heidelberg, 404–421. https://doi.org/10.1007/11853565_24

[52] Wendy Roldan, Ziyue Li, Xin Gao, Sarah Kay Strickler, Allison Marie Hishikawa, Jon E. Froehlich, and Jason Yip. 2021. Pedagogical Strategies for Reflection in Project-based HCI Education with End Users. In *Designing Interactive Systems Conference 2021 (DIS '21)*. Association for Computing Machinery, New York, NY, USA, 1846–1860. https://doi.org/10.1145/3461778.3462113

[53] Richard M Ryan and Edward L Deci. 2000. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American psychologist* 55, 1 (2000), 68.

[54] Richard M Ryan and Heather Patrick. 2009. Self-determination theory and physical. *Hellenic journal of psychology* 6 (2009), 107–124.

[55] John Sadauskas, Daragh Byrne, and Robert K. Atkinson. 2015. Mining Memories: Designing a Platform to Support Social Media Based Writing. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. Association for Computing Machinery, New York, NY, USA, 3691–3700. https://doi.org/10.1145/2702123.2702383

[56] Herman Saksono and Andrea G. Parker. 2017. Reflective informatics through family storytelling: Self-discovering physical activity predictors. *Conference on Human Factors in Computing Systems - Proceedings* 2017-May (2017), 5232–5244. https://doi.org/10.1145/3025453.3025651

[57] Alan W. Salmoni, Richard A. Schmidt, and Charles B. Walter. 1984. Knowledge of Results and Motor Learning: A Review and Critical Reappraisal. *Psychological Bulletin* 95, 3 (1984), 355–386. https://doi.org/10.1037/0033-2909.95.3.355

[58] Joey Scarr, Andy Cockburn, Carl Gutwin, and Philip Quinn. 2011. Dips and Ceilings: Understanding and Supporting Transitions to Expertise in User Interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11)*. Association for Computing Machinery, New York, NY, USA, 2741–2750. https://doi.org/10.1145/1978942.1979348

[59] D.A. Schön. 1983. *The Reflective Practitioner: How Professionals Think in Action*.

[60] Katie Seaborn and Deborah I Fels. 2015. Gamification in theory and action: A survey. *International Journal of human-computer studies* 74 (2015), 14–31.

[61] Abigail Sellen, Yvonne Rogers, Richard Harper, and Tom Rodden. 2009. Reflecting human values in the digital age. *Commun. ACM* 52, 3 (March 2009), 58–66. https://doi.org/10.1145/1467247.1467265

[62] Kennon M Sheldon, Andrew J Elliot, Youngmee Kim, and Tim Kasser. 2001. What is satisfying about satisfying events? Testing 10 candidate psychological needs. *Journal of personality and social psychology* 80, 2 (2001), 325.

[63] Roberta Shroyer. 2000. Actual Readers versus Implied Readers: Role Conflicts in Office 97. *Technical Communication* 47, 2 (2000), 238–240. jstor:43748856 http://www.jstor.org/stable/43748856

[64] C-H. Su and C-H. Cheng. 2015. A mobile gamification learning system for improving the learning motivation and achievements. *Journal of Computer Assisted Learning* 31, 3 (2015), 268–286. https://doi.org/10.1111/jcal.12088 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/jcal.12088

[65] Gustavo F. Tondello, Alberto Mora, and Lennart E. Nacke. 2017. Elements of Gameful Design Emerging from User Preferences. *Proceedings of the Annual Symposium on Computer-Human Interaction in Play - CHI PLAY '17* (2017), 129–142. https://doi.org/10.1145/3116595.3116627

[66] Lauren Wilcox, Betsy DiSalvo, Dick Henneman, and Qiaosi Wang. 2019. Design in the HCI Classroom: Setting a Research Agenda. In *Proceedings of the 2019 on Designing Interactive Systems Conference (DIS '19)*. Association for Computing Machinery, New York, NY, USA, 871–883. https://doi.org/10.1145/3322276.3322381

[67] Jacob O. Wobbrock, Leah Findlater, Darren Gergle, and James J. Higgins. 2011. The Aligned Rank Transform for Nonparametric Factorial Analyses Using Only Anova Procedures. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11)*. ACM, New York, NY, USA, 143–146. https://doi.org/10.1145/1978942.1978963

[68] Jonathan Zittrain. 2008. Ubiquitous human computing. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 366, 1881 (Oct. 2008), 3813–3821. https://doi.org/10.1098/rsta.2008.0116