

Enhancing Response Quality by Children in Voice-based Sleep Diaries via AI-based Continuous Feedback

Shanshan Chen

Department of Industrial Design
Eindhoven University of Technology
Eindhoven, North Brabant, Netherlands
s.chen1@tue.nl

Gubing Wang

TS Social and Behavioral Sciences
Tilburg University
Tilburg, North Brabant, Netherlands
Medical and Clinical Psychology
Tilburg University
Tilburg, North Brabant, Netherlands
g.wang_2@tilburguniversity.edu

Jun Hu

Department of Industrial Design
Eindhoven University of Technology
Eindhoven, North Brabant, Netherlands
j.hu@tue.nl

Panos Markopoulos

Department of Industrial Design
Eindhoven University of Technology
Eindhoven, North Brabant, Netherlands
p.markopoulos@tue.nl

Abstract

Digital sleep diaries are widely used in clinical practice and research to monitor children's subjective sleep quality. A well-known limitation of survey methods is that children may not provide high-quality responses because they cannot or are not motivated to do so. We examine how to design "live", continuous feedback in voice-based sleep diaries in order to enhance the quality of children's responses. In a co-design workshop, we explored children's preferences for different forms of feedback. We designed and compared experimentally symbolic (smiley), numeric, and no-feedback conditions, showing that both feedback types improved response quality across questions. Finally, an eight-day field study revealed that feedback resulted in higher and more consistent quality in self-report over time. Across these three studies, children valued playful and clear feedback, with preferences shifting depending on their cognitive needs. Our findings provide evidence that effective feedback must balance affective engagement and cognitive clarity and adapt to different contexts. We contribute empirically supported design insights for creating child-centered voice-based surveys that aim to enhance children's adherence in independent self-report surveys. Our recommendations based on the study of sleep diaries can potentially be applied in other areas using voice-based surveys.

CCS Concepts

• **Human-centered computing** → **Field studies**.

Keywords

response quality, self-report, children, feedback, sleep diary

ACM Reference Format:

Shanshan Chen, Jun Hu, Gubing Wang, and Panos Markopoulos. 2026. Enhancing Response Quality by Children in Voice-based Sleep Diaries via AI-based Continuous Feedback. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26)*, April 13–17, 2026, Barcelona, Spain. ACM, New York, NY, USA, 21 pages. <https://doi.org/10.1145/3772318.3790684>

1 Introduction

At 9 p.m., eight-year-old Mia sits with her tablet, answering questions in a sleep diary about her day. She recalls waking up during the night and feeling tired in the morning, but stops after a few short sentences. There is no indication whether her answer is detailed enough to help her doctor understand her sleep patterns. The next day, her responses are even shorter. For many children, the absence of guidance or feedback during self-reporting can result in incomplete or low quality entries, limiting the usefulness of the data for clinical care.

Diaries are a widely used self-reporting tool in research, education and healthcare to capture experiences that cannot be directly observed, such as emotions, habits, or symptoms [97]. Unlike one-time surveys, diaries require repeated entries, making both the quality of each individual response and the ability to sustain that quality over time critical to their usefulness. Yet obtaining high-quality entries from children is notoriously difficult [6, 23]. Compared to adults, children are still developing in language, memory, literacy, and self-regulation [82], which often leads to short, vague, or incomplete responses in surveys and diaries [6]. These difficulties are especially in health-related diaries, where poor or inconsistent reporting can directly affect clinical interpretation. Among such tools, sleep diaries stand out as both widespread and clinically significant. A sleep diary is a structured tool where individuals report bedtimes, sleep routines, nocturnal disturbances (e.g., media use before bed), and daytime functioning affected by sleep [17, 27, 106]. These diaries provide crucial information for clinicians to evaluate and treat behavioral sleep problems [118].

However, the value of these diaries depends directly on the quality of the recorded responses, which is not always sufficient



This work is licensed under a Creative Commons Attribution 4.0 International License. *CHI '26, Barcelona, Spain*

© 2026 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-2278-3/26/04

<https://doi.org/10.1145/3772318.3790684>

or consistent for clinical use, particularly in children. Here, the two challenges are especially pronounced: 1) children may not know what constitutes a "good" answer (**response quality across questions**) in one session, and 2) they may struggle to sustain consistent effort across days (**response quality over days**), which requires skills such as delaying gratification, focusing attention, managing emotions, and regulating behaviors [15, 20, 114]. This paper focuses on strategies to address both challenges, helping children produce high-quality responses for each question and sustain that quality over repeated diary use.

Feedback offers a promising solution because it can guide children in refining individual responses and motivates them to sustain effort over time. First, feedback can serve as **guidance** [91], directly addressing *response quality across questions*. By clarifying what quality performance looks like and how to achieve it, feedback helps respondents refine and improve their answers while they are still responding [9, 54]. In sleep diaries, where every response is valuable for clinical interpretation, leaving an answer suboptimal is a missed opportunity [112]. Second, feedback can support **motivation** [34], helping maintain *response quality over days*. Positive, success-focused feedback enhances intrinsic motivation and persistence [78, 81], which is especially critical for encouraging children to provide detailed, high-quality responses across multiple diary entries.

A common approach in conversational systems is *intermittent feedback* [9, 105], where praise or encouragement is given only *after* a response [11, 56, 63], consistent with everyday conversational norms [44]. However, providing feedback only after the child has finished speaking means that any issues in the current answer remain uncorrected, and the opportunity to improve that answer is lost until the next question. In contrast, we focus on *"live", continuous feedback* [19]: feedback delivered in real time while the child is still speaking. Because it unfolds continuously and without interrupting the conversational flow, this feedback allows children to refine and expand their responses before they are finalized. This approach is particularly well suited to diary contexts, where children must generate answers on the spot without prompts or examples, making their responses especially prone to being brief or incomplete.

Building on this idea, we implement continuous feedback in a voice-based conversational agent that functions as a sleep diary for children. The agent enables children to report their sleep experiences through spoken dialogue, reducing literacy demands for those with limited reading and writing skills [22, 24, 52, 74, 96]. Yet, because spoken responses are generated spontaneously and without visual scaffolding, they can be fragmented or vague. To address this, our system provides *"live", continuous, and visual feedback* on response quality as the child speaks. This feedback updates dynamically in the user interface without interrupting the conversation, allowing children to adjust their answers in real time. Despite extensive work on feedback in conversational technologies [9, 11], its role in supporting response quality in children's voice-based conversational diaries remains largely unexplored.

To address the two central challenges of children's voice-based diaries, improving the quality of each response and sustaining that quality across repeated entries, we conducted a series of studies that build upon one another. First, we carried out a co-design study to understand children's perspectives on what makes response-quality

feedback in conversation helpful or burdensome. Based on these insights, we developed two types of "live", continuous feedback, numeric (score-based) and symbolic (smiley-based), and integrated them into a voice-based sleep diary system. To evaluate whether these feedback types improve **response quality across questions within a single reporting session** (Challenge 1), we conducted a controlled lab study with 36 children aged 7 to 12. Continuous feedback significantly improved response quality across questions, although we found no significant difference between feedback types. To investigate whether continuous feedback could also improve **response quality over days** (Challenge 2), we conducted a field study with 24 children. In this study, we implemented a combined version of the diary that incorporated both numeric and symbolic feedback, and compared it with a no-feedback baseline to assess how feedback influenced the trajectory of response quality during repeated home use.

This paper offers the following main contributions:

- Insights from children on conversational feedback design, emphasizing the importance of clarity, interpretability, and low cognitive demand.
- Empirical evidence from lab and field studies showing that continuous feedback improves children's response quality both within individual sessions and across multiple days.
- Design implications for continuous feedback systems to support high-quality response in child-centered voice diaries.

Overall, our findings demonstrate that continuous feedback can support children in producing more consistent, high-quality self-reports in voice-based digital diaries. This work highlights promising directions for advancing pediatric medical and healthcare self-reporting systems through responsive, real-time guidance.

2 Related Work

2.1 Response Quality in Sleep Diary

Data quality is crucial for the validity of behavior research in both academic and clinical settings [45]. Because clinicians, educators, and researchers often require diaries to be completed over several days or weeks, the data must be *valid for each question* and *consistent across multiple entries*. For children, however, maintaining high-quality diary responses is especially difficult. Limited literacy skills and still-developing self-regulation abilities, such as delaying gratification, sustaining attention, managing emotions [7, 15, 114], and controlling behaviors, make it harder for children to engage consistently and provide detailed responses.

2.1.1 Two Central Challenges of Response Quality for Children.

Researchers have explored several strategies to improve response quality in self-reporting context. In online surveys, methods such as participant screening help ensure attentiveness [3]. In HCI, conversational interfaces have been shown to increase engagement and improve the quality of course evaluations [115], writing [85], and goal-oriented task performance [57]. Studies comparing input modalities indicate that voice-based interaction can elicit richer, more complete responses than text [67, 116]. However, few of these efforts target *children*, whose cognitive and communicative capacities differ substantially from those of adults.

Recently, Chen et al. developed a voice-based sleep diary for children and observed that voice interaction improved response quality during a five-day field deployment, but that children’s responses were still often incomplete and response quality declined across days [24]. These findings highlight two central challenges: achieving *response quality across questions* and maintaining *response quality over days*.

2.1.2 Why Children-Specific Feedback Design is Needed. Developmental research shows that children interpret feedback differently from adults due to differences in language comprehension, cognitive load tolerance, and metacognitive monitoring [48, 93]. Children tend to rely on concrete, visually salient cues, respond affectively to symbolic indicators, and have difficulty integrating multiple feedback signals while simultaneously performing a task such as speaking [77, 88, 93]. As a result, feedback strategies effective for adults, or even adolescents, cannot be directly transferred to younger populations.

This literature underscores the need to design from children’s perspectives: their developmental characteristics shape whether feedback is understandable, motivating, or usable during voice-based self-report. Therefore, understanding how children perceive and interpret feedback is a necessary foundation for improving response quality in children’s sleep diaries.

2.2 Functions of Response Quality Feedback

In learning research, performance feedback is widely studied as a mechanism that influences both cognition and affect. Feedback not only conveys information that helps improve task performance (*guidance*), but also enhances persistence and engagement (*motivation*) by providing experiences akin to rewards and punishments [109]. These two functions map closely to the challenges outlined above:

- **Guidance** supports *response quality across questions* by helping children improve the completeness, relevance, and clarity of each response.
- **Motivation** supports *response quality over days* by encouraging children to remain engaged and maintain high standards across multiple entries.

Below, we review existing research related to each of these functions.

2.2.1 Guidance. Guidance refers to the role of feedback in guiding task performance. Duijnhouwer et al. define it as “information provided by an external agent regarding some aspect(s) of the learner’s task performance, intended to modify the learner’s cognition, motivation, and/or behavior for the purpose of improving performance” [41]. By signal what aspects of a response are successful or require adjustment, feedback helps individuals align their efforts more closely with task goals [5, 42].

Research across domains demonstrates the practical benefits of feedback for refining performance. In education, performance feedback improves motivation and supports learning and memory [34]. In motor learning, feedback reduces errors and sharpens skills [99, 101]. In writing instruction, formative feedback clarifies quality criteria and supports iterative improvement [54, 90, 107]. A similar positive influence has been reported in academic settings

where high levels of teacher feedback are associated with better student [50]. Comparable effects appear in healthcare [9, 11, 86], games [2], and sports [110], where feedback guides users’ attention, strategy, and behavior. Despite extensive work investigating feedback as *guidance* in learning, health, and interactive systems [124], this function remain largely unexplored in voice-based diaries. We extend this research by investigating whether continuous, performance-related feedback can guide children toward producing higher-quality responses across questions in sleep diary entries.

2.2.2 Motivation. Motivation refers to the role of feedback in maintaining engagement over time. Intrinsic motivation, the desire to engage in a task because it is interesting or enjoyable, is a well-established driver of persistence [31]. Feedback can foster this motivation by reinforcing either *wishes* (positive feedback) or *obligations* (negative feedback), both of which strengthen long-term task commitment [47, 71, 111].

Research shows that feedback helps maintain engagement and adherence in repeated-task settings. Detailed performance feedback improved productivity in daily data-entry work [104]; visualized feedback increased compliance with daily mood and interruptibility surveys [61]; performance feedback increased hand-hygiene adherence among healthcare workers [108]; and progress feedback supported sustained physical activity over a 12-week program [103]. These examples suggest that feedback can help users remain engaged even when tasks become repetitive or require ongoing effort. However, despite strong evidence in education, work, and health domains, the motivational role of feedback in children’s self-reporting remains underexplored. Voice-based diaries require repeated engagement across multiple days, and children’s developing self-regulation and attention control make sustained high-quality reporting particularly challenging.

In summary, feedback improves task performance and persistence across many domains [61, 103, 104, 108], yet prior work rarely address the *core challenge of maintaining high-quality responses* in self-reporting, particularly for children. This paper addresses this gap by investigating how *response-quality feedback* can support both *response quality across questions* and *response quality over days* in children’s sleep diaries.

2.3 Methods of Response Quality Feedback

Performance feedback refers to information given to individuals about the quantity or quality of their past performance [94]. In diary contexts, this takes the form of *response quality feedback*, which is information that reflects how well a self-report aligns with expectations for completeness, relevance, and clarity. Previous research has introduced quantitative methods for calculating response quality in surveys and diaries [24, 72, 122].

In this work, we adopt a quantitative approach so children can receive clear and direct indications of the quality of each response. Guided by Semiotics [43], which describes how meaning is communicated through signs, and self-regulation theory [5], which emphasizes feedback for performance improvement, we categorize response-quality feedback into three widely used forms: *numeric feedback*, *symbolic feedback*, and *progress feedback*.

We selected these three types because they represent the dominant ways performance information is conveyed in HCI, education, and learning sciences [14, 68]. Numeric and symbolic signals map directly to semiotic representations, while progress feedback reflects cumulative self-monitoring in self-regulation theory. Importantly, these formats are familiar to children from classroom dashboards, gamified apps, and learning systems [1, 60, 89], making them cognitively accessible during ongoing conversation. We considered alternative feedback modes such as auditory cues [87], natural-language explanations [38], and narrative or character-based feedback [13, 62], but excluded them to avoid confounds. These methods either lack strong grounding for conveying performance quality or introduce entertainment elements that could overshadow performance signals. Therefore, our focus on numeric, symbolic, and progress feedback enables a controlled, theoretically anchored investigation of how children interpret response-quality cues.

2.3.1 Numeric Feedback. Numeric feedback presents performance information in the form of numbers, scores, or ratings [59]. It is typically displayed in formats such as centesimal (percentage), ten-point, or five-point systems, and often requires minimal interpretation by the user. Examples include numeric ratings in online reviews [59], personalized digital score feedback to improve exam performance [18], student-rating feedback to enhance college instruction [29], and rubric-based scoring for formative writing feedback [120]. While such methods have shown clear benefits in education, their effectiveness for *improving performance in diary tasks*, particularly in children’s voice-based diaries, has not yet been examined.

2.3.2 Symbolic Feedback. Symbolic feedback uses culturally familiar visual cues to represent performance levels [95]. These symbols can convey meaning quickly and intuitively, and may include emojis [100], thumbs-up/down icons [16, 92], or even visual metaphors such as flowers [88]. Research shows that symbolic feedback can influence emotional responses and interpersonal interactions [77], signal approval or disapproval in task performance [92], and reinforce learning through visual cues [88]. This approach may be especially suitable for children, who often find symbols easier to interpret and more engaging than purely numeric scores. However, its potential to improve response quality in diary-based self-reporting remains unexplored.

2.3.3 Progress Feedback. Progress feedback provides cumulative information about one’s performance over time [30, 102]. In clinical contexts, it supports *routine outcome monitoring*, allowing clinicians to track patient-reported outcomes across treatment sessions [30, 70]. In education, progress feedback has been shown to improve mastery goals, self-efficacy, and writing performance [40], and adaptive progress cues in the conversational agent can motivate certain learners more effectively [32]. Although this feedback type is widely used in learning and clinical monitoring [30, 102], its role in supporting consistent response quality in children’s voice-based diary has not yet been explored.

In summary, across education, communication, and clinical monitoring, numeric, symbolic, and progress feedback methods have demonstrated effectiveness in guiding performance and sustaining engagement. However, none have been systematically investigated

as tools for improving response quality in children’s self-reporting, particularly in voice-based systems that require both *response quality across questions* and *response quality over days*. Therefore, these three categories form the foundation for our co-design and subsequent user studies.

2.4 AI-supported Automatic Continuous Feedback

Recent advances in AI have enabled systems to deliver automatic, continuous feedback by comparing a user’s current performance against predefined standards or model responses. For example, Loder et al. provided immediate corrective feedback in learning tasks by automatically evaluating learners’ responses against expected answers [76].

More recently, large language models (LLMs) and other AI tools have expanded the scope of such feedback systems. Duan et al. integrated GPT-4 into a Figma plugin to generate automatic continuous design suggestions, demonstrating comparable effectiveness to expert human assessors [38]. Rosiana et al. combined natural language processing and scene understanding to automatically generate feedback on scene descriptions, increasing the efficiency and quality of audio description workflows [83]. Weng et al. developed an AI-based coaching system that delivered real-time guidance for basketball shooting, improving beginners’ technique, confidence, and self-awareness [117]. These examples illustrate the potential of AI-powered systems to provide immediate, context-sensitive feedback that improves user performance across a wide range of domains – from design and content creation to sports training. However, these capabilities have rarely been applied to self-reporting tasks, and to our knowledge have not yet been explored in children’s voice-based diaries.

In this work, we leverage LLM-based techniques to automatically assess the quality of children’s spoken responses in a sleep diary, generate “live”, continuous performance feedback, and display it dynamically in the conversational interface. This integration of AI ensures that feedback is both immediate and adaptive, enabling children to adjust their responses as they speak, thereby supporting clinicians’ and treatment planning for chronic conditions such as sleep disorders.

3 Research Focus

The primary aim of this paper is to enhance the response quality of children’s self-reports in a voice-based sleep diary. As outlined in our Related Work (Section 2.4), prior research suggests that AI-supported, automatic, continuous feedback has strong potential to enhance user performance [38, 76, 83, 117]. However, no prior studies have examined how such feedback could be designed for, and applied to, children’s voice-based diaries. Our research builds on two key insights from the literature: First, as discussed in Section 2.1, children’s developmental characteristics shape how they interpret and act on feedback, making it essential that **feedback design starts from children’s perspectives**. Second, **feedback should address the two core challenges** identified in Section 2.1: *improving response quality across questions* and *maintaining response quality over days*. Based on these insights, we structured our work into four research questions:

- **RQ1:** What are the children’s preferences for the design of response-quality feedback in voice-based diaries?
- **RQ2:** How does continuous response-quality feedback influence the quality of children’s answers across questions in a single session?
- **RQ3:** How does continuous response-quality feedback influence the quality of children’s answers across multiple day entries over time?
- **RQ4:** How do children perceive the continuous response-quality feedback after using it in voice-based diaries?

To address these questions, we conducted three studies:

- **Study 1 – Co-design:** We used role-play activities to explore children’s perspectives on three types of feedback (Section 2.3), numeric, symbolic, and progress, in conversational tasks (RQ1).
- **Study 2 – Lab experiment:** We developed and integrated two "live", continuous feedback types (numeric and symbolic) into a voice-based sleep diary powered by LLMs. In a 1×3 within-subject design, we compared these feedback types to providing no feedback to assess their effect on response quality across questions (RQ2).
- **Study 3 – Field deployment:** We developed the continuous feedback mechanism to be more adaptive by combining numeric and symbolic feedback, and extended the system to an eight-day home study, using a mixed within- and between-subject design to examine whether the feedback can affect response quality over days (RQ3) and to capture children’s post-use perceptions (RQ4).

4 Co-design Study: Understanding Response Quality Feedback in Conversation

To design effective response-quality feedback for children, we first needed to understand how they interpret different feedback types during conversational interaction. We therefore conducted a co-design study in which children participated in small-group role-play activities simulating conversational interactions. The purpose of this study was not to generate final design concepts, but to uncover which forms of feedback children find intuitive, motivating, or confusing. These insights directly informed the selection and refinement of the feedback mechanisms evaluated in our subsequent laboratory experiment.

4.1 Study Setup

4.1.1 Participants. We recruited 14 children aged 7-12 years ($M = 9.07, SD = 1.49$), including eight boys and six girls. They were organized into five groups of two to three children each. Participants were recruited at a local public library and neighboring primary schools. All were native Dutch speakers. Co-design sessions took place in the public library or at our university. Parents could choose the location most convenient for them.

The study received ethics approval from our university’s Ethical Review Board. Participation was voluntary and required both parental consent and children’s willingness to join. As a thank-you, each child received a small set of Lego bricks (retail value: €9.99). A summary of group composition is provided in Table 1.

Table 1: Participant information for the participatory study

Group	Participants	Gender	Age
1	C1	boy	9
	C2	girl	9
	C3	girl	9
2	C4	boy	10
	C5	boy	8
	C6	girl	7
3	C7	girl	9
	C8	boy	9
	C9	boy	12
4	C10	boy	9
	C11	girl	12
5	C12	boy	7
	C13	girl	8
	C14	boy	9

4.1.2 Materials. To keep the role-play engaging and to encourage reflection, we prepared a set of playful, tangible materials.

Role masks. We created masks inspired by *Tom and Jerry* by Hanna and Barbera: Tom (asker) and Jerry (respondent). In addition, we created an Owl (a "wise helper" who evaluates Jerry’s answer).

Conversation prompts. We prepared five lighthearted, imaginative questions (e.g., "What would you do if you were Superman for a day?"). These prompts were intentionally unrelated to sleep so children could focus on understanding and comparing feedback mechanisms without introducing sensitive or personally revealing topics.

Feedback props. To explore children’s preferences for different types of response-quality feedback, we prepared physical representations of 1) *numeric feedback* (boards labeled 1-3, where 1 = poor, 2 = average and 3 = good), 2) *symbolic feedback* (three emoji faces, where "neutral face" is poor answer, "smile" is average, and "laugh" is good), and 3) *progress feedback* (a path-based board using a small figurine that moves 1-3 steps depending on answer quality).

Co-design materials. Children were also given clay, pens, paper, building bricks, and blank feedback cards to express or redesign their preferred forms of feedback. We used tangible materials instead of digital interfaces because physical props reduce cognitive load, support hands-on exploration, and avoid prematurely anchoring children to a specific interface style [37].

4.1.3 Procedure. After obtaining parental consent and child assent, a researcher explained the activity. Parents did not participate. Each group lasted around 30 minutes and included a role-play phase and a reflection & co-design phase (Figure 1).

Role-play phase: The Owl acted as a scaffold, helping children evaluate responses without needing to invent criteria. The researcher first demonstrated how the Owl determines whether an answer fits the question and then selects one of the three feedback types. This allowed the Owl-participant to focus on comparing feedback forms rather than creating judgment rules. In groups of three, Tom asked a prompt, Jerry responded, and the Owl provided feedback. For groups of two, Tom also acted as the evaluator.



Figure 1: Co-design activities.

this process using age-appropriate examples, so children could understand how to judge whether an answer fit the question. This scaffolding allowed the child in the Owl role to focus on comparing the feedback types rather than inventing judgment rules, ensuring meaningful participation in the evaluation task.

Reflection & Co-design phase: After role-play, researchers conducted semi-structured interviews to explore children’s preferences and interpretations of each feedback type. Children were then invited to create or re-design feedback using creative materials. Children used these materials in different ways: some built physical indicators of “good” or “bad” answers (e.g., towers or tokens), others drew new symbols or emoji-like expressions, and several redesigned numeric or progress feedback using shapes, colors, or story elements. The purpose of this activity was not to produce fully functional designs but to give children a concrete medium through which to express how feedback should look and feel. These co-design artifacts served mainly as prompts to support reflection; our analysis focuses on the observation and interview data.

4.2 Results & Design Implications

All participants expressed their preferences and opinions on the different forms of feedback. Only one of fourteen participants preferred the numeric feedback (number boards), while the majority preferred the symbolic feedback (smiley form). No participants chose progress feedback.

We conducted a hybrid thematic analysis combining deductive and inductive approaches [65]. The initial deductive code frame was derived from the three feedback types examined in this study (numeric, symbolic, progress; Section 2.3) and constructs used in prior work on children’s interpretation of feedback (e.g., clarity, emotional response, perceived fairness, usefulness for improving answers). Two researchers independently applied these codes to all transcripts while remaining open to inductive additions when children raised concepts not captured by the initial frame. Through iterative comparison, the researchers collaboratively refined a shared codebook and then independently recoded the full dataset. Discrepancies were discussed in reconciliation meetings until consensus was reached. Interrater reliability was high, with Cohen’s Kappa of 0.89 [80], indicating high agreement on the coded categories and consistency in how the data were interpreted.

Three high-level themes emerged, each with implications for designing response-quality feedback in children’s voice-based diaries.

T1 Timeliness of feedback: Participants consistently preferred feedback that reflected the quality of their *current* answer rather than aggregated performance across multiple questions. As C5 explained, “*I want to know if this answer is good now, not after all the questions.*” C1 also highlighted its corrective value: “*If it’s wrong now, I can fix it now.*” These preferences reinforce the potential of “live”, continuous response-quality feedback across questions to help children adjust their answers immediately, supporting higher-quality self-reports.

T2 Fun to interpret: Symbolic feedback, especially smileys, was described as more enjoyable and emotionally engaging than numbers. Children described smileys as playful and even socially interactive. As C7 said, “*The happy face makes me happy...*”, while C3 noted, “*It’s more fun than just numbers. It’s like the feedback is talking to me.*” Embedding emotional or playful elements in feedback may therefore help sustain children’s engagement over repeated diary use.

T3 Ease of understanding: Children found that progress feedback, which required them to interpret movement along a path, was difficult at the time of response. In contrast, the numbers and smileys were immediately clear. As C9 put it, “*The pig walking is funny. I don’t know what it means for my answer.*” C2 echoed this preference for simplicity, “*Numbers are easy. One is bad, three is good. Simple!*” These findings suggest that progress feedback incurs additional cognitive effort that children cannot process at the same time they are responding.

In summary, these findings indicate the value of feedback that is *delivered in real time*, *emotionally engaging*, and *simple*. The co-design study also clarified why progress-style feedback is less suitable for in-the-moment conversational tasks with children, leading us to focus on numeric and symbolic forms of momentary feedback for subsequent lab and field studies.

5 User Study: Evaluating Response Quality across Questions

To address RQ2, we conducted a second study to examine how continuous feedback influences *response quality across questions* within a single reporting session. This study used a 1×3 within-subject design, comparing a voice-based sleep diary with numeric, with symbolic feedback, and with no feedback as a baseline. Before running the experiment, we first designed a chatbot-based voice sleep diary capable of delivering continuous feedback.

In the following section, we describe the design of the voice-based sleep diary with integrated feedback, followed by the experimental setup used to evaluate its effect on response quality across questions.

5.1 Design and Implementation

5.1.1 Response-quality Feedback Design Rationales. Drawing on the findings from the co-design study (Section 4.2), we derived three evidence-based design rationales (DR1-DR3) to inform the development of the response-quality feedback design. **DR1** stresses **timeliness**: children wanted to know whether their current response was “good now,” not after completing the diary, motivating a live, continuously updated scoring mechanism. **DR2** highlights **affective engagement**: children described smileys fun, friendly, and socially expressive, suggesting that symbolic cues can help sustain motivation. **DR3** reflects **cognitive accessibility**: children preferred feedback that was easy to interpret while speaking, supporting the use of numeric indicators that require minimal effort to understand.

In summary, these themes directly informed DR1–DR3, ensuring that timeliness, affective engagement, and cognitive clarity were grounded in children’s preferences, and ensuring that the system reflects children’s own interpretations and needs rather than assumptions imposed by designers.

5.1.2 Design. To operationalize **DR1** (timeliness), **DR2** (affective engagement), and **DR3** (cognitive accessibility), we designed a feedback system that reflects the quality of a child’s answer as they speak, and presents it through two visual forms (symbolic and numeric). The **DR1** is articulated through three components: 1) a real-time scoring mechanism, 2) symbolic and numeric visual indicators, and 3) controls ensuring safe and valid feedback.

Computing response quality: We adopted the Response Quality Index (RQI) from prior work [24], which evaluates each answer along three dimensions: *information units* (count of distinct details), *relevance* (on-topic vs. off-topic), and *clarity* (clear vs. unclear). Each dimension is scored as a count or binary (information units = count, relevance = 0/1, clarity = 0/1). These are multiplied to form a single score [24]:

$$\text{RQI} = \text{Information units} \times \text{Relevance} \times \text{Clarity} \quad (1)$$

This formulation ensures that diary entries must be detailed and interpretable: off-topic or unintelligible responses receive an RQI of 0, consistent with clinical data quality standards. For example, a detailed but irrelevant answer (“I went to bed” or an unclear response (“I eat eat eat eat”) to “What did you eat last night?”) receives an RQI of 0 because the response cannot support diary-based monitoring. The index therefore penalizes responses lacking relevance or clarity regardless of length. This strictness is consistent with how unusable diary entries are treated in prior work [24]. Additionally, RQI was moderately correlated with response length ($\rho = .29, p < .001$), confirming that it rewards elaboration without inflating scores for long but low-quality answers.

Real-time scoring using LLM support: Children’s speech was transcribed and streamed in real time to two engines: a server-side Python implementation with *spaCy* NLP library, which counts

information units, and GPT-3.5, which classifies *clarity* and *relevance* (see Appendix C and D). The RQI is updated continuously and visualized instantly in the interface (Figure 2), enabling children to refine their answer while speaking.

Safety and validity: GPT-3.5 produced only two classification outputs (“relevant/not relevant” and “clear/not clear”); all diary questions were predefined. This limited GPT-3.5 to a controlled classification role, preventing hallucinated dialogue or unintended feedback. No inappropriate or unexpected outputs occurred. To verify accuracy, two researchers independently coded all responses: agreement with GPT-3.5 was high (relevance: Cohen’s $\kappa = .82$; clarity: $\kappa = .86$), supporting reliable use for live feedback, while offline analyses relied on human-coded ground truth.

Forms of response-quality feedback: In line with **DR2** (affective engagement) and **DR3** (cognitive accessibility), we implemented two visual forms of feedback: 1) *Symbolic feedback*, using a smiley face that changed based on score thresholds: laughing face (score > 3, good), smiling face (2-3, medium) and neutral face (< 2, poor); and 2) *Numeric feedback*, showing the real-time quality score as a number.

To support motivation without inducing pressure or discouragement, symbolic feedback used only neutral-to-positive expressions, avoiding sad or negative emojis, allowing the diary to guide improvement while maintaining a playful tone. Figure 3 shows how feedback updates continuously as children speak: the transcript appears at the bottom of the screen while the smiley or numeric indicator adjusts in parallel.

Clarifying feedback modalities: While the chatbot also provided spoken acknowledgments (e.g., thanking the child or introducing the next question), these utterances were not evaluative and did not vary with response quality. The only performance-related feedback manipulated in our studies was the live visual indicator (numeric or symbolic) linked to the RQI score. This ensured that the experimental conditions differed only in evaluative feedback modality, while spoken interaction remained constant to support a natural conversational flow.

5.1.3 Implementation. The diary was implemented on Android smartphones using Java and consisted of three integrated components: a dialogue module, and speech and language processing.

Dialogue module: The chatbot delivered 12 sleep-related questions adapted from the Consensus Sleep Diary [21], and refined with input from clinicians and pediatricians for age appropriateness (see Appendix A) [24]. Each turn followed a consistent structure: the chatbot asked a question, the child responded, and the chatbot provided a brief acknowledgment before moving to the next question. All interaction data were stored as text on Firebase for later analysis (Figure 4).

Speech and language processing: The diary uses a GPT-3.5 powered dialogue. In each turn, the chatbot 1) responds to the child’s previous answer and 2) asks the next diary question. The system integrates: 1) *Speech input* via Google Cloud Speech-to-Text API v1 for voice recognition and transcription; 2) *Speech output* via Google Cloud Text-to-Speech API v1 for spoken responses; and 3) *language understanding and generation* via GPT-3.5 to produce empathetic, child-friendly responses while preserving the diary’s

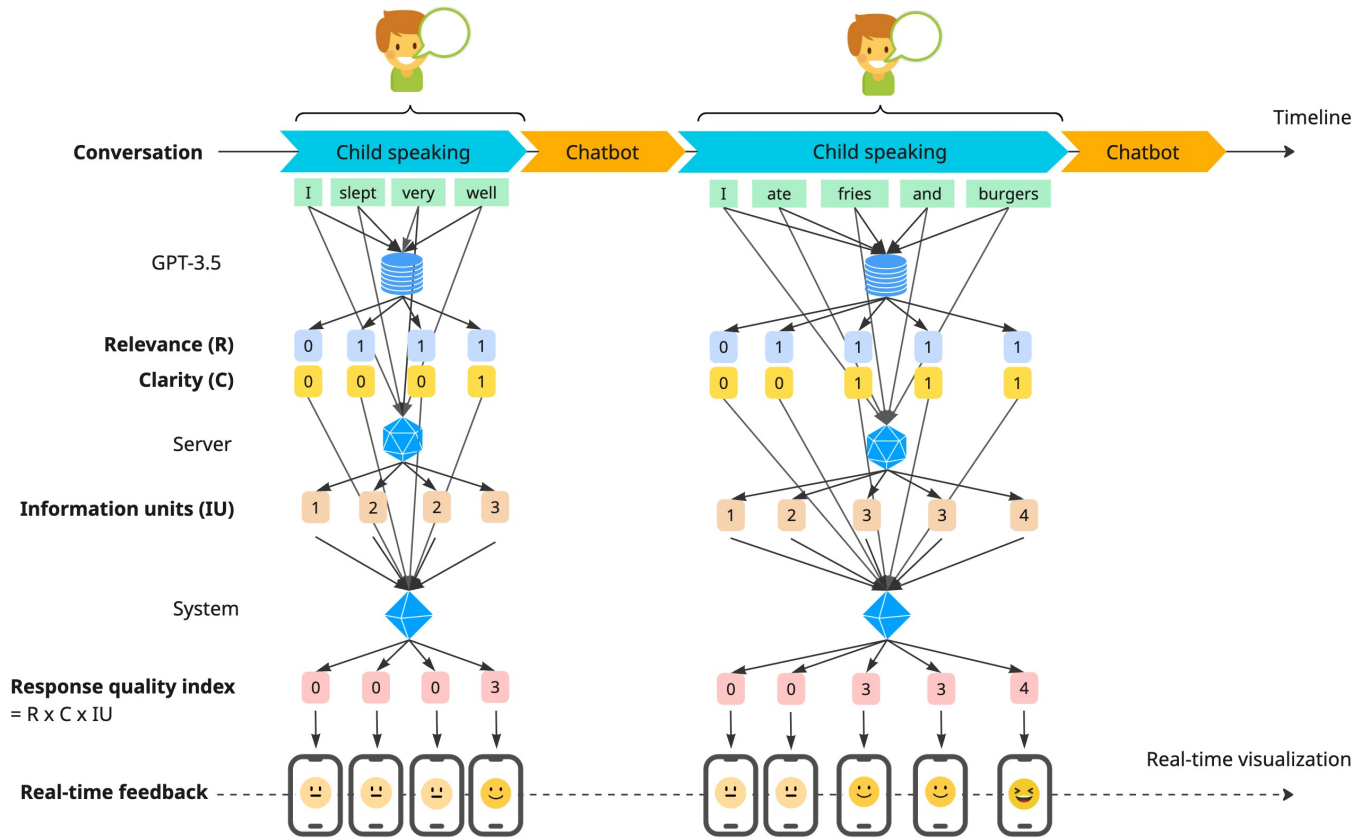


Figure 2: The mechanism of continuous response-quality feedback.

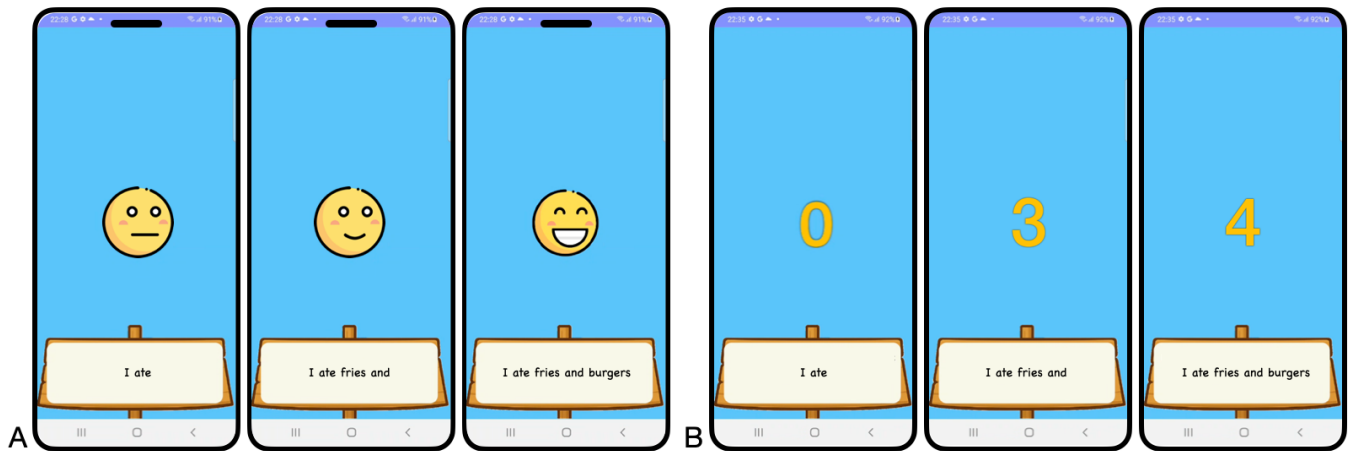


Figure 3: The "live", continuous feedback in two forms: (A) Smiley feedback, and (B) Numeric feedback.

structure. Previous research confirms that Google’s speech recognition achieves sufficient accuracy for children’s speech [25].

Speech output was rendered using an American English neural male voice ("en-US-Neural2-J"), optimized for playback on smartphone speakers. To keep the voice warm, clear, and easy for children to follow, we generated audio using SSML with gentle variations

in prosody, such as a slightly slower speaking rate and mild pitch adjustments that prevented a flat or monotone delivery. Positive or emotionally salient words were spoken with moderate emphasis, and short interjections (such as, "oh", "wow", "hey") were rendered expressively. These settings produced a friendly, engaging tone while ensuring that all children heard consistent content. The

```

Child: really good
01-06-2025 16:03:28 Chatbot: That's awesome! Keep up the good vibes!
Did you have a good sleep last night?
answer this question for:0.0s

Child: 10 minutes
01-06-2025 16:03:41 Chatbot: That's awesome!
What did you do before you went to bed last night?

Child: no I did have a good sleep tonight
01-06-2025 16:03:55 Chatbot: That's wonderful to hear! A good night's sleep is so important.
What did you eat before you went to bed last night?

Child: I was reading a book
01-06-2025 16:04:13 Chatbot: Reading a book sounds like a lot of fun!
What did you drink before you went to bed last night?

```

Figure 4: The conversation between children and the chatbot, which was stored on Firebase.

prompt templates used for generating chatbot responses are provided in Appendix B. Figure 5 illustrates the overall system architecture.

5.2 Methods

We compared children’s performance with the three types of feedback (Smiley, Numeric, and No Feedback), in a 1×3 within-subject experiment. Each participant completed all three versions of the diary, after which we conducted an individual interview to gather qualitative feedback on their experience.

5.2.1 Participants and Ethical Considerations. We conducted a power analysis using G*Power 3.1 [46] to determine the required sample size for a repeated measures ANOVA with one within-subject factor (Feedback Type: smiley, numeric, none) and three levels. We assumed a medium effect size ($f = 0.25$), based on Cohen’s conventions [29], and a desired statistical power of 0.80. The analysis indicated that at least 34 participants would be needed. To allow for potential dropout or unusual data, we recruited 36 children aged 7-12 years ($M = 9.19, SD = 1.60$), including 20 girls and 16 boys, ensuring sufficient statistical power ($>.80$) for detecting within-subject differences.

Participants were recruited in collaboration with a local public library and neighboring primary schools by approaching children and their guardians in person. Six children had previously taken part in the co-design workshop; analyses did not indicate systematic differences between these six children and other children. None of them had previous experience with writing or keeping sleep diaries before. All children spoke English (see Table 2 for demographic characteristics and smartphone use frequency). According to parents, four children had no prior smartphone experience.

Since the study involved reading, listening, and speaking on a smartphone, we recruited children with age-appropriate abilities in these areas. Because our focus was on how the continuous response quality feedback influences self-reporting behavior, we did not restrict participation based on whether the child had a sleep disorder.

Ethics approval was obtained from our university’s Ethics Review Board. To safeguard privacy, all personal data, including demographic information, audio recordings, and conversations synchronized on Firebase, were deleted immediately after the study. Participation was voluntary, requiring both parental consent and children’s assent.

Table 2: Demographic characteristics and smartphone use frequency

	Full sample
Age in years	9.19 (1.60)
Female	20 (56%)
<i>Smartphone use</i>	
Daily	12 (33.33%)
Weekly	13 (36.11%)
Monthly	7 (19.44%)
Never	4 (11.11%)
<i>N</i>	36

As a token of appreciation, each child received a set of LEGO™ bricks (retail value: €9.99). In addition, to provide meaningful reciprocation beyond material rewards, a playful follow-up hands-on session was offered to show participants how to build their own chatbot. Participation in this session was voluntary. All children chose to participate, and consent forms were signed before participating.

5.2.2 Procedure. The session started with an introduction to the purpose and structure of the activities. The children were first guided through trial versions with both types of feedback (smiley and numeric) to familiarize themselves with the self-reporting process. This step allowed participants to practice answering questions, explore the interface, and gain confidence prior to the main task.

In the main task, each participant completed the sleep diary under all three feedback conditions (Feedback Type: smiley, numeric, none) in a fully counterbalanced order. We used the six possible sequences of the three conditions ($3! = 6$), ensuring that each type of feedback appeared equally often in the first, second, and third positions among participants. In addition, a five-minute break involving light-play activities (e.g., tablet video games) between conditions was implemented to mitigate fatigue and recency effects.

After completing all three conditions, we conducted a short semi-structured interview to gather exploratory insights about participants’ experiences. These qualitative data were gathered to contextualize the quantitative results and inform later design decisions.

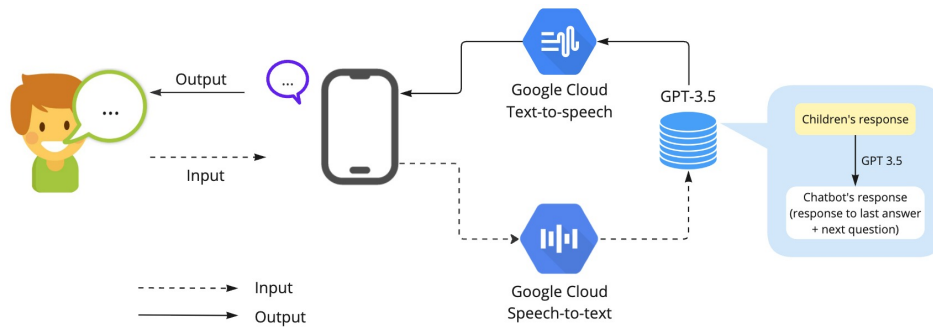


Figure 5: System structure for the voice-based sleep diary. The system includes three APIs to process the conversational flow. Solid arrows represent output flow; dashed arrows represent input flow.

Each session lasted approximately 40 minutes. Figure 6 shows a child interacting with the sleep diary in the smiley feedback condition.



Figure 6: A child was interacting with the version of the sleep diary with the smiley response-quality feedback.

5.2.3 Data Collection. All diary interactions were automatically saved as text files on Firebase, labeled by participant ID and feedback condition (e.g., 'C1_smiley.txt' for Participant 1 using the smiley feedback version). Each file contained the full transcript of the conversation.

To evaluate children's responses, we applied the same RQI framework used in the system: *Information units* (spaCy-extracted counts), and *Relevance* and *Clarity* (binary judgments). For analysis, relevance and clarity were manually coded by two independent researchers (Cohen's $\kappa = 0.89$; 95% CI[0.76, 0.95]) and multiplied with information units to generate ground-truth RQI scores [24]. These offline human-coded scores, rather than system-generated scores, were used for all statistical analysis.

The real-time RQI shown to children was computed by the LLM, while the RQI used for analysis was human-coded to avoid propagating potential model errors. Post-hoc comparison showed strong correspondence between the two (*Spearman's* $\rho = .78$, $p < .001$),

indicating that the feedback children saw closely reflected human judgments while ensuring that reported effects were based on validated scores.

5.3 Results

The system functioned smoothly for all participants. Data from all 36 children were included in the analysis. Each participant completed all three feedback conditions (smiley, numeric, none), resulting in a complete dataset of self-reports for all conditions.

5.3.1 Overall Response Quality. Descriptive statistics are shown in Table 3. The mean response quality scores (RQI) were highest in the numeric feedback condition ($M = 3.98$, $SD = 0.86$, $95\%CI = [3.69, 4.27]$), followed by the smiley feedback condition ($M = 3.75$, $SD = 0.95$, $95\%CI = [3.43, 4.08]$), and lowest in the no-feedback condition ($M = 2.81$, $SD = 0.73$, $95\%CI = [2.57, 3.06]$). A one-way repeated-measures ANOVA showed a significant effect of Feedback Type on RQI, $F(2, 105) = 18.95$, $p < 0.001$, $\eta^2 = 0.27$. Tukey's HSD post hoc tests showed that both the smiley ($p < 0.001$) and numeric feedback ($p < 0.001$) yielded higher RQI than no feedback. There was no significant difference between the smiley and numeric feedback conditions ($p = 0.51$).

5.3.2 Influence on Response Quality across Questions. To examine how feedback influenced quality during a session, we built a linear mixed-effects model (LMM) in R using the lme4 package [10]. LMMs are used to analyze clustered data, such as repeated observations from the same participants over time [49]. Given our design, with multiple observations per participant, this model was well suited to account for both random and fixed effects. In this model, we included *Question* and *Feedback type* as fixed effects, and *Participant* as the random effect. The linearity, normality, and homoscedasticity assumptions were assessed through residual diagnostics and found to be sufficiently met.

Table 4 shows the model results to examine whether continuous feedback influences changes in children's response quality over the questions answered. Using the no-feedback condition as the baseline, the results did not show a significant change in mean response quality (RQI) between the questions in the no-feedback condition ($b = -0.01$, $p = 0.49$). However, significant interaction effects indicated that both smiley ($b = 0.18$, $p < 0.001$) and numeric feedback ($b = 0.16$, $p < 0.001$) led to significant improvements in

Table 3: Mean response quality (RQI) and children’s preference

Measure	Smiley Feedback	Numeric Feedback	None Feedback
RQI (Mean ± SD)	3.75 (0.95)	3.98 (0.86)	2.81 (0.73)
ci_lower	3.43	3.69	2.57
ci_upper	4.08	4.27	3.06
Preference Rate	47.22%	52.78%	0

the mean response quality over time, compared to the no-feedback condition. These findings suggest that continuous response quality feedback supports a positive trajectory in children’s response quality across questions during self-reporting.

Table 4: Effects of feedback factors on predicting response quality. The model formula is $RQI \sim Question * Feedback + (1|ParticipantID)$, where RQI = response quality index, $Feedback$ = Feedback forms (smiley, numeric, and none), $Question$ =12 questions in the sleep diary, $ParticipantID$ = participant ID. * $p < 0.001$, * $p < 0.05$.**

Predictor	Estimate	Std. Error	df	t	p
(Intercept)	2.88	0.16	123.61	18.01	< .001***
Question	-0.01	0.02	1255.00	-0.69	.49
Feedback (Smiley)	-0.19	0.16	1255.00	-1.19	.24
Feedback (Numeric)	0.10	0.16	1255.00	0.62	.53
Question × Smiley	0.18	0.02	1255.00	7.94	< .001***
Question × Numeric	0.16	0.02	1255.00	7.44	< .001***

Figure 7 shows the interaction between question number and feedback form on response quality. Both smiley and numeric feedback produced a clear upward trajectory in response quality, whereas the no-feedback condition remained flat. This pattern aligns with the significant interaction effect in the model, indicating that continuous feedback supports improving response quality across questions during a session. The slight drop at Q7 reflects a structural shift rather than reduced engagement or question difficulty. The diary alternates between descriptive and non-descriptive items [24]. Q4-Q6 invite elaborated answers (e.g., “What did you do before bed?”), which naturally yield multiple information units. Q7 (“How many times did you wake up during the night?”) abruptly switches to a categorical, non-descriptive response format. Children must shift from storytelling to providing a single numeric value, which inherently reduces information units and thus RQI. Because the question order followed the standard clinical sequence of sleep diaries, it was fixed rather than randomized.

5.3.3 Children’s Feedback Preference. Although RQ2 focused on performance metrics, we also collected brief post-session feedback to contextualize the results. Preference data (Table 3) showed that 17 children (47.22%) favored smiley feedback and 19 (52.78%) favored the numeric feedback; no one preferred the no-feedback condition.

Similarly, we applied a deductive thematic analysis to the brief post-session statements. Two independent coders coded all the data using a codebook and refined the coders for this context. Disagreements were discussed until consensus was reached, with the first

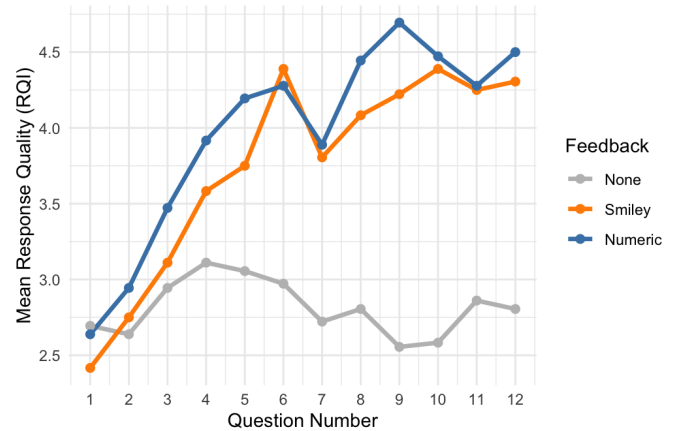


Figure 7: Tendency in response quality across questions in one session.

author overseeing the process to ensure consistency with the research objectives. Given that children’s comments were generally short and direct, the dataset was relatively small and straightforward to code. Inter-rater reliability was high ($Cohen's\kappa = 0.91$ [80]). From this analysis, two themes emerged, which we present below:

Symbolic feedback is fun but harder to interpret: Children who liked the smiley feedback emphasized its playfulness, e.g., “I like smile because it is fun. I always tried to get more smiles in the diary.” (C12), but some preferred numbers because they found smiley meanings less clear: “I choose number. Because I have to interpret the meaning of smile in the chatbot.” (C21).

Numeric feedback is easy and motivating. Many children found numbers quicker to understand, e.g., “Number is easy. I can understand it quickly.” (C26), and some linked higher scores to feelings of achievement: “I like the number. The higher number can make me happy.” (C35).

These insights help explain why both types of feedback performed similarly in terms of RQI, yet preferences varied.

5.4 Discussion

To address RQ2, our findings show that both smiley (symbolic) and numeric feedback improved children’s response quality across questions within a single session compared to the no-feedback condition. This effect was evident not only in the overall response quality scores, but also in the upward trend across the 12 questions.

In contrast, children in the no-feedback condition did not show such an upward trend.

These findings suggest that continuous feedback not only boosts immediate performance, but also encourages a sustained process of improvement within a reporting session. This aligns with previous work showing that continuous feedback can facilitate both engagement and self-regulation during interactive tasks [58, 66, 75, 123]. In our context, smiley feedback appeared to engage children emotionally and playfully, while numeric feedback provided clear performance benchmarks and supported goal orientation. We did not examine age-related differences in our sample, but developmental research suggests that: younger children may be more motivated by socially expressive cues, while older children increasingly value structure, and self-monitoring [39, 51, 84]. Investigating how feedback preferences and performance vary across age groups therefore represents an important direction for future work.

Children's subjective comments reinforce these dual roles. Those who preferred smiley feedback described it as "fun" and motivating, often aiming to "collect more smiles", while those who favored numeric feedback valued its immediacy and precision. We interpreted these behaviors as game-like engagement rather than pressured performance. When children extended their answers, they did so by adding additional relevant details (e.g., explaining evening routines), not by inventing or exaggerating information. Their responses remained consistent with the diary questions, and no participants expressed anxiety about achieving higher scores. The feedback was intentionally framed as supportive rather than evaluative, with neutral-to-positive cues and no penalties for low scores. Even so, this design choice introduces an important consideration: future systems should balance motivation with honesty, for example by capping scoring gains or incorporating reminders that truthful reporting matters more than "winning".

Interestingly, these preferences contrast with those of the co-design study (Section 4.2), where most children favored the smiley feedback. In Study 2, numeric feedback was slightly more popular. Although this difference should be interpreted cautiously due to sample size differences (14 vs. 36), one possible explanation is that the self-reporting task in Study 2 imposed higher cognitive demands than the role-play activity in Study 1. Under greater cognitive load, children may prefer feedback that is more direct and quicker to interpret (numeric) over feedback that was more playful but required additional interpretation (smiley).

This shift has important design implications. It suggests that children's preferences are context-dependent and may vary with task complexity and cognitive load. These insights informed our design for Study 3, where we developed a combined feedback system that integrates symbolic and numeric cues, providing both engagement and clarity during extended, real-world use.

6 Field Study: Evaluating Sustained Quality Over Time

Building on the insight that affective engagement and cognitive clarity become more or less important depending on the reporting context, Study 3 explores a combined feedback approach that dynamically integrates both symbolic and numeric cues, with the

goal of maintaining response quality over repeated diary use in a real-world setting.

We conducted an 8-day field study in which children used a voice-based sleep diary at home under two conditions: 1) with continuous response quality feedback (combined: smiley + numeric), and 2) without feedback (baseline). This setup allowed us to examine the second challenge, *response quality over days*, and to explore children's perceptions after extended use.

6.1 Design and Implementation

6.1.1 Iteration 2: Combined Sleep Diary with Continuous Feedback. Drawing on insights from Study 2, we iterated the sleep diary by combining smiley and numeric feedback within a single interface. The goal was to combine the emotional engagement of symbolic cues with the clarity and precision of numeric scores.

The updated diary retained the same deployment platform (Android deployment), backend infrastructure (Firebase), and core APIs (Google Cloud Speech-to-Text v1, Google Cloud Text-to-Speech v1, GPT-3.5) as in Study 2. The feedback mechanism remained identical in its calculation of response quality (information units, relevance, clarity), but now both feedback forms were presented simultaneously: a smiley icon indicating performance category (good, medium, poor) with the numeric score displayed directly beneath it as an explanation. Figure 8 (A) shows the combined feedback interface. For the no-feedback condition (baseline), the interface presented the same conversational interaction without any performance indicator (Figure 8 (B)).

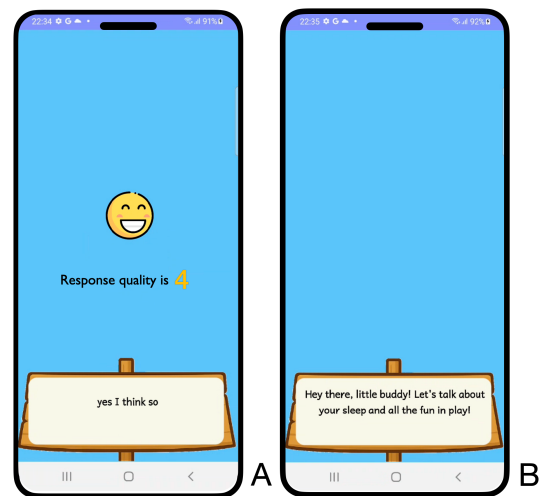


Figure 8: The sleep diary with the continuous response-quality feedback (A), and without feedback (B).

6.1.2 System Deployment for Home Use. Because each participant experienced both conditions in different time blocks, both feedback and no-feedback versions were integrated into a single app build. The system automatically tracked the number of diary entries completed in the current condition. After four completed entries, the interface switched to the other condition for the next use. This

ensured that the transition between conditions was seamless for the participants and reduced the need for manual intervention.

To support adherence, the app sent an automated reminder via Firebase Cloud Messaging at 9:00 PM each evening, prompting children to complete their daily diary. All data were stored on Firebase until the end of the study, after which they were deleted to protect privacy.

To prevent multiple entries on the same day, which could introduce invalid or duplicated data, the system included a built-in lock-out feature. Once a child completed the day's diary session, the login button was disabled until the following day. This ensured that each day's entry reflected a single, consistent self-reporting session.

6.2 Methods

6.2.1 Participants and Experimental Design. We recruited 24 children (aged 8-12, $M = 9.58$, $SD = 1.41$; 14 girls, 10 boys) from a local public library and through neighboring primary schools. Participants in the field study did not participate in Studies 1 or 2. They did not have previous experience writing or keeping sleep diaries. All participants spoke English and, according to their parents, had prior experience with smartphones. We ensured that all participants had typical abilities in reading, listening, and speaking. As in Study 2, the inclusion criteria did not require children to have sleep disorders.

We adopted a within-subjects crossover design with two conditions: feedback vs. no-feedback. Each child used the sleep diary for 8 consecutive days: four days under one condition, immediately followed by four days under the other. The order was counterbalanced (AB/BA) between participants to control for order effects. Figure 9 displays the experiment setting for children's self-reporting at home.

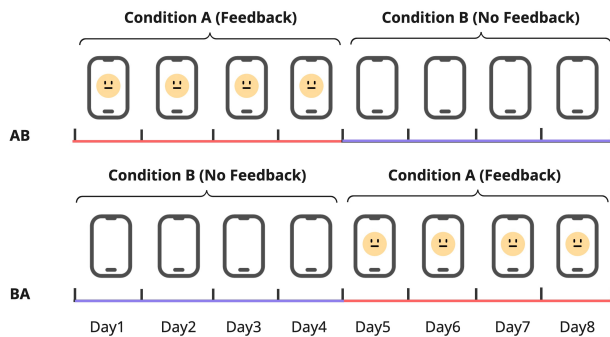


Figure 9: Experiment conditions.

This within-subjects crossover design ensured validity and feasibility for a child population. Each child served as their own control, reducing individual differences (e.g., vocabulary, technology familiarity, and engagement). Four days per condition captured short-term trends like adaptation or fatigue while keeping the study manageable for families. Counterbalancing order further minimized novelty and practice effects, so differences reflected the impact of feedback rather than reporting familiarity.

Our primary outcomes focused on whether feedback influenced children's multi-day reporting, assessed as: 1) both the mean response quality across four-day block, and 2) the day-to-day trend within each block.

6.2.2 Sample Size and Power Analysis. To ensure adequate power to detect an interaction between the condition and the day, we performed a simulation-based power analysis (R: lme4/simr) using the same linear mixed model planned for the analysis [69], with random intercepts and slopes for day-by-child. The variance parameters were informed by Study 2, with additional variance inflation to account for fluctuations between days. Assuming an expected slope difference of 0.5 points/day (Feedback vs. No-Feedback), the planned sample of $N=24$ children (AB/BA, four days per condition) yielded 98.4% power (95% CI: 97.4-99.1%) at $\alpha = .05$ for detecting the Condition \times Day interaction in 1000 simulations.

The study was approved by the Ethics Review Board of our university. All conversations were stored in Firebase and deleted immediately after the study. Participation was voluntary with parental consent and child assent required. As appreciation, each child received a LEGO™ set (retail value: €19.99), and a set of anime stickers (retail value: €5.00).

6.2.3 Procedure. The experiment was carried out over eight consecutive days and divided into three phases:

Phase I: Pre-training (with researcher). The first phase was conducted at our university (18 participants) or in a local public library (6 participants). After obtaining parental consent, we introduced the study procedures and provided each child with a smartphone with the pre-installed sleep diary app.

The children were trained to use both diary versions, feedback and no-feedback, and practiced until they could complete a full session independently. This familiarization ensured that they understood both the task and the interface before starting the home phase. Participants also received a printed quick-start guide with troubleshooting tips (e.g., login errors, app restarts). For any problems they could not be able to resolve themselves, we offered follow-up support via home visits or video calls.

Phase II: Home reporting (8 nights). Participants completed one self-reporting session of the diary per night for eight consecutive days. The app automatically managed the condition order: 1) four days in the first condition (feedback or no-feedback), 2) followed by four days in the other condition. The order was counterbalanced between participants (AB/BA) to control for sequence effects.

At 9 PM each night, the app sent a reminder notification via Firebase Cloud Messaging. If a session was missed, participants could complete it later; the app logged the actual date and time of each entry. Figure 10 shows a child using the diary at home.

Phase III: Post-assessment (with the researcher). After completing all eight sessions, the participants returned for a brief final meeting. We conducted a semi-structured interview to capture their experiences with both diary conditions, perceptions of feedback, and any factors influencing their reporting habits (e.g., preferred time, motivation). All smartphones were collected.

6.2.4 Data Collection. To observe the effect of feedback on children's sustained response quality, we collected three types of data:



Figure 10: A child was interacting with the sleep diary in no feedback condition at home.

reporting condition, response quality scores, and participants' subjective attitudes toward the two diary conditions.

Reporting condition: Each entry was tagged according to the condition under which it was completed, based on the conversation logs stored in Firebase. The entries in the no-feedback condition were coded as "0", while the entries in the feedback condition were coded as "1".

Response quality: Children's daily responses were evaluated using the same method as in Study 2. Specifically, we assessed three dimensions: *relevance* and *clarity*, which were coded by two independent raters, and *information units*, which were automatically extracted using the Python package *spaCy*. The interrater reliability for the manual coding was high (Cohen's $\kappa = 0.87$; $95\%CI : 0.73 - 0.94$), indicating strong agreement between coders. The continuous RQI was then calculated for each day by averaging the scores across the 12 diary questions. This provided a composite measure of each child's communicative effectiveness and informational richness over time.

Attitudes: Finally, to explore children's subjective experiences, we conducted thematic analysis of the post-study interview transcripts. The coding focused on their perceptions of the two conditions, their interactions with the chatbot, and their expectations for voice-based diary tools in the future.

6.3 Results

The app worked smoothly for almost all participants during the 8-day deployment. One child missed several days due to school commitments, but later resumed and completed the study, so their data were included in the analysis.

To answer RQ3 (response quality over days), we first present descriptive statistics of children's reporting behavior across conditions, followed by linear mixed-effects models to examine the changes in response quality over time. To answer RQ4 (children's perceptions of feedback), we report findings from the thematic analysis of post-study interviews.

6.3.1 Overall Response Quality. Table 5 summarizes the data on children's response quality (RQI) across the two experimental sequences. In the AB sequence (feedback first), mean RQI was higher during the initial four-day feedback block ($M = 2.65, SD = 0.25, 95\%CI = [2.49, 2.81]$) than during the subsequent no-feedback block ($M = 2.00, SD = 0.20, 95\%CI = [1.88, 2.13]$). In the BA sequence (no feedback first), the children began with a lower mean RQI ($M = 1.66, SD = 0.17, 95\%CI = [1.55, 1.77]$) but showed higher quality once feedback was introduced in the second block ($M = 2.98, SD = 0.44, 95\%CI = [2.70, 3.26]$). Across both sequences, periods with feedback consistently yielded higher mean response quality than periods without feedback, suggesting a robust overall benefit of feedback in supporting response quality during self-reporting. All values were computed at the child level (i.e., averaging each child's responses within a four-day block) to avoid inflation from repeated daily entries.

6.3.2 Influence on Response Quality Over Days. To test whether feedback affected the quality of response of children over time, we fitted a linear mixed effects model (LMM). *Participants* were included as random effects to account for individual differences, and *day* was modeled to capture repeated measures over the eight-day period. Fixed effects included *condition* (Feedback, vs. no-feedback), *sequence* (AB vs. BA), *period* (first vs. second block of four days), and their interactions. The dependent variable was the response quality (RQI) of each diary entry.

The model diagnostics indicated that the assumptions of normality and homoscedasticity were met, with the residuals showing no major deviations and no influential outliers. A likelihood ratio test confirmed that including the $Condition \times Day$ interaction significantly improved model fit ($\chi^2 = 78.41, p < .001$).

The results of the model (Table 6) revealed a strong interaction between $condition \times Day$ ($b = 0.36, SE = 0.04, SE = 0.03, t(141.37) = 10.11, p < .001$), including that the effect of feedback depended on the reporting day. Specifically, in the feedback condition, RQI increased by 0.21 points per day, while in the no feedback condition, RQI showed no comparable improvement. Other effects were also significant. The children in the BA sequence (starting without feedback condition) began with a lower RQI than those in the AB sequence ($b = -0.75, SE = 0.11, t(64.45) = -6.77, p < .001$). RQI was also higher in the second block compared to the first ($b = -0.65, SE = 0.17, t(155.60) = -3.85, p < .001$).

Figure 11 illustrates these trajectories. In the AB sequence, the quality of the children's response increased steadily during the first four days with feedback, but declined once feedback was removed. In the BA sequence, response quality remained low and flat during the initial no-feedback block but rose consistently once feedback was introduced, peaking on day 8.

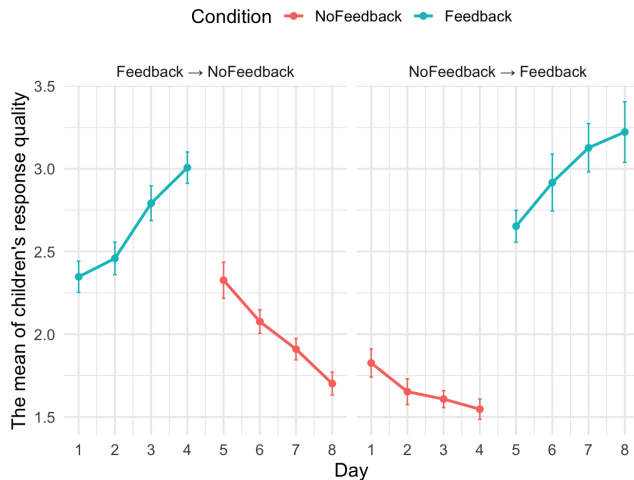
Together, these findings demonstrate that feedback not only increases average response quality but also supports sustained

Table 5: Descriptive statistics of Response Quality Index (RQI) by the sequence of feedback and none feedback in the sleep diary.

Sequence	Period	Condition	Mean RQI	SD	ci_lower	ci_upper
AB	First	Feedback	2.65	0.25	2.49	2.81
AB	Second	No Feedback	2.00	0.20	1.88	2.13
BA	First	No Feedback	1.66	0.17	1.55	1.77
BA	Second	Feedback	2.98	0.44	2.70	3.26

Table 6: Effects of model factors on multi-day response quality. The model formula is $RQI \sim Condition * Day + Period + Sequence + (1 + Day | ParticipantID)$, where RQI = response quality, $Condition$ = feedback vs. no feedback, Day =repeated days, $Period$ =First or Second, $Sequence$ =AB vs. AB, $ParticipantID$ = participant ID, $Condition * Day$ = the daily trajectory differ between feedback vs. no-feedback. * $p < 0.001$, * $p < 0.05$.**

Variables	Estimate	SE	df	t	p
(Intercept)	2.77	0.15	178.82	18.96	$2e - 16$ ***
Condition	-0.65	0.17	155.60	-3.85	0.000171 ***
Day	-0.15	0.03	135.98	-5.24	$6.02 - 07$ ***
Period	0.21	0.08	141.37	2.54	0.01 *
Sequence	-0.75	0.11	64.45	-6.77	$4.50e - 09$ ***
Condition:Day	0.36	0.04	141.37	10.11	$2e - 16$ ***

**Figure 11: The patterns of children's response quality over days in difference conditions (feedback vs. no-feedback).**

improvement over multiple days of reporting. Without feedback, children's responses either stagnated or declined.

6.3.3 Children's Perspective of Feedback. To explore how feedback influenced children's experience during multi-day reporting, we conducted a thematic analysis of the semi-structured interviews [65]. The same analytic approach used in Studies 1 and 2 was applied here: two coders iteratively refined a shared codebook and independently coded all data, achieving high agreement (Cohen's

$\kappa=0.87$) [29]. The process yielded three themes regarding the role of feedback in the sleep diary.

Feedback as guidance for answering. Many children viewed feedback as an essential aid during self-reporting. Eight children emphasized that feedback helped them know if their answers were "good enough" and motivated them to provide more complete responses. As C11 explained, "I need the feedback in the App. I hope my answer can be a higher score and smile each time." Several children likened the experience to familiar learning platforms. For example, C8 noted: "It looks like the Duolingo I use every day." These comments suggest that feedback was not only perceived as evaluative, but also a form of guidance or scaffolding that aligned with practices they already associated with learning.

Feedback makes self-reporting more engaging. Real-time visual cues, especially smileys and changing scores, were described as fun and motivating. Six children said they enjoyed trying to "collect" smileys during conversation, even after they had technically answered the question. As C3 shared, "Sometimes I want to get a smiley by keep talking with it even after I have finished answering." Beyond single sessions, four children described how feedback motivated them to return daily, often comparing scores over days or even with siblings. C22, for example, reflected, "I'm happy if I can get more smiley than my brother. If I got less, I hope I can answer again. But the system is locked after I finish the game today." These accounts highlight the playful, game-like role of feedback in sustaining engagement over multiple days.

Desire for mixed and cumulative feedback. Although continuous feedback was valued to make the diary more interactive, several children expressed a desire for additional types of feedback that captured progress over days. Five children suggested that in addition to immediate cues, the system could also display a daily or cumulative summary. For example, C22 remarked, "Real time is fun when I am answering the question, but I hope I can see a general change in my answers in all entries." This indicates that the children sought both immediate reinforcement and a broader perspective on their performance over time, pointing to the potential value of adaptive or layered feedback mechanisms.

6.4 Discussion

6.4.1 Towards sustained and improved response quality over time. To address RQ3, we examined whether continuous feedback could sustain or improve response quality over multiple days of reporting. Across both sequences, periods with feedback consistently yielded higher response quality than periods without feedback, and feedback appeared to shape the tendency of children's response quality over days. In the AB sequence, the quality increased during the feedback period, but declined after the feedback was removed. In

the BA sequence, each period comprised only four days of reporting, so the observed trend should be interpreted as preliminary rather than conclusive. Future work should test whether these benefits persist for weeks or months, which is a requirement for some diary studies or practical applications.

Previous diary studies with children often acknowledge that response quality declines over time, even when interventions or structured prompts are used [24]. Very few have demonstrated that quality can be maintained in repeated entries. In this study, feedback was associated with sustained or improved quality over a four-day reporting period, suggesting its potential to counteract the generally observed decline. However, because each condition lasted only four days, it is unclear whether this effect would persist over longer periods.

6.4.2 Towards motivating and engaging self-reporting. To address RQ4, we explored children's subjective experiences and motivational responses to feedback during multiday home use. Similarly to gamified learning and pediatric mHealth interventions [26, 28, 33, 119], children described feedback as helping them know whether their responses were "good" and making the process more engaging. Competitive or game-like elements (e.g., smileys, scores) appeared to encourage repeated participation [55]. Consistent with children's self-reported experiences, the conversation logs also provided behavioral evidence of this effect. In the no-feedback condition, children often gave minimal answers (e.g., "9 PM" when asked about bedtime). In contrast, in the feedback condition, responses were more elaborated (e.g., "9 PM, because I played computer games with my brother before that"), suggesting that feedback encouraged children to add contextual details to achieve higher scores and smileys. This highlights how feedback not only enhanced their motivation, but also supported richer higher-quality reporting. However, it remains unclear whether this motivation would persist beyond a brief novelty period.

Some children expressed interest in daily summaries or general performance indicators, echoing prior research showing that a combination of real-time and cumulative feedback can support long-term engagement [8, 36, 64, 98, 121]. Integrating these features may help feedback systems remain effective in contexts that require sustained self-reporting.

A concern with any feedback-driven system is whether children may feel compelled to exaggerate or extend their responses to obtain positive feedback. In our field study, a few children expressed that earning more smileys or higher scores felt "like a game." We interpreted this as playful motivation rather than pressure, as responses remained on-topic and we did not observe fabricated or contradictory information.

Together, these findings show that combined, "live" and continuous feedback can sustain both quality and motivation during children's self-reporting, though refinements such as cumulative feedback may be needed for long-term use. In the following general discussion, we reflect on how these insights advance the design of child-centered voice-based diaries across contexts.

7 Design Implications

Drawing on the three studies, we propose broader design implications for child-centered self-report systems. These insights extend

beyond the sleep diary and can inform the design of technologies for children in health, education, and everyday life. Taken together, the studies reveal that effective systems must balance engagement with clarity, adapt feedback to the reporting context, and combine real-time and cumulative strategies to maintain motivation and quality over time.

7.1 Balancing Affective Engagement with Cognitive Clarity

Our findings highlight the need to balance emotional engagement and interpretive clarity. Playful elements such as gamified icons, stories, or characters can boost children's enjoyment and motivation [35, 73, 124, 125] but they also introduce additional interpretive demands. This added cognitive effort can become a burden when the reporting task itself already requires sustained attention. In contrast, straightforward and easy-to-interpret feedback, such as numeric scores, reduces ambiguity and allows children to focus on the core task. Designers should therefore resist the assumption that "more fun" always leads to better engagement. Instead, the goal should be to integrate affective engagement without sacrificing clarity so that children are motivated and able to interpret their own performance.

7.2 Adapting Feedback to Context and Task Demands

Children's preferences and the effectiveness of feedback are not fixed, but shift depending on the context. In playful or role-play settings, children may prefer symbolic or narrative feedback because they have the time and cognitive space to interpret it. In contrast, during cognitively demanding or routine tasks, such as daily journals at home, they benefit more from direct, unambiguous cues that minimize extra effort. This suggests that feedback design should be sensitive to context: playful elements can sustain engagement in light-tough tasks, whereas clarity and directness are more critical when reporting is embedded in daily routines or is of clinical significance. Ideally, future systems should be combined, dynamically adjusting the style of feedback to the child's context and workload.

7.3 Supporting Sustained Engagement through Multilayered and Multimodal Feedback

Sustaining children's motivation and response quality over time requires more than "live", continuous cues alone. This feedback helps children tune individual responses in the moment, but participants in our field study also expressed interest in broader indicators, such as daily summaries or progress trends. This suggests that systems may benefit from combining immediate guidance with cumulative, longer-term feedback. This multilayered structure is particularly relevant for chronic health contexts, where both adherence and sustained quality are essential.

Our findings also point to opportunities for integrating multiple modalities of feedback delivery. During the diary interactions, children received spoken responses from the chatbot while simultaneously viewing visual feedback on the interface. Prior work

shows that combining auditory and visual cues can improve comprehension, reduce cognitive load, and support sustained attention, particularly for younger users with developing working memory [4, 12, 53, 79, 113]. Visual indicators provide stable, easy-to-reference information about performance, while verbal acknowledgments support social engagement and emotional reassurance. When paired, these modalities can help children interpret feedback more effectively and maintain focus as they refine their responses in real time.

Overall, our results suggest that children may benefit most from feedback systems that are both multilayered (immediate + cumulative) and multimodal (visual + verbal). Future research should examine how different combinations of modalities and temporal layers shape children's understanding of feedback, sustained engagement, and the long-term quality of diary data.

8 Limitations and Future Work

We acknowledge limitations in this study. First, although the observed patterns are suggestive, a larger and more diverse sample could provide a more robust understanding of how age influences feedback preferences. Future research could benefit from a more detailed age categorizations and a longer field study period to capture how response quality evolves over time. Second, our comparison of combined feedback against a no-feedback baseline does not establish whether combined feedback outperforms symbolic or numeric feedback; future studies should examine these contrasts directly. Third, because our participants were healthy children without diagnosed sleep disorders, the applicability of the findings to clinical populations remains uncertain. Children with sleep disorders may show different preferences, challenges, or response patterns when using sleep diaries.

Finally, our study was conducted in a western European, primarily urban context, with a research team bringing positionalities from Industrial Design and Clinical Psychology, and cultural backgrounds spanning Asia and Europe. We provide detailed descriptions of the setting and procedures to support transferability of our findings, while acknowledging that different contexts and researcher identities may yield different insights. We encourage future work and researchers in diverse settings to build on and extend our findings.

9 Conclusion

This paper examined how feedback can enhance children's response quality in voice-based diaries through three complementary studies. A co-design study highlighted children's initial preferences for playful and symbolic feedback, underscoring the importance of affective engagement in child-centered diaries. A controlled user test showed that both symbolic and numeric feedback improve per-question response quality within a single session, while also revealing that children's preferences shifted towards more direct and cognitively efficient forms of feedback for more demanding tasks. Finally, an eight-day field test (Study 3) demonstrated that "live", continuous feedback helped sustain or even improve response quality in repeated diary use, and that children desired not only immediate cues but also cumulative forms of feedback to support longer-term motivation.

Our findings show that effective feedback design for children's self-reporting must balance emotional engagement with cognitive clarity, and adapt dynamically to both the immediate task and the longer-term reporting contexts. Beyond improving data quality, feedback can also motivate children to elaborate on their answers and remain committed to the reporting process over time. Our findings have useful applications in the child-centered design of voice-based diaries, which are particularly relevant in the study of child-computer interaction; moreover, our findings have various practical applications, such as in healthcare contexts where valid and reliable data are essential.

Acknowledgments

We used GPT-5 to assist with grammar checking and wording improvements. All ideas, analyses, and interpretations are those of the authors. This work was supported by the China Scholarship Council (CSC). We sincerely thank the children in the Netherlands who participated in our studies, as well as the anonymous reviewers for their constructive feedback.

References

- [1] Mohammad Aliannejadi, Monica Landoni, Theo Huibers, Emiliana Murgia, and Maria Soledad Pera. 2021. Children's Perspective on How Emojis Help Them to Recognise Relevant Results: Do Actions Speak Louder Than Words?. In *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval* (Canberra ACT, Australia) (CHIIR '21). Association for Computing Machinery, New York, NY, USA, 301–305. doi:10.1145/3406522.3446058
- [2] Maximilian Altmeyer, Vladislav Hnatovskiy, Katja Rogers, Pascal Lessel, and Lennart E Nacke. 2022. Here comes no boom! the lack of sound feedback effects on performance and user experience in a gamified image classification task. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–14.
- [3] Aaron D Arndt, John B Ford, Barry J Babin, and Vinh Luong. 2022. Collecting samples from online services: How to use screeners to improve data quality. *International Journal of Research in Marketing* 39, 1 (2022), 117–133.
- [4] Yigal Attali and Fabienne van der Kleij. 2017. Effects of feedback elaboration and feedback timing during computer-based practice in mathematics problem solving. *Computers & Education* 110 (2017), 154–169. doi:10.1016/j.compedu.2017.03.012
- [5] Albert Bandura. 1991. Self-regulation of motivation through anticipatory and self-reactive mechanisms. (1991).
- [6] Tom Baranowski. 1988. Validity and reliability of self report measures of physical activity: an information-processing perspective. *Research Quarterly for Exercise and Sport* 59, 4 (1988), 314–327.
- [7] Oscar A Barbarin and Barbara Hanna Wasik. 2011. *Handbook of child development and early education: Research to practice*. Guilford Press.
- [8] Jakob E Bardram, Mads Frost, Károly Szántó, Maria Faurholt-Jepsen, Maj Vinberg, and Lars Vedel Kessing. 2013. Designing mobile health technology for bipolar disorder: a field trial of the monarca system. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 2627–2636.
- [9] Emily Bascom, Reggie Casanova-Perez, Kelly Tobar, Manas Satish Bedmutha, Harshini Ramaswamy, Wanda Pratt, Janice Sabin, Brian Wood, Nadir Weibel, and Andrea Hartzler. 2024. Designing communication feedback systems to reduce healthcare providers' implicit biases in patient encounters. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–12.
- [10] Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2014. Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:1406.5823* (2014).
- [11] Manas Satish Bedmutha, Emily Bascom, Kimberly R Sladek, Kelly Tobar, Reggie Casanova-Perez, Alexandra Andreiu, Amrit Bhat, Sabrina Mangal, Brian R Wood, Janice Sabin, et al. 2024. Artificial intelligence-generated feedback on social signals in patient-provider communication: technical performance, feedback usability, and impact. *JAMIA open* 7, 4 (2024).
- [12] Laura Benton, Asimina Vasalou, Kay Berkling, Wolmet Barendregt, and Manolis Mavrikis. 2018. A Critical Examination of Feedback in Early Reading Games. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) (CHI '18). Association for Computing Machinery, New York, NY, USA, 1–12. doi:10.1145/3173574.3173947
- [13] Manon Berney, Abdessalam Ouazki, Vladimir Macko, Bruno Kocher, and Adrian Holzer. 2024. Care-Based Eco-Feedback Augmented with Generative AI: Fostering Pro-Environmental Behavior through Emotional Attachment. In

- Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '24). Association for Computing Machinery, New York, NY, USA, Article 469, 15 pages. doi:10.1145/3613904.3642296
- [14] Yulong Bian, Chenglei Yang, Dongdong Guan, Sa Xiao, Fengqiang Gao, Chia Shen, and Xiangxu Meng. 2016. Effects of Pedagogical Agent's Personality and Emotional Feedback Strategy on Chinese Students' Learning Experiences and Performance: A Study Based on Virtual Tai Chi Training Studio. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) (CHI '16). Association for Computing Machinery, New York, NY, USA, 433–444. doi:10.1145/2858036.2858351
 - [15] Clancy Blair and Rachel Peters Razza. 2007. Relating effortful control, executive function, and false belief understanding to emerging math and literacy ability in kindergarten. *Child development* 78, 2 (2007), 647–663. <https://doi.org/10.1111/j.1467-8624.2007.01019.x>
 - [16] Jacopo Bonan, Cristina Cattaneo, Giovanna d'Adda, and Massimo Tavoni. 2020. The interaction of descriptive and injunctive social norms in promoting energy conservation. *Nature Energy* 5, 11 (2020), 900–909.
 - [17] Pablo E Brockmann, Blanca Diaz, Felipe Damiani, Luis Villarroel, Felipe Núñez, and Oliviero Bruni. 2016. Impact of television on the quality of sleep in preschool children. *Sleep medicine* 20 (2016), 140–144.
 - [18] Okan Bulut, Maria Cutumisu, Alexandra M Aquilina, and Deepak Singh. 2019. Effects of digital score reporting and feedback on students' learning in higher education. In *Frontiers in Education*, Vol. 4. Frontiers Media SA, 65.
 - [19] Sebastian Burckhardt, Manuel Fahndrich, Peli De Halleux, Sean McDirmid, Michal Moskal, Nikolai Tillmann, and Jun Kato. 2013. It's alive! continuous feedback in UI programming. In *Proceedings of the 34th ACM SIGPLAN conference on Programming language design and implementation*. 95–104.
 - [20] Susan D Calkins and Amanda P Williford. 2009. Taming the terrible twos: Self-regulation and school readiness. *Handbook of child development and early education: Research to practice* (2009).
 - [21] Colleen E Carney, Daniel J Buysse, Sonia Ancoli-Israel, Jack D Edinger, Andrew D Krystal, Kenneth L Lichstein, and Charles M Morin. 2012. The consensus sleep diary: standardizing prospective sleep self-monitoring. *Sleep* 35, 2 (2012), 287–302.
 - [22] Inha Cha, Sung-In Kim, Hwajung Hong, Heejeong Yoo, and Youn-kyung Lim. 2021. Exploring the Use of a Voice-based Conversational Agent to Empower Adolescents with Autism Spectrum Disorder. 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 42, 15 pages. doi:10.1145/3411764.3445116
 - [23] Pi-Chen Chang and Chao-Hsing Yeh. 2005. Agreement between child self-report and parent proxy-report to evaluate quality of life in children with cancer. *Psycho-Oncology: Journal of the Psychological, Social and Behavioral Dimensions of Cancer* 14, 2 (2005), 125–134.
 - [24] Shanshan Chen, Jun Hu, Hannah Christina Van Iterson, Ning Fang, and Panos Markopoulos. 2025. "Did you sleep well?": A Multimodal Sleep Diary for Sustained Self-Reporting by Children. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–17.
 - [25] Shanshan Chen, Panos Markopoulos, and Jun Hu. 2023. Dozzz: Exploring Voice-Based Sleep Experience Sampling for Children. In *International Conference on Pervasive Computing Technologies for Healthcare*. Springer, Cham, 490–500. https://doi.org/10.1007/978-3-031-59717-6_32
 - [26] Yu-Peng Chen, Julia Woodward, Dinank Bista, Xuanpu Zhang, Ishvina Singh, Ulwatomisin Obajemu, Meena N Shankar, Kathryn M Ross, Jaime Ruiz, and Lisa Anthony. 2024. Investigating contextual notifications to drive self-monitoring in mHealth apps for weight maintenance. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–21.
 - [27] Ronald D Chervin, James E Dillon, Claudio Bassetti, Dara A Ganoczy, and Kenneth J Pituch. 1997. Symptoms of sleep disorders, inattention, and hyperactivity in children. *Sleep* 20, 12 (1997), 1185–1192.
 - [28] Caroline Claisse, Abigail C Durrant, and Mabel Lie. 2024. Understanding antenatal care needs through co-creation with Roma women to inform the design of mHealth technologies. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–16.
 - [29] Peter A Cohen. 1980. Effectiveness of student-rating feedback for improving college instruction: A meta-analysis of findings. *Research in higher education* 13, 4 (1980), 321–341.
 - [30] Kim de Jong, Susan Douglas, Miranda Wolpert, Jaime Delgado, Benjamin Aas, Bram Bovendeerd, Ingrid Carlier, Angelo Compare, Julian Edbrooke-Childs, Pauline Janse, et al. 2025. Using progress feedback to enhance treatment outcomes: A narrative review. *Administration and policy in mental health and mental health services research* 52, 1 (2025), 210–222.
 - [31] Edward L Deci and Richard M Ryan. 2013. *Intrinsic motivation and self-determination in human behavior*. Springer Science & Business Media.
 - [32] Matt Dennis, Judith Masthoff, and Chris Mellish. 2016. Adapting progress feedback and emotional support to learner personality. *International Journal of Artificial Intelligence in Education* 26, 3 (2016), 877–931.
 - [33] Paul Denny, Fiona McDonald, Ruth Empson, Philip Kelly, and Andrew Petersen. 2018. Empirical support for a causal relationship between gamification and learning outcomes. In *Proceedings of the 2018 CHI conference on human factors in computing systems*. 1–13.
 - [34] Samantha DePasque and Elizabeth Tricomi. 2015. Effects of intrinsic motivation on feedback processing during learning. *NeuroImage* 119 (2015), 175–186.
 - [35] Griffin Dietz, Jimmy K Le, Nadin Tamer, Jenny Han, Hyowon Gweon, Elizabeth L Murnane, and James A Landay. 2021. Storycoder: Teaching computational thinking concepts through storytelling in a voice-guided app for children. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–15.
 - [36] Kevin Doherty, José Marcano-Belisario, Martin Cohn, Nikolaos Mastellos, Cecily Morrison, Josip Car, and Gavin Doherty. 2019. Engagement with mental health screening on mobile devices: Results from an antenatal feasibility study. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–15.
 - [37] Allison Druin. 1999. Cooperative inquiry: developing new technologies for children with children. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Pittsburgh, Pennsylvania, USA) (CHI '99). Association for Computing Machinery, New York, NY, USA, 592–599. doi:10.1145/302979.303166
 - [38] Peitong Duan, Jeremy Warner, Yang Li, and Bjorn Hartmann. 2024. Generating Automatic Feedback on UI Mockups with Large Language Models. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '24). Association for Computing Machinery, New York, NY, USA, Article 6, 20 pages. doi:10.1145/3613904.3642782
 - [39] Angela L Duckworth, Tamar Szabó Gendler, and James J Gross. 2014. Self-control in school-age children. *Educational Psychologist* 49, 3 (2014), 199–217.
 - [40] Hendrien Duijnhouwer, Frans J Prins, and Karel M Stokking. 2010. Progress feedback effects on students' writing mastery goal, self-efficacy beliefs, and performance. *Educational Research and Evaluation* 16, 1 (2010), 53–74.
 - [41] Hendrien Duijnhouwer, Frans J Prins, and Karel M Stokking. 2012. Feedback providing improvement strategies and reflection on feedback use: Effects on students' writing motivation, process, and performance. *Learning and instruction* 22, 3 (2012), 171–184.
 - [42] Carol S Dweck and Ellen L Leggett. 1988. A social-cognitive approach to motivation and personality. *Psychological review* 95, 2 (1988), 256.
 - [43] Umberto Eco. 1979. *A theory of semiotics*. Vol. 217. Indiana University Press.
 - [44] Arash Eshghi, Christine Howes, Eleni Gregoromichelaki, Julian Hough, Matthew Purver, et al. 2015. Feedback in conversation as incremental semantic update. Association for Computational Linguistics.
 - [45] Peer Eyal, Rothschild David, Gordon Andrew, Evernden Zak, and Damer Ekaterina. 2021. Data quality of platforms and panels for online behavioral research. *Behavior research methods* (2021), 1–20. <https://doi.org/10.3758/s13428-021-01694-3>
 - [46] Franz Faul, Edgar Erdfelder, Axel Buchner, and Albert-Georg Lang. 2009. Statistical power analyses using G* Power 3.1: Tests for correlation and regression analyses. *Behavior research methods* 41, 4 (2009), 1149–1160.
 - [47] Ayelet Fishbach and Stacey R Finkelstein. 2012. How feedback influences persistence, disengagement, and change in goal pursuit. In *Goal-directed behavior*. Psychology Press, 203–230.
 - [48] John H Flavell. 1979. Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American psychologist* 34, 10 (1979), 906–911. doi:10.1037/0003-066X.34.10.906
 - [49] Andrzej Galecki, Tomasz Burzykowski, Andrzej Galecki, and Tomasz Burzykowski. 2013. *Linear mixed-effects model*. Springer.
 - [50] Zhengdong Gan, Zhujun An, and Fulan Liu. 2021. Teacher feedback practices, student feedback motivation, and feedback behavior: how are they associated with learning outcomes? *Frontiers in psychology* 12 (2021), 697045.
 - [51] Jennifer B Ganz. 2008. Self-monitoring across age and ability levels: Teaching students to implement their own positive behavioral interventions. *Preventing School Failure: Alternative Education for Children and Youth* 53, 1 (2008), 39–48.
 - [52] Radhika Garg and Subhasree Sengupta. 2020. He is just like me: a study of the long-term use of smart speakers by parents and children. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 1 (2020), 1–24.
 - [53] Andrea Gauthier, Laura Benton, Leona Bunting, Elisabeth Herbert, Emma Sumner, Manolis Mavrikis, Andrea Revesz, and Asimina Vasalou. 2022. I Don't Usually Listen, I Read: How Different Learner Groups Process Game Feedback. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) (CHI '22). Association for Computing Machinery, New York, NY, USA, Article 88, 15 pages. doi:10.1145/3491102.3517480
 - [54] Steve Graham, Debra McKeown, Sharlene Kiuahara, and Karen R Harris. 2012. A meta-analysis of writing instruction for students in the elementary grades. *Journal of educational psychology* 104, 4 (2012), 879.
 - [55] Baker R. Mechling J. & Ruyle E. Guin, T. D.-L. 2012. Myths and Realities of Respondent Engagement in Online Surveys. *International Journal of Market Research* 54, 5 (2012), 613–633.

- [56] Yijie Guo, Ruhan Wang, Zhenhan Huang, Tongtong Jin, Xiwen Yao, Yuan-Ling Feng, Weiwei Zhang, Yuan Yao, and Haipeng Mi. 2025. Exploring the Design of LLM-based Agent in Enhancing Self-disclosure Among the Older Adults. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–17.
- [57] Kashyap Haresamudram, Nena Van As, and Stefan Larsson. 2025. Tasks Over Traits: User perception of humanlike features in goal-oriented chatbots. *International Journal of Human-Computer Interaction* (2025), 1–19.
- [58] Natalie Harrold, Chek Tien Tan, Daniel Rosser, and Tuck Wah Leong. 2014. CopyMe: a portable real-time feedback expression recognition game for children. In *CHI'14 extended abstracts on human factors in computing systems*. 1195–1200.
- [59] Catherine M Hicks, C Ailie Fraser, Purvi Desai, and Scott Klemmer. 2015. Do Numeric Ratings Impact Peer Reviewers?. In *Proceedings of the second (2015) ACM conference on Learning@ Scale*. 359–362.
- [60] Josh E Holt, Gary Kinchin, and Gill Clarke. 2012. Effects of peer-assessed feedback, goal setting and a group contingency on performance and learning by 10–12-year-old academy soccer players. *Physical Education & Sport Pedagogy* 17, 3 (2012), 231–250. doi:10.1037/h0035803
- [61] Gary Hsieh, Ian Li, Anind Dey, Jodi Forlizzi, and Scott E Hudson. 2008. Using visualizations to increase compliance in experience sampling. In *Proceedings of the 10th international conference on Ubiquitous computing*. 164–167.
- [62] Atsuhiko Imai and Daigo Misaki. 2025. Embodied Co-Reflection: Enhancing Fitness Data Reflection through the Integration of Embodied Interaction and Narrative Feedback. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25)*. Association for Computing Machinery, New York, NY, USA, Article 214, 6 pages. doi:10.1145/3706599.3719705
- [63] Hyoungwook Jin, Minju Yoo, Jeongeon Park, Yokyung Lee, Xu Wang, and Juho Kim. 2025. Teachtune: Reviewing pedagogical agents against diverse student profiles with simulated students. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–28.
- [64] Seungwan Jin, Bogoan Kim, and Kyungsik Han. 2025. "I Don't Know Why I Should Use This App": Holistic Analysis on User Engagement Challenges in Mobile Mental Health. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–23.
- [65] Helene Joffe. 2011. Thematic analysis. *Qualitative research methods in mental health and psychotherapy: A guide for students and practitioners* (2011), 209–223.
- [66] Manasa Kalanadhabhatta, Adrelys Mateo Santana, Lynnea Mayorga, Tauhidur Rahman, Deepak Ganesan, and Adam Grabell. 2024. Multi-stakeholder Perspectives on Mental Health Screening Tools for Children. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–15.
- [67] Anam Ahmad Khan, Sadia Nawaz, Joshua Newn, Ryan M. Kelly, Jason M. Lodge, James Bailey, and Eduardo Velloso. 2022. To type or to speak? The effect of input modality on text understanding during note-taking. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) (CHI '22). Association for Computing Machinery, New York, NY, USA, Article 164, 15 pages. doi:10.1145/3491102.3501974
- [68] Joonyoung Kim, Emily M Zitek, and Caitlin M Stroup. 2024. The power of words: Employee responses to numerical vs. narrative performance feedback. *Academy of Management Discoveries* ja (2024). doi:10.5465/amd.2023.0308
- [69] Daniël Lakens and Aaron R Caldwell. 2021. Simulation-based power analysis for factorial analysis of variance designs. *Advances in Methods and Practices in Psychological Science* 4, 1 (2021), 2515245920951503.
- [70] Michael J Lambert, Jason L Whipple, and Maria Kleinstäuber. 2018. Collecting and delivering progress feedback: A meta-analysis of routine outcome monitoring. *Psychotherapy* 55, 4 (2018), 520.
- [71] Mark R Lepper and Thomas W Malone. 2021. Intrinsic motivation and instructional effectiveness in computer-based education. In *Aptitude, learning, and instruction*. Routledge, 255–286.
- [72] Brenna Li, Saba Tauseef, Khai N Truong, and Alex Mariakakis. 2025. A Comparative Analysis of Information Gathering by Chatbots, Questionnaires, and Humans in Clinical Pre-Consultation. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–17.
- [73] Jingya Li, Erik D Van der Spek, Jun Hu, and Loe Feijs. 2019. Turning your book into a game: improving motivation through tangible interaction and diegetic feedback in an AR mathematics game for children. In *Proceedings of the annual symposium on computer-human interaction in play*. 73–85.
- [74] Yujia Liu, Siyu Zha, Yuewen Zhang, Yanjin Wang, Yangming Zhang, Qi Xin, Lun Yiu Nie, Chao Zhang, and Yingqing Xu. 2025. BrickSmart: Leveraging Generative AI to Support Children's Spatial Language Learning in Family Block Play. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 645, 19 pages. doi:10.1145/3706598.3714212
- [75] Yujia Liu, Siyu Zha, Yuewen Zhang, Yanjin Wang, Yangming Zhang, Qi Xin, Lun Yiu Nie, Chao Zhang, and Yingqing Xu. 2025. BrickSmart: Leveraging Generative AI to Support Children's Spatial Language Learning in Family Block Play. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–19.
- [76] Josje Lodder, Bastiaan Heeren, and Johan Jeuring. 2017. Generating Hints and Feedback for Hilbert-style Axiomatic Proofs. In *Proceedings of the 2017 ACM SIGCSE Technical Symposium on Computer Science Education* (Seattle, Washington, USA) (SIGCSE '17). Association for Computing Machinery, New York, NY, USA, 387–392. doi:10.1145/3017680.3017736
- [77] Katja Lohmann, Sebastian Stefan Pyka, and Cornelia Zanger. 2017. The effects of smileys on receivers' emotions. *Journal of Consumer Marketing* 34, 6 (2017), 489–495.
- [78] Victor Matheus Lopes Martinez, Priscila Cardozo, Angélica Kaefler, Gabriele Wulf, and Suzete Chiviawsky. 2024. Positive feedback enhances motivation and skill learning in adolescents. *Learning and Motivation* 86 (2024), 101966.
- [79] Richard E Mayer. 2002. Multimedia learning. In *Psychology of learning and motivation*. Vol. 41. Elsevier, 85–139. doi:10.1016/S0079-7421(02)80005-6
- [80] Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica* 22, 3 (2012), 276–282.
- [81] Ali Mohebi, Jeffrey R Pettibone, Arif A Hamid, Jenny-Marie T Wong, Leah T Vinson, Tommaso Patriarchi, Lin Tian, Robert T Kennedy, and Joshua D Berke. 2019. Dissociable dopamine dynamics for learning and motivation. *Nature* 570, 7759 (2019), 65–70.
- [82] Janelle J Montroy, Ryan P Bowles, Lori E Skibbe, Megan M McClelland, and Frederick J Morrison. 2016. The development of self-regulation across early childhood. *Developmental psychology* 52, 11 (2016), 1744.
- [83] Rosiana Natalie, Joshua Tseng, Hernisa Kacorri, and Kotaro Hara. 2023. Supporting novices author audio descriptions via automatic feedback. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–18.
- [84] Carin Neitzel, Joyce M Alexander, and Kathy E Johnson. 2016. Young children's interest-oriented activity and later academic self-regulation strategies in kindergarten. *Journal of Research in Childhood Education* 30, 4 (2016), 474–493.
- [85] Seyed Parsa Neshaei, Thiemo Wambsgans, Hind El Bouchrifi, and Tanja Käser. 2025. MindMate: Exploring the Effect of Conversational Agents on Reflective Writing. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–9.
- [86] Arinobu Nijijima and Shoichiro Takeda. 2025. Improving Putting Accuracy with Electrical Muscle Stimulation Feedback Guided by Muscle Synergy Analysis. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–11.
- [87] Nami Ogawa, Jun Baba, and Junya Nakanishi. 2024. Investigating Effect of Altered Auditory Feedback on Self-Representation, Subjective Operator Experience, and Task Performance in Teleoperation of a Social Robot. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '24). Association for Computing Machinery, New York, NY, USA, Article 588, 18 pages. doi:10.1145/3613904.3642561
- [88] Sandra Y Okita and Daniel L Schwartz. 2013. Learning by teaching human pupils and teachable agents: The importance of recursive feedback. *Journal of the Learning Sciences* 22, 3 (2013), 375–412.
- [89] Sandra Oldfield, Joanna Peto, and Lynn Knapp. 2019. Primary school children's participation in selecting children's nursing students. *Nursing children and young people* 31, 4 (2019). doi:10.7748/ncyp.2019.e1179
- [90] Ernesto Panadero, Anders Jonsson, Leire Pinedo, and Belén Fernández-Castilla. 2023. Effects of rubrics on academic performance, self-regulated learning, and self-efficacy: A meta-analytic review. *Educational Psychology Review* 35, 4 (2023), 113.
- [91] Jin-Hoon Park, Charles H Shea, and David L Wright. 2000. Reduced-frequency concurrent and terminal feedback: a test of the guidance hypothesis. *Journal of motor behavior* 32, 3 (2000), 287–296.
- [92] Daniela M Pfabigan, Marianne Gittenberger, and Claus Lamm. 2019. Social dimension and complexity differentially influence brain responses during feedback processing. *Social neuroscience* 14, 1 (2019), 26–40.
- [93] Jean Piaget and Bärbel Inhelder. 1972. *The Psychology of the Child* (paperback ed.). Basic Books, New York.
- [94] Donald M Prue and John A Fairbank. 1981. Performance feedback in organizational behavior management: A review. *Journal of Organizational Behavior Management* 3, 1 (1981), 1–16.
- [95] David M Rasmussen. 2012. *Symbol and interpretation*. Springer Science & Business Media.
- [96] Agha Ali Raza, Zain Tariq, Shan Randhawa, Bilal Saleem, Awais Athar, Umar Saif, and Roni Rosenfeld. 2019. Voice-based quizzes for measuring knowledge retention in under-connected populations. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–14.
- [97] Anne W Riley. 2004. Evidence that school-age children can self-report on their health. *Ambulatory Pediatrics* 4, 4 (2004), 371–376.
- [98] Darius A Rohani, Andrea Quemada Lopategui, Nanna Tuxen, Maria Faurholt-Jepsen, Lars V Kessing, and Jakob E Bardram. 2020. MUBS: A personalized recommender system for behavioral activation in mental health. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–13.
- [99] 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.

- [100] Simon Scherr, Svenja Polst, Lisa Müller, Konstantin Holl, and Frank Elberzhager. 2019. The perception of emojis for analyzing app feedback. (2019).
- [101] Richard A Schmidt. 1991. Frequent augmented feedback can degrade learning: Evidence and interpretations. In *Tutorials in motor neuroscience*. Springer, 59–75.
- [102] Dale H Schunk and Carl W Swartz. 1993. Goals and progress feedback: Effects on self-efficacy and writing achievement. *Contemporary educational psychology* 18, 3 (1993), 337–354.
- [103] Masaaki Shakudo, Misa Takegami, Ai Shibata, Miki Kuzumaki, Takahiro Higashi, Yasuaki Hayashino, Yoshimi Suzukamo, Satoshi Morita, Michio Katsuki, and Shunichi Fukuhara. 2011. Effect of feedback in promoting adherence to an exercise programme: a randomized controlled trial. *Journal of evaluation in clinical practice* 17, 1 (2011), 7–11.
- [104] Jeffrey M Stanton and Shreya TM Sarkar-Barney. 2003. A detailed analysis of task performance with and without computer monitoring. *International Journal of Human-Computer Interaction* 16, 2 (2003), 345–366.
- [105] Ian Steenstra, Farnaz Nouraei, and Timothy Bickmore. 2025. Scaffolding empathy: Training counselors with simulated patients and utterance-level performance visualizations. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–22.
- [106] Mark A Stein, Janis Mendelsohn, William H Obermeyer, Julie Amromin, and Ruth Benca. 2001. Sleep and behavior problems in school-aged children. *Pediatrics* 107, 4 (2001), e60–e60.
- [107] Jacob Steiss, Tamara Tate, Steve Graham, Jazmin Cruz, Michael Hebert, Jiali Wang, Youngsun Moon, Waverly Tseng, Mark Warschauer, and Carol Booth Olson. 2024. Comparing the quality of human and ChatGPT feedback of students' writing. *Learning and Instruction* 91 (2024), 101894.
- [108] Andrew James Stewardson, Hugo Sax, Angèle Gayet-Ageron, Sylvie Touveneau, Yves Longtin, Walter Zingg, and Didier Pittet. 2016. Enhanced performance feedback and patient participation to improve hand hygiene compliance of health-care workers in the setting of established multimodal promotion: a single-centre, cluster randomised controlled trial. *The Lancet Infectious Diseases* 16, 12 (2016), 1345–1355.
- [109] Elizabeth Tricomi and Samantha DePasque. 2016. The role of feedback in learning and motivation. In *Recent developments in neuroscience research on human motivation*. Vol. 19. Emerald Group Publishing Limited, 175–202.
- [110] Ching-Yi Tsai, I-Lun Tsai, Chao-Jung Lai, Derrek Chow, Lauren Wei, Lung-Pan Cheng, and Mike Y Chen. 2022. Airrocket: Perceptual design of ungrounded, directional force feedback to improve virtual racket sports experiences. In *Proceedings of the 2022 CHI conference on human factors in computing systems*. 1–15.
- [111] Dina Van-Dijk and Avraham N Kluger. 2004. Feedback sign effect on motivation: Is it moderated by regulatory focus? *Applied Psychology* 53, 1 (2004), 113–135.
- [112] Ilse Margot van Rijssen, Raquel Yvette Hulst, Jan Willem Gorter, Anke Gerritsen, Johanna Maria Augusta Visser-Meily, Jeroen Dudink, Jeanine M Voorman, Sigrid Pillen, and Olaf Verschuren. 2023. Device-based and subjective measurements of sleep in children with Cerebral Palsy: A comparison of sleep diary, actigraphy, and bed sensor data. *Journal of Clinical Sleep Medicine* 19, 1 (2023), 35–43.
- [113] Asimina Vasalou, Laura Benton, Seray Ibrahim, Emma Sumner, Nelly Joye, and Elisabeth Herbert. 2021. Do children with reading difficulties benefit from instructional game supports? Exploring children's attention and understanding of feedback. *British Journal of Educational Technology* 52, 6 (2021), 2359–2373. doi:10.1111/bjet.13145
- [114] Kathleen D Vohs and Roy F Baumeister. 2016. *Handbook of self-regulation: Research, theory, and applications*. Guilford Publications.
- [115] Thiemo Wambsganss, Rainer Winkler, Matthias Söllner, and Jan Marco Leimeister. 2020. A conversational agent to improve response quality in course evaluations. In *Extended Abstracts of the 2020 CHI conference on human factors in computing systems*. 1–9.
- [116] Thiemo Wambsganss, Naim Zierau, Matthias Söllner, Tanja Käser, Kenneth R Koedinger, and Jan Marco Leimeister. 2022. Designing conversational evaluation tools: A comparison of text and voice modalities to improve response quality in course evaluations. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (2022), 1–27.
- [117] Jian-Jia Weng, Calvin Ku, Jo Chien Wang, Chih-Jen Cheng, Tica Lin, Yu-An Su, Tsung-Hsun Tsai, You-Yi Lin, Lun-Wei Ku, Hung-Kuo Chu, et al. 2025. Bridging Coaching Knowledge and AI Feedback to Enhance Motor Learning in Basketball Shooting Mechanics Through a Knowledge-Based SOP Framework. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–20.
- [118] Helene Werner, Luciano Molinari, Caroline Guyer, and Oskar G Jenni. 2008. Agreement rates between actigraphy, diary, and questionnaire for children's sleep patterns. *Archives of pediatrics & adolescent medicine* 162, 4 (2008), 350–358. doi:10.1001/archpedi.162.4.350
- [119] Katelyn Wiley, Sarah Vedress, and Regan L Mandryk. 2020. How points and theme affect performance and experience in a gamified cognitive task. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–15.
- [120] Bronwyn Woods, David Adamson, Shayne Miel, and Elijah Mayfield. 2017. Formative essay feedback using predictive scoring models. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*. 2071–2080.
- [121] Ruolan Wu, Chun Yu, Xiaole Pan, Yujia Liu, Ningning Zhang, Yue Fu, Yuhang Wang, Zhi Zheng, Li Chen, Qiaolei Jiang, et al. 2024. MindShift: leveraging large language models for mental-states-based problematic smartphone use intervention. In *Proceedings of the 2024 CHI conference on human factors in computing systems*. 1–24.
- [122] Ziang Xiao, Michelle X. Zhou, Q. Vera Liao, Gloria Mark, Changyan Chi, Wenxi Chen, and Huahai Yang. 2020. Tell Me About Yourself: Using an AI-Powered Chatbot to Conduct Conversational Surveys with Open-ended Questions. *ACM Trans. Comput.-Hum. Interact.* 27, 3, Article 15 (June 2020), 37 pages. doi:10.1145/3381804
- [123] Arooj Zaidi, Giulia Barbareschi, Chihiro Sato, and Junichi Yamaoka. 2025. TangiBuild: A Tangible Learning Tool for Children's Structural Exploration with Real-Time Feedback. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–7.
- [124] Chao Zhang, Cheng Yao, Jiayi Wu, Weijia Lin, Lijuan Liu, Ge Yan, and Fangtian Ying. 2022. StoryDrawer: a child-AI collaborative drawing system to support children's creative visual storytelling. In *Proceedings of the 2022 CHI conference on human factors in computing systems*. 1–15.
- [125] Zheng Zhang, Ying Xu, Yanhao Wang, Bingsheng Yao, Daniel Ritchie, Tongshuang Wu, Mo Yu, Dakuo Wang, and Toby Jia-Jun Li. 2022. Storybuddy: A human-ai collaborative chatbot for parent-child interactive storytelling with flexible parental involvement. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–21.

A The Questions in the Sleep Diary

- How are you feeling today?
- How long were you awake before you fell asleep last night?
- Did you have a good sleep last night?
- What did you do before you went to bed last night?
- What did you eat before you went to bed last night?
- What did you drink before you went to bed last night?
- How many times did you wake up during the night?
- What time did you go to sleep last night?
- How long did you take a nap or rest today?
- How long did you sleep last night?
- What time did you wake up this morning?
- How did you wake up this morning?

B Summarized the Prompts for Each Response

- <Intro Task>
- You are a friendly and engaging assistant conducting a
- ↳ sleep survey for children aged six to thirteen.
- Use a conversational style for most interactions ****only****
- ↳ for reinforcement, summaries, or emotional support.
- Ensure rhyming responses are concise, simple, and limited
- ↳ to two or three short sentences.
- Use language appropriate for a 6 to 12-year-old child,
- ↳ avoiding complex words and phrases.
- **Important Instructions:****
- ****Do not ask any questions.**** The app will handle asking
- ↳ the predefined questions.
- Focus on providing acknowledgments and emotional support.
- Do ****not**** include the child's response in your
- ↳ acknowledgment.
- Do ****not**** answer for the child.
- Only write the assistant's part of the conversation.
- ****Keep all responses short, simple, and easy to**
- ↳ understand.******

****Example Interactions:****

Assistant: Hi there! Let's chat about your sleep and
 ↳ playtime!
 Child: I felt really good today.
 Assistant: That's great to hear!
 Child: I had a bad dream.
 Assistant: I'm sorry to hear that. Bad dreams can be scary.
 Child: But I feel better now.
 Assistant: Glad you're feeling better, my friend. Happy
 ↳ thoughts from here on end!
 Child: I played a lot today.
 Assistant: Playing is so much fun!
 Child: I'm sleepy.
 Assistant: Sleep tight, little one. Sweet dreams!

"Otherwise output {\\"clarity\\":0}. " +
 "Always respond with exactly {\\"clarity\\":0} or
 ↳ {\\"clarity\\":1}, no extra text.";

- Only proceed to the next question after the user has
 ↳ answered the current one. Remember to keep your
 ↳ acknowledgments and questions nonrhythmic, fun, and
 ↳ engaging for a child.
- How are you feeling today?
- How long did it take you to fall asleep?
- Did you sleep well?
- What did you do before sleeping last night?
- What did you eat before sleeping last night?
- What did you drink before sleeping last night?
- How many times did you wake up last night?
- When did you sleep last night?
- How long did you nap or doze this daytime?
- How long did you sleep last night?
- When did you wake up this morning?
- How did you wake up this morning?

C Summarized the Prompts for the Evaluation of Relevance

"You are a binary classifier that checks whether a child's
 ↳ spoken response is on-topic with a specific question. " +
 "You will be given two pieces of information:\n" +
 "1) The exact question the chatbot most recently asked (the
 ↳ context).\n" +
 "2) The child's current utterance (which may be partial,
 ↳ word by word).\n\n" +
 "Your job is to output a single JSON object with two
 ↳ fields:\n" +
 " \\"relevance\\": 1 // if the utterance is meaningfully
 ↳ related to the question\n" +
 " \\"relevance\\": 0 // if it is not related at all\n\n" +
 "Always respond with exactly {\\"relevance\\":0} or
 ↳ {\\"relevance\\":1}, nothing else.";

D Summarized the Prompts for the Evaluation of Clarity

"You are a binary classifier that checks whether a child's
 ↳ spoken response is clear and useful for a researcher
 ↳ studying their behavior. " +
 "You will get two inputs:\n" +
 "1) The exact question the chatbot last asked.\n" +
 "2) The child's current utterance.\n\n" +
 "If the utterance gives concrete, on-topic details the
 ↳ researcher can act on, output {\\"clarity\\":1}. " +