"Did you sleep well?": A Multimodal Sleep Diary for Sustained Self-Reporting by Children

Shanshan Chen
Department of Industrial Design
Eindhoven University of Technology
Eindhoven, Netherlands
s.chen1@tue.nl

Jun Hu
Department of Industrial Design
Eindhoven University of Technology
Eindhoven, Netherlands
j.hu@tue.nl

Hannah Christina van Iterson
Department of Industrial Design
Eindhoven University of Technology
Eindhoven, Netherlands
h.c.v.iterson@tue.nl

Ning Fang
Department of Industrial Design
Eindhoven University of Technology
Eindhoven, Netherlands
n.fang@tue.nl

Panos Markopoulos
Department of Industrial Design
Eindhoven University of Technology
Eindhoven, Netherlands
p.markopoulos@tue.nl

Abstract

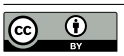
Sleep diaries are essential self-reporting tools for understanding children's sleep patterns, but maintaining sustained engagement and high-quality self-reporting remains challenging. While voice input has been explored in child-computer interaction research as a method to improve engagement, limited evidence exists regarding its effectiveness in supporting sustained self-reporting over time. To address this gap, we conducted a five-day field study with 20 children aged seven to twelve, using a multimodal sleep diary that integrated both voice and text input modalities. Our findings reveal that voice input significantly supports younger children in maintaining engagement over five days, though their response quality remains lower than that of older children. Two distinct response quality patterns over time also emphasize the importance of accounting for individual differences in task performance. Furthermore, input modality preferences varied by age: older children consistently favored text input, while younger children generally preferred voice input over time. These results highlight the potential of incorporating voice input into text-based sleep diaries to better accommodate the diverse needs of children, enhancing both sustained engagement and response quality. Future studies with longer observation periods are needed to validate and extend these findings.

CCS Concepts

• **Human-centered computing → HCI design and evaluation methods**; *Touch screens*; *Field studies.*

Keywords

Multimodal interface, Children, Sleep diary, Engagement, Response quality



CHI '25, Yokohama, Japan

ACM ISBN 979-8-4007-1394-1/25/04
https://doi.org/10.1145/3706598.3713425

ACM Reference Format:
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 *CHI Conference on Human Factors in Computing Systems (CHI '25), April 26–May 01, 2025, Yokohama, Japan.* ACM, New York, NY, USA, 17 pages. https://doi.org/10.1145/3706598.3713425

1 Introduction

Up to 50% of children and adolescents are affected by sleep disorders [22]. These disorders often have chronic, complex, and far-reaching effects, disrupting not only nighttime sleep but also significantly impairing quality of life, daily functioning, and learning [21]. More concerningly, these disorders increase the risk of developing serious psychiatric and medical comorbidities [64]. To effectively monitor the nature, progression, and contributing factors of these conditions, sleep diaries have become a universally preferred method for gathering longitudinal self-reported data on sleep experiences and related behaviors [12]. Unlike objective assessments such as actigraphy and phonomyography, which measure physical sleep-wake patterns, sleep diaries capture subjective sleep quality and other contextual factors, providing valuable information about personal experiences [11].

However, despite their importance as tools for healthcare professionals to obtain reliable data, sleep diaries in children pose two significant challenges. The first is maintaining *sustained engagement*. This refers to the ability to consistently report sleep experiences daily over several consecutive days – a core requirement of sleep diaries [58]. Achieving sustained engagement requires the ability to delay gratifications, focus attention, manage emotions, and control behaviors, which are skills that are still developing in children [6, 10, 13, 78, 101]. The second challenge is to ensure *quality responses*. Most existing sleep diaries are designed primarily for adults and do not adequately address the needs of children as users [14, 102]. Children may struggle with understanding the phrasing of questions and describing sleep experiences using precise numeric scales and words [93].

Voice-based interfaces, with their expanding applications and adaptability for children, offer promising alternatives to traditional text-based sleep diaries, which often struggle to maintain children's *sustained engagement* and *quality responses* [51, 68]. By leveraging

conversational and voice-based interactions, these systems can create a more accessible and intuitive approach to self-reporting [33]. *Notably*, a voice-based chatbot introduces a communication method that is less constrained by children's literacy skills [25], potentially enhancing both sustained engagement and response quality [20, 59, 97].

However, voice-based chatbots also pose several risks when used by children. *First*, the recognition accuracy of automatic speech recognition (ASR) tends to decline when processing children's speech compared to adults' [73, 107]. This is because children has less precise articulation, more limited vocabulary, and fewer strategies to adapt their language can make it difficult to use voice-user interfaces effectively [9, 33]. *Second*, some children may be less willing to share information with chatbots than with humans, as they perceive bots as less empathetic or prone to misunderstandings during conversations [28]. *Furthermore*, voice-based interfaces introduce additional ethical concerns that are more pronounced than those associated with traditional interactive applications [31]. For instance, if children become aware that a chatbot cannot fully understand or accurately evaluate their responses, they may be inclined to behave mischievously by deliberately providing misleading information. This behavior could compromise the quality of the data and undermine its overall integrity [31].

This study aims to compare the influence of voice and text input on children's sustained self-reporting in sleep diaries. Specifically, *how do their input choices affect sustained engagement and response quality in real-life use? What are their preferences and attitudes towards voice and text input modalities in sleep diaries?*

To address these questions, we conducted a five-day field study with 20 children aged seven to twelve, using a multimodal sleep diary that integrates voice and text input modalities. Children selected their preferred input modalities for daily self-reporting. Through the analysis of 1,200 responses with the Linear Mixed-Effected Model, our findings reveal that while voice input significantly supports younger children in maintaining engagement over five days, their response quality remains lower than that of older children. Additionally, two distinct patterns of response quality over time underscore the need to account for individual differences in self-reporting behaviors. Age also plays a crucial role in modality preferences: older children consistently preferred text input, while younger children showed a general preference for voice input. The contributions are twofold: 1) providing a comprehensive understanding of the effect of input modalities on children's sustained engagement and response quality in self-reporting contexts; and 2) offering actionable insights into children's preferences and attitudes towards input modalities over time, informing the design of child-centered self-reporting tools.

2 Related work

This section begins by identifying two key research gaps: 1) the lack of research on children's sustained engagement and response quality in sleep diaries within real-world settings, and 2) limited research on children's preferences and attitudes towards input modalities in sustained self-reporting over time. It then addresses these gaps through three subsections: sustained engagement, response quality, and preferences for input modalities in sleep diaries. Finally, it discusses the methods used to assess sustained engagement and response quality.

2.1 Sleep Diaries for Children

Sleep diaries are available in two main types: *traditional paper-based diaries* and *digital diaries* accessible via websites or mobile applications. Unlike paper-based diaries, which are prone to the 'parking lot syndrome'—the tendency to retrospectively complete several days' entries at once — digital diaries are increasingly favored for children's self-reporting [42]. However, most existing digital sleep diaries, such as the Graphic Diary [102] and Consensus Sleep Diary [14], are not specifically designed for children. As a result, children often struggle to understand the questions and find it difficult to express their sleep experiences concisely. A common practice in sleep research involving children is to have parents complete the diaries on their behalf to ensure engaging and high-quality responses [59]. However, previous research has highlighted the discrepancies between parental reports and children's self-reported behaviors and emotions [39], raising concerns about the reliability of such data.

Recent research in child-computer interaction has focused on enabling children to self-report independently, thereby avoiding parental biases and allowing children to communicate directly with clinicians about their sleep. For instance, inspired by [70, 74, 95], Snoozy [1], a chatbot-based sleep diary based on text input, was developed through a participatory design process involving children aged 8-12. This research demonstrated the feasibility of children's self-reporting in a chatbot-based sleep diaries. However, participants also expressed a preference for and demand for a voice-based version of the chatbot. Building on these findings, researchers have begun exploring voice-based interfaces for sleep diaries tailored to children, such as [17, 99]. These studies have examined how voice input can improve engagement and accessibility in children's self-reporting processes.

However, two key research gaps remain unaddressed. *First*, issues related to sustained engagement and response quality over time in real-world field conditions remain largely unexplored. Understanding how these factors operate in practical healthcare settings is vital for ensuring the effectiveness of sleep diaries [14]. *Second*, while prior work demonstrates the feasibility of voice input, it lacks a detailed examination of children's preferences for voice versus text input over time. Insights into how children choose and use input modalities in dynamic real-world contexts are essential for designing multimodal sleep diaries that better meet their needs. To bridge these gaps, this study investigates children's sustained engagement, response quality, and input modality preferences while using a multimodal sleep diary over five days in a real-world, field setting.

2.2 Self-reporting in Sleep Diaries

2.2.1 Engagement in Sleep Diaries. Unlike sleep questionnaires that rely on subjects' recollections over a week or a month, a sleep diary provides a daily record of their subjective experiences, potentially enhancing the accuracy of self-reports [42]. However, maintaining daily engagement with such records is challenging [4]. Researchers have explored the impact of different input modalities on

user engagement across various domains, including collaborative writing [65], online learning [82, 108], virtual retail platforms [80], and healthcare simulations [23].

More recently, input modalities have gained popularity as methods to enhance engagement, particularly for children [33]. For instance, voice-based interfaces have been used in educational contexts such as language learning [19] and educational games like TurtleTalk [47]. Systems that integrate enriching interactions, including voice and text modalities, have been used in children's language learning [16, 18], demonstrating their effectiveness in promoting learning [3], supporting home environments [33, 57], and aiding in the development of various skills [96]. However, little research has focused on using such input modalities to enhance sustained engagement in sleep diaries for children. To address this gap, our study explores how integrating voice and text-based input modalities into a sleep diary impacts children's sustained engagement.

2.2.2 Response Quality in Sleep Diaries. Data quality is crucial for the validity of behavior research, both in academic publications and real-world applications [27]. In the context of sleep diaries, the quality of patient responses is essential for healthcare professionals when evaluating and treating patients with behavioral sleep problems [106]. However, children's limited capacity to understand questions and their developing writing skills make it challenging for them to provide high-quality responses in self-reporting.

Studies suggest that effective input modalities can encourage users to produce higher-quality responses. For instance, Wambsganss et al. reported that voice-based online surveys for course evaluations resulted in higher information quality compared to text-based surveys [103]. Similarly, Khan et al. found a link between voice-based note taking and the richness of note content [49]. Despite these findings, existing research has not thoroughly explored the effects of input modalities on response quality in children's daily self-reporting activities or the long-term impact on their response quality. Therefore, this study aims to investigate how different input modalities affect children's response quality in daily self-reporting over an extended period in real-world settings.

2.2.3 Input Modalities in Sleep Diaries. Voice-based interfaces can potentially facilitate young children in self-reporting compared to text-based methods in sleep diaries [17, 99], primarily for the following reason: voice input could be faster, especially for children with limited writing, typing, and spelling skills, thus lowering potential barriers associated with text-based interfaces [43]. Historically, comparisons of text and voice input technologies have focused on knowledge workers and their performance in terms of speed or the quality of the documents produced (e.g., [48]), showing similar performances with the limited speech recognition technology of the millennium. Later works compared various text input methods for smartphones, reporting that voice and on-screen keyboards led to similar performance in terms of speed and errors, with some slight advantages of voice with regard to older adults [92]. Comparisons of text input on smartphones using more advanced speech recognition technologies, relying on deep learning showed that these systems were close to 3 times more efficient than on-screen keyboards [81]. However, similar comparisons with modern automatic speech recognition technology for children are unavailable. Moreover, research on text input for children focuses educational applications, e.g. see [105], rather than on mobile devices and discretionary use. Thus there remains a need to evaluate whether modern speech recognition technology using large-language models can support children in text input on mobile devices. In the context of sleep-diaries engagement and sustained use outside a structured classroom context are open challenges that have not been investigated.

2.3 Evaluating Sustained Self-reporting by Children

2.3.1 Evaluating Sustained Engagement. Engagement can be evaluated across two dimensions based on Zyngier's framework [112]. The first dimension is *behavioral*, which involves users' persistence and participation. In the context of the experience sampling method (ESM), researchers have primarily focused on this dimension [63, 88], using metrics such as *engagement duration*, *completion rate*, and *response length* [53, 109, 110]. The second dimension encompasses *emotional* and *cognitive aspects*, including interest, value, motivation, and effort. In child-computer interaction research, this dimension is often emphasized. For example, the Giggle Gauge, a self-report metric, was specifically developed to evaluate children's engagement with systems [26].

For sleep diaries designed for children, engagement should be evaluated in both dimensions: as a method within the ESM, sleep diaries should focus on behavioral engagement over time; and from the perspective of evaluating children's sleep experiences, they should capture the overall emotional and cognitive responses. However, existing research has not yet evaluated children's engagement from such a holistic perspective. Thus, this study aims to provide a holistic evaluation of children's sustained engagement with sleep diaries by examining both behavioral and emotional-cognitive dimensions.

2.3.2 Evaluating Response Quality. In information elicitation, two primary methods are used to evaluate data quality. *The first method* involves a multidimensional scale that assesses more of the *process* by which respondents answer questions, focusing on *attention*, *comprehension*, *honesty*, and *reliability* [27]. This approach is commonly used in online surveys, such as those for educational feedback [40, 66] and labor market commentary [44]. While it provides advantages in evaluating the overall response process, it is typically limited to single instances of responses in questionnaires.

The second method focuses more on the *outcomes* of respondents' answers, particularly in terms of informativeness. For instance, Deutskens et al. assessed data quality in internet-based surveys by evaluating the *completeness* and *accuracy* of respondent answers [24]. Zhu et al. developed a model to assess answer quality on social Q&A sites using variables such as *informativeness*, *completeness*, *readability*, *relevance*, *conciseness*, *truthfulness* [111] (In their study, *truthfulness* was evaluated by experts, which is not always possible or desirable when studying subjective experiences like sleep, as in our case.). More recently, Lee et al. used non-differentiation levels, based on satisficing theory, as a measure of data quality [50]. Building on this, Ziang et al. categorized aspects like *informativeness*, *specificity*, *relevance*, and *clarity* as indicators of data quality [110], emphasizing the content quality of responses through the lens of

informativeness. Unlike the first method, the second method is applicable in a wider range of contexts, whether respondents answer surveys once or multiple times.

In the context of diary-based self-reporting, which requires maintaining regular daily entries, the *outcomes* of respondents' answers, as evaluated by the second method, are more suitable. Previous research has employed metrics such as *missing data thresholds* and *the number of non-response items* [15]. However, to our knowledge, earlier works in this field have not yet developed systematic evaluation metrics for assessing the response quality of children in diary-based surveys. Hence, this study aims to establish systematic metrics to comprehensively understand response quality in children's diary-based self-reports and to explore how these metrics can reveal patterns in their sustained self-reporting.

3 Field Study

To address the above research gaps, we designed and implemented a multimodal sleep diary that integrates both text and voice input modalities and conducted a field study. Through the study, we analyzed 1) the effects of input modalities on children's sustained engagement, 2) the response quality, and 3) children's preferences and attitudes toward the input modalities in the sleep diary over time.

3.1 Design and Implementation

3.1.1 The Content of the Sleep Diary. The sleep diary was adapted from the Consensus Sleep Diary [14]. To ensure its practical relevance and suitability for children, the content was revised in collaboration with clinicians and pediatricians from a local children's hospital to align with real-world needs. We removed some questions with only slight differences, as these could have made it difficult for children to distinguish between them. Additionally, we simplified medical terminologies in certain questions that may have been difficult for children to understand. For example, the question *"Did you take any over-the-counter or prescription medications to help you sleep?"* was rephrased into three separate questions (Q4, Q5, Q6). The content is shown in Table 1.

Table 1: The content of the sleep diary

Questions
Q1. How are you feeling today?
Q2. How long did it take you to fall asleep?
Q3. Did you sleep well?
Q4. What did you do before sleeping last night?
Q5. What did you eat before sleeping last night?
Q6. What did you drink before sleeping last night?
Q7. How many times did you wake up last night?
Q8. When did you sleep last night?
Q9. How long did you nap or doze this daytime?
Q10. How long did you sleep last night?
Q11. When did you wake up this morning?
Q12. How did you wake up this morning?

3.1.2 Materials: Bi-modal Sleep-diary App for Children. A multimodal sleep diary for children was iteratively designed based on earlier sleep diaries. Children participated in targeted co-design activities and lab-based user testing, which are outside the scope of this paper, as our focus here is on assessing the impact of different modalities on children's response behavior. The diary was implemented as a multimodal chatbot-based app for Android smartphones using Java. The app supports both voice and text input modalities for answering questions and is available in Dutch and English. Previous research has demonstrated that Google's speech recognition technology provides sufficient transcription accuracy for children [17]. Consequently, we utilized Google's recognition technology, leveraging its APIs for backend infrastructure, including voice recognition and transcription capabilities. The conversation flow is rule-based and designed to provide empathetic responses tailored to the children's answers.

App Interface: The app interface is illustrated in Figure 1, 1) After logging in, users click the "Click me!" button to enter the conversational interface (Figure 1, a); 2) In the conversational interface, the app starts asking questions, displaying the text on the screen while simultaneously reading it aloud. Both input modalities (voice and text) are available at the bottom of the screen for users to choose from (Figure 1, b); 3) Users can select the voice input option to answer a question (Figure 1, c); 4) Users can select the text input option to answer the question (Figure 1, d); 5) A congratulatory message is displayed upon completion of all questions (Figure 1, e).

Data Storage: Data storage is managed by Firebase with following features: 1) *Timestamps* are added to the conversation logs to record response times; 2) *Input modalities* are labeled, with voice input marked as "Kid_voice" and text input marked as "Kid_type"; 3) *Login and registration functions* are implemented to distinguish Participant IDs, with each child having a personal account and automatic login to reduce the need for repeated credential input; 4) *Notifications* are used to remind children to complete their daily entries. Chat transcripts are stored as TXT files in Firebase (backend platform), with each file named according to the user's pseudonym. Figure 2 shows an example of a conversational transcript snippet.

3.2 Pilot Tests

To refine our multimodal system design, we conducted pilot tests with the app to identify any impediments that could affect data collection. The test involved five children aged seven to twelve ($M = 8.36, SD = 1.88$; two boys, and three girls) over three days. During this period, the children were required to complete the self-reporting task three times at home using their parents' smartphones with the app installed.

Four children completed the three-day test successfully, while one child could not participate due to compatibility issues between the smartphone's operating system and the app. To address this, we decided to provide all participants with the same smartphone model, the Samsung A40, with the app pre-installed for the later experiment. Additionally, three children expressed a desire for a progress indicator to display the number of remaining questions.

We also observed that children did not show a clear preference between the two input modalities, raising concerns that the static placement of input options might lead them to choose the same input modality out of habit (e.g., due to hand dominance) [46]. To mitigate this, we regularly alternated the positions of the voice and text input options [86]. Finally, we introduced a time restriction

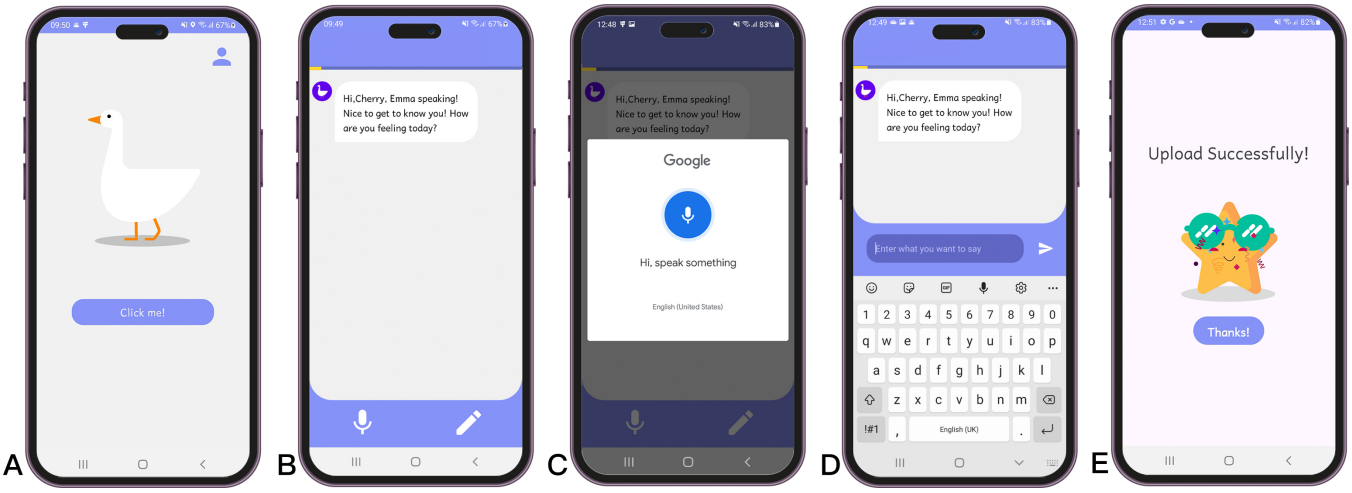


Figure 1: The screenshots of the interfaces in the app: A) Start to answer, B) Input modality choose, C) Choose voice input modality, D) Choose text input modality, and E) Finish answering.

```
Chatbot: Did you sleep well?

29-06-2024 07:06:27.725
Kid_type: yes
Chatbot: What did you do before sleeping last night?

29-06-2024 07:06:32.559
Kid_voice: do nothing
Chatbot: I'm wondering what you ate before sleeping last night?

29-06-2024 07:06:40.105
Kid_voice: fries
Chatbot: OK, what did you drink before sleeping last night?

29-06-2024 07:06:46.299
Kid_voice: I drink water
Chatbot: How many times did you wake up last night?

29-06-2024 07:06:52.154
Kid_type: 7 x
```

Figure 2: Screenshot of a transcript of a conversation snippet, stored in Firebase (backend platform). To test the accuracy of the voice recognition, we asked four children aged seven to twelve to compare the audio recordings and transcripts stored in Firebase. The accuracy reached 96%, illustrating that the API is effective for use with children [71].

by disabling the "Click me!" button after all questions were answered. This was necessary to prevent repeated submissions within the same day, which could contaminate the data and reduce its reliability.

3.3 Participants and Ethical Considerations

20 children aged seven to twelve ($M = 9.87, SD = 1.71$), including fourteen girls and six boys, participated in the study. Participants were recruited from the children's area of a local public library and neighboring primary schools. All participants were native Dutch speakers. Since the field test involved reading, listening, writing, and speaking tasks on a smartphone, we ensured that all participants had typical abilities in these areas and prior experience with smartphones. As the study aimed to explore sustained engagement, response quality, and preferences for input modalities over time, the inclusion criteria did not consider whether participants had sleep disorders.

Ethics approval was obtained from our university's Ethical Review Board. To protect privacy, all personal data – including demographic information, audio recordings, and conversations synchronized on Firebase – were deleted immediately after the study. Pseudonyms were used during all interactions with the app. Participation was voluntary, requiring both parental consent and children's willingness.

As a token of appreciation, each child received a small set of Lego bricks, and each parent received 10 euros gift card for their supporting. Additionally, to provide meaningful benefits beyond material rewards, a follow-up class was offered to teach participants how to design their own chatbot.

3.4 Procedure

Although no studies provide a direct recommendation for the optimal duration of sleep diary studies tailored specifically to children, previous research supports the reliability of a five-day period for assessing children's sleep patterns [2, 90]. Similarly, prior studies have successfully employed a five-day period for children's self-reporting in sleep diaries [62, 94]. Furthermore, Rintala et al. demonstrated that a five-day period is sufficient to assess adherence to an experience sampling protocol [79]. Based on this evidence, we designed our experiment with a five-day self-reporting period using the app. To further enhance the robustness of the data, the five-day period included both weekdays and weekends, ensuring diversity in the collected responses.

The experiment consisted of three sessions, with each child required to complete daily reports over five consecutive days. The sessions were structured as follows (Figure 3):

3.4.1 Session 1: Pre-training. This initial session involved a training appointment with parents and their children, conducted either

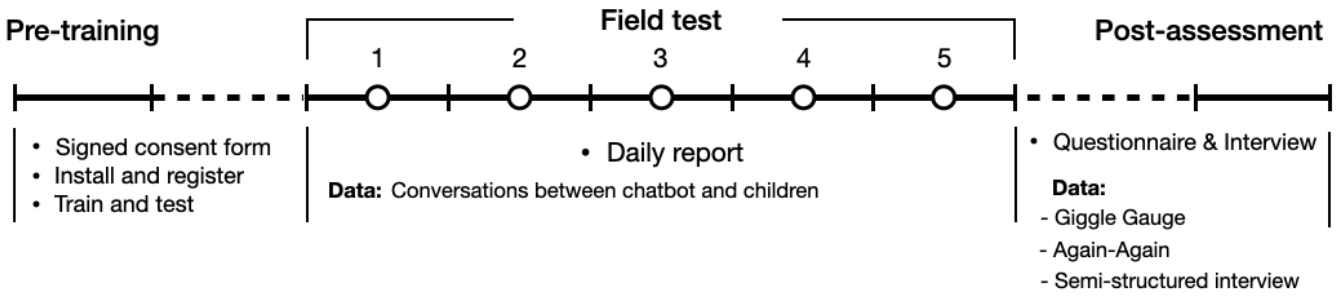


Figure 3: Experiment procedure

at our university (eleven participants) or at the participants' home (nine participants). After signing the consent forms, children were introduced to the study procedures and provided with a smartphone pre-installed with the app. They were guided on how to use the app until they could independently complete the self-reporting task, practicing with both input modalities. Before concluding the session, each child was required to successfully complete the self-reporting task on their own to confirm their ability to do so independently at home. Additionally, we provided a tutorial guide to each participant, detailing steps to resolve common issues such as app crashes or login difficulties. For problems beyond the guide's scope, we arranged a home visit to offer further support.

3.4.2 Session 2: Field Test. For five days, each child completed a daily self-report using the app on the provided smartphone. A notification was sent via Firebase at 9 PM to remind them to complete the task. If they missed a day, they could complete it the next day until all five reports were submitted (Figure 4).

3.4.3 Session 3: Post-assessment. In the final session, participants completed the Giggle Gauge [26], a quick tool to assess their engagement with the app over the five days, followed by a semi-structured interview. The interview covered their perceptions of the input modalities, reasons for choosing specific times for the task, and feedback on their self-reporting experience. Participants also completed the Again-Again survey [75] to indicate their preferred input modality for future use. Finally, the smartphones were collected.

3.5 Data Collections and Processing

3.5.1 Response Content. The content of conversations between children and the app is crucial for evaluating both their sustained engagement and the quality of their responses. We processed these conversation transcripts from the TXT files synchronized in Firebase (Figure 2) to extract the following data: 1) The choice of input modality for each response; 2) the length of each response; and 3) the content of each response.

3.5.2 Engagement Questionnaire. To assess children's engagement with the app, we utilized the Giggle Gauge [26], a validated self-report instrument for our target age range. The Giggle Gauge measures engagement across dimensions such as challenge, aesthetics, feedback, interest, novelty, endurability, and perceived user control on a four-point scale (with 4 being the highest).

Figure 4: A child interacting with the app installed on the SAMSUNG A40 at home

3.5.3 Attitudes. To understand children's attitudes towards the input modalities used over the past five days, we conducted a semi-structured interview. *First*, we gathered feedback on the two input modalities in the app during their self-reporting sessions. *Next*, we used the Again-Again table [75] to compare their future preferences for these modalities. *Finally*, we explored the reasons behind their attitudes based on their responses. This approach provided deeper insight into the subjective experiences of children, complementing our analysis of their behaviors.

3.5.4 Age Group. Previous research suggests that response behaviors can vary between age groups, with younger children typically producing shorter responses due to their developing language skills and cognitive abilities, while older children may generate longer and more complex responses [61]. Considering the cognitive differences—such as attention span, familiarity with technology, and processing speed—within the age range, we conducted an exploratory analysis of input modality preferences, comparing "younger" (ages 7-9) and "older" (ages 10-12) groups [69]. Each age group consisted of seven children, allowing us to observe how these developmental factors might influence their choice between voice and text input modalities.

3.5.5 Answering Type. Considering the effects of question type (open-ended or closed-ended [89]) on the input modalities (text and voice) [41], we categorized children's answers according to their type: Descriptive Answer Type (DAT) and Non-Descriptive Answer Type (NDAT) [67, 88]. DAT answers, like *"What did you eat before sleeping last night?"*, require descriptive information. NDAT answers, like *"When did you sleep last night?"*, involve time or numerical responses. In the sleep diary (Table 1), answers to six questions (Q1, Q3, Q4, Q5, Q6, and Q12) are DAT, while the other six are NDAT.

We focused our analysis on DAT responses because they are the most critical aspect of sleep diaries, providing the primary source of subjective sleep experiences and contextual factors. In contrast, NDAT responses primarily involve numbers and times, limiting their potential for capturing more detailed information. However, NDAT responses can be effectively supplemented with objective measures from actigraphy or wearable devices like smartwatches, providing a comprehensive view of sleep patterns.

3.5.6 Preferences. To understand participants' preferences of input modalities, we calculated the frequency of choosing each input modality over five days. A higher frequency of choosing one modality indicates a preference [104], allowing us to minimize personal bias and satisficing in subjective responses [56].

4 Measures

To address our research questions, we built metrics to analyze children's *sustained engagement* and *response quality*.

4.1 Assessing Sustained Engagement

Sustained engagement in sleep diaries encompasses both behavioral and emotional/cognitive dimensions, as outlined in Section 2.4.1. For *behavioral engagement*, previous studies have utilized metrics like *engagement duration*, *completion rate*, and *response length* [53, 109, 110]. In this paper, all children successfully completed their self-reporting over the five-day period, and no time limitations were imposed on individual responses. As a result, metrics like *engagement duration* and *completion rate* (100%) were not considered. Instead, we focused on *response length*, measured as the word count of a child's response to each question [30], calculated directly from the chat transcript. For *the second dimension* – emotional and cognitive engagement – we used the Giggle Gauge [26].

4.2 Assessing Response Quality

To assess the quality of children's responses, we drew inspiration from Gricean Maxims [35], focusing on four key aspects [110]: *informativeness*, *specificity*, *relevance*, and *clarity*. These principles help ensure that "cooperative" participants provide high-quality responses. Prior research has drawn from Shannon's information theory to calculate *informativeness* based on the *surprisal* of each word – the inverse of its expected frequency in modern English [38, 110]. However, due to children's limited vocabulary, which often results in low surprisal values, informativeness is not a suitable metric for evaluating their responses. Therefore, we focused on these three metrics: *information units (specificity)* [110], *relevance*, and *clarity*:

Information Units (IU). An IU represents the smallest integral piece of new information in a response, focusing on the elements that directly answer the question [98]. For instance, in response to the question, *"What did you eat before sleeping last night?"*, the answer: *"I ate apple and chocolate"* contains three IUs: *"I"*, *"apple"*, and *"chocolate"* (Table 2). In this context, the verb "ate" is not counted as an IU because it does not add new information beyond what is implied by the question, which specifically asks about the subject and food items. We used the *Spacy* package in Python to automatically count the number of IUs in each response based on this approach.

Relevance (R). In the context of sleep diaries, a quality response must be relevant to the question. Irrelevant responses add no value and complicate clinicians' analysis. Two researchers manually and independently assessed relevance on a three-level scale: 0 – Irrelevant, 1 – Somewhat Relevant, and 2 – Relevant (Table 3). Inter-rater reliability was calculated for these scores: $\kappa = 0.84$ (95% CI, 0.70 to 0.98).

Clarity (C). According to the Gricean Maxim of clarity, an effective response is easily understood without ambiguity. Clarity was scored on a three-level scale: 0 – illegible text, 1 – incomplete sentences or blur answer, 2 – clearly articulated response (Table 4). Two researchers manually and independently scored the clarify of each text response, with an inter-rater reliability of $\kappa = 0.91(95\%\ CI, 0.91 to 0.91)$.

Response Quality Index. We created an overall *RQI* for each response by aggregating the three quality metrics, following the approach of XIAO et al.,[110].

$$\text{RQI} = \text{IU} \times R \times C \tag{1}$$

This formula 1 allows us to quantify response quality, recognizing that even responses with high information content are not useful if they are irrelevant. For example, if the app asks: *"What did you do before sleeping last night?"*, and the child responds, *"I want to go to school."*, the answer, though informative, provides no useful data for clinicians. The value of RQI serves an indicator of the response quality rather than an absolute measure, offering a relative assessment of how well the response aligns with the intended question [110].

5 Result

The app functioned smoothly for most participants, with the exception of five children who accidentally deleted the app during their usage at home. We visited their homes to reinstall the app, ensuring they could continue with the experiment. Ultimately, all 20 children completed the study. A power analysis for repeated measures ANOVA, conducted with two age groups (younger vs. older children), indicated that a sample size of 20 participants achieves up to 86% power [100].

We collected a total of 1,200 responses over the five-day period, addressing three research objectives: 1) examining the effect of input modalities on children's *sustained engagement*, 2) evaluating *response quality*, and 3) understanding children's *preference and attitudes* toward input modalities in the sleep diary.

Table 2: Example of *Information Units* in a response to a question

Response	Information units (IU)	The value of IU
"I ate apple and chocolate."	I - apple - chocolate	3

Table 3: Example of the code to the *Relevance* in responses to a question

Question	Answer	The value of Relevance (R)
What did you eat before sleeping last night?	Irrelevant: *"I slept early."*	0
	Somewhat relevant: *"dinner"*	1
	Relevant: *"Apple and chocolate."*	2

Table 4: Example of the code to the *Clarity* in responses to a question

Question	Answer	The value of Clarity (C)
What did you eat before sleeping last night?	Illegible text: *"nij"*	0
	Incomplete sentences: *"dinner"*	1
	Clearly articulated response: *"Apple and chocolate."*	2

5.1 Influence of Input Modalities on Sustained Engagement

As outlined in Section 4.1, we examined children's sustained engagement across two dimensions: *response length* as a measure of behavioral engagement, and the results of the Giggle Gauge for emotional and cognitive engagement.

5.1.1 Behavioral Engagement. The results focus on: 1) the overall effect on response length, 2) the patterns of response length over time, 3) the effect of input modalities over time, and 4) the combined effects of input modalities and age groups over time.

The overall effect on response length. To investigate the effect of age groups and input modalities on children's engagement, we built a linear mixed-effects model (LMM) in R using lme4 package [7]. LMMs are used to analyze clustered data, such as repeated observations from the same participants over time [29]. Given our design, with multiple observations per participant, this model was well-suited to account for both random and fixed effects. In this model, *participants* were treated as a random effect to control for individual differences. The dependent variable was the *response length* of each answer. *Input modalities* (text vs. voice) and *age groups* (younger vs older group) were treated as fixed effects. The final model examined the effects of age groups, the answer type, and their interaction on the preference for input modality (Table 5).

Assumption test confirmed that the LMM met the necessary criteria. The residuals vs. fitted values plot indicated linearity, and the Q-Q plot showed approximate normality of residuals, despite the Shapiro-Wilk test ($W = 0.78, p < 0.05$) suggesting deviation. This aligns with previous findings that LMMs are robust to minor deviations from normality, particularly with larger datasets [85]. Visual inspections of residuals vs. predictor plots and random effects, which were approximately normal and centered around zero, further confirmed that the model's assumptions were adequately met.

The model revealed a statistically significant main effect of the input modality on response length ($p < 0.001$), with voice input resulting in responses that were, on average, 1.66 words longer than those using text input modality. The main effect of *AgeGroup* on response length was not significant ($p = 0.18$), indicating no substantial difference between age groups.

Table 5: Effects of model factors on predicting response length. The model formula is $RL \sim AgeGroup + InputModality + (1|ParticipantID)$**, where** RL **= response length,** $InputModality$ **= input modality (voice vs text),** $AgeGroup$**=the groups of age (younger vs older),** $ParticipantID$ **= participant ID.** *****p < 0.001,** ***p < 0.05.**

Variables	Estimate	SE	df	t	p
(Intercept)	1.64	0.12	55.87	14.29	$2e-16$ ***
InputModality	1.66	0.10	1075.36	4.54	$6.34e-06$ ***
AgeGroup	0.20	0.14	25.38	1.38	0.18

Patterns of response length over time. Our analysis using a LMM examined the effect of time (Day 1 to Day 5) on children's response length. Assumption testing confirmed that the LMM met the necessary criteria, as indicated by residual plots and random effects analysis. The results showed a significant negative effect of day on response length ($\beta = -0.07, p < 0.05$), indicating that response length tended to decrease slightly over the five-day period (Figure 5). However, this decline was not linear. The response length generally decreased over the first few days, then stabilized, with a slight recovery on Day 5, forming a gentle "U-shape" pattern.

The effect of input modalities over time. The model revealed a significant main effect of input modality on children's response length over time, with longer responses observed for voice input compared to text input ($\beta = 0.41, p < 0.05$). Figure 6 shows the response lengths over the five days for both text and voice input modalities. Specifically, on Day 1, voice input resulted in slightly longer responses than text input. A notable drop in response length was observed for text input on Day 2, followed by a decline in response length for voice input on Day 3. By Days 4 and 5, response

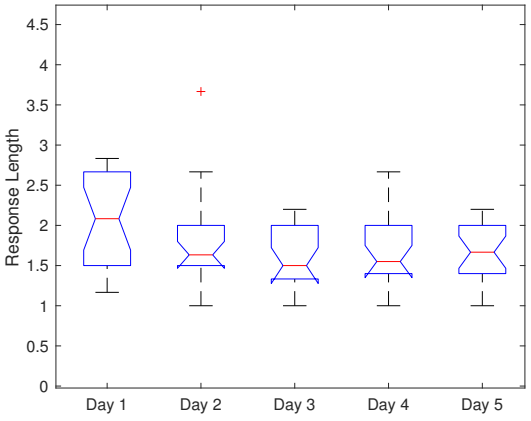


Figure 5: The distribution of response length over time.

lengths for the voice input modality had obviously recovered compared to text input.

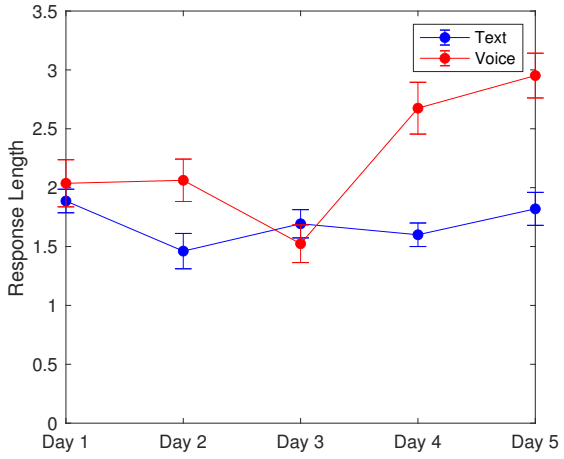


Figure 6: Patterns between two input modalities. Blue line = text input; red line = voice input; Y-axis = mean response length to each DAT.

The combined effects of input modalities and age groups over time. We observed a significant interaction effect between input modality and age groups ($\beta = -0.11, p < 0.05$), suggesting that the effect of input modality on response length varies by age group. Figure 7 displays the effect of input modalities (voice vs. text) and age groups (older vs. younger) on response length over five days. For voice input, responses from younger children showed a progressive increase in length, peaking on Day 5. In contrast, responses from older children were more variable, with a dip on Days 3, followed by a slight increase on Day 5. For text input, both age groups generally produced shorter responses. Responses from older children remained relatively low and stable, with a slight decrease toward the end. Responses from younger children also showed a decline in text input response length after Day 1, with a slight recovery on Day 5.

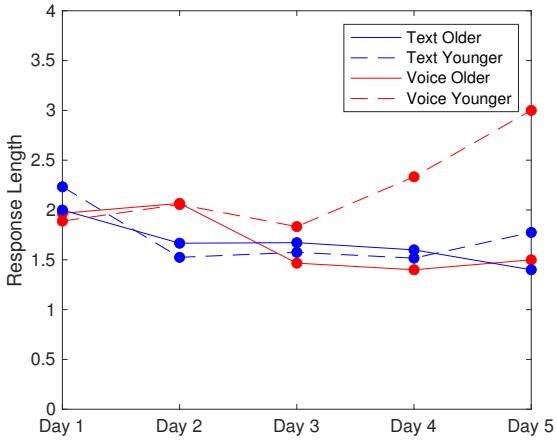


Figure 7: Patterns between two input modalities in two age groups. Red solid line = responses from older children with voice input; red dashed line = responses from younger children with voice; blue solid line = responses from older children with text input; blue dashed line = responses from younger children with text input. Y-axis = mean response length to each DAT.

5.1.2 Emotional and Cognitive Engagement. Twelve out of twenty children exhibited high engagement ($E \geq 3.6$), and seven showed moderate engagement ($3.0 \leq E < 3.6$). Additionally, seventeen out of twenty children rated 4 ($M = 3.82, SD = 0.39$) for *"I like how the app looked and felt"*, fourteen rated 4 ($M = 3.68, SD = 0.59$) for *"I enjoy using it"*, and fourteen rated 4 ($M = 3.65, SD = 0.67$) for *"I would like to do this again sometime"*. These results suggest that the app's aesthetic and sensory appeal, endurability, and overall interest are key factors in attracting and engaging children [26].

5.2 Impact of Input Modalities on Response Quality

Using Formula 1, we calculated the Response Quality Index (RQI). Our analysis primarily focused on: 1) the overall effect on response quality, 2) patterns of response quality over time, 3) the effects of input modalities over time, and 4) the combined effects of input modality and age groups over time.

Overall effect on response quality. To explore the effects of input modality and age group on response quality, we built a LMM. ParticipantID was treated as a *random effect*, while input modality (voice vs. text) and age group were *fixed effects*. Assumption testing was conducted, including checks for linearity (residuals vs. fitted value plots), normality of residuals (Q-Q plots), homoscedasticity, and the distribution and independence of random effects. These results confirmed that the model's assumptions were satisfied.

The findings revealed a statistically significant main effect of input modality on response quality ($\beta = 0.54, p < 0.05$), with voice input producing responses that were, on average, 0.41 units higher in quality than text input. In this context, one unit represents a change in the composite score derived from Formula 1 (the product of information units, relevance, and clarity), reflecting a meaningful improvement in the overall quality of the responses. The effect of age group on response quality was not statistically significant ($\beta = 0.50, p = 0.12$), indicating no significant difference between the different age groups.

Patterns of response quality over time. Figure 8 illustrates two distinct patterns of change in the response quality of 20 children over five days, categorized based on the fluctuations in their responses. These patterns are presented in two subplots, with each line representing an individual child's pattern: the first subplot reveals a "U-shaped" pattern in response quality. It starts at a relatively high level On Day 1, declines sharply, and then shows a notable recovery on Days 4 or 5. The second subplot presents a more stable pattern, where response quality remains relatively consistent over five days, with only minor fluctuations that are less pronounced than those in the first subplot.

The effect of input modalities over time. The model revealed a significantly effect of input modality on children's response quality over time, with higher responses quality observed for voice input compared to text input ($\beta = 0.33, p < 0.05$). Figure 9 illustrates two distinct "U-shape" patterns in response quality for voice and text input over five days. The response quality for voice input shows a gradual recovery, while the pattern for text input exhibits an opposite fluctuation. Specifically, voice input quality starts higher than with text input but drops to similar levels on Day 2. However, by Day 5, voice input quality recovers and significantly increases. In contrast, text input quality, which begins at a lower level, briefly stabilizes before declining further by the end of the study.

The combined effects of input modalities and age groups over time. We observed a significant interaction effect between input modality and age groups ($\beta = -0.06, p < 0.05$), suggesting that the effect of input modality on response quality varies by age group over time. Figure 10 shows how ages and input modalities affect response quality over time. Responses from older children using voice input displayed significant fluctuation – starting high, followed by a sharp decline, and then a strong recovery by Day 5. In contrast, the quality of responses from younger children using voice input exhibited a steadier pattern, maintaining consistent quality early on with gradual improvement by Day 5.

5.3 Children's Preferences and Attitudes towards Input Modalities

5.3.1 Children's Preference of Input Modalities in the Sleep Diary. To investigate the effects of age groups and answer types on children's preferences for input modalities, we built a linear mixed-effects model. In this model, participants were treated as a random effect to control for individual differences. The dependent variable was the preference for input modalities. Age groups (younger vs older), answer type (DAT vs NDAT), and their interaction were treated as fixed effects.

Our analysis focused on: 1) the overall effect of input modality, 2) overall preferences for input modality, 3) patterns in input modality preferences over time.

Overall effects of input modality. Analysis of 1,200 individual responses revealed a statistically significant main effect of *AgeGroup* on input modality preference ($\beta = -0.54, p < 0.01$). The negative estimate indicates that older children showed a lower preference for the voice modality compared to younger children. *AnswerType* also had a significant effect ($\beta = 0.04, p < 0.05$), with DAT being slightly more preferred. However, the interaction between *AgeGroup* and *AnswerType* was not statistically significant ($p = 0.70$), indicating that the effect of *AnswerType* on modality preference did not differ significantly between age groups. Furthermore, the result of the Again-Again table revealed that a significant interaction between input modality and age group ($F = 18.70, p < 0.001$), indicating that the durability and preferences for input modalities are age-dependent.

Overall preferences for input modality. We further measured the effect of *AgeGroup* and *AnswerType* on input modality preference, reporting the results in terms of overall choice. The t-values and p-values were calculated using Satterhwaite's approximation of degree of freedom [84]. Figure 11 shows a bar chart of modality preferences across the two age groups. The Chi-square test revealed significant differences in modality choice between age groups, $\chi^2_3 = 256.18, p < 0.001$, suggesting that age plays a crucial role in how children prefer to input information. Older children favored text input significantly more than younger ones. The large Chi-square value ($\chi^2_3 = 256.18$) indicates substantial differences in preference, likely due to factors like technology familiarity, cognitive processing speed, or comfort with different input methods.

Patterns in input modality preferences over time. The findings revealed distinct patterns in input modality preferences between younger and older children over the five-day period. Younger children increasingly favored the voice modality ($\beta = 0.02, p = 0.001 < 0.05$), while their preference for text input significantly declined over time ($\beta = -0.54, p < 0.001$). In contrast, older children initially preferred text input, showing a gradual but significant decrease in this preference over time ($\beta = 0.04, p < 0.05$), alongside an increase in their likelihood of selecting the voice modality ($\beta = -0.86, p < 0.001$). Figure 12 illustrates these trends, depicting the mean choice of input modality (voice vs. text) over five days for both age groups. While younger children showed a strong initial preference for voice that increased further, older children maintained a preference for text but gradually adopted voice input more frequently.

5.3.2 Children's Attitudes Toward Multimodal Sleep Diaries. We applied thematic analysis [45] to analyze the interview transcripts. The transcripts were coded to capture participants' experiences with the two input modalities, interactions with the chatbot, and their expectations of self-reporting activities at home.

Two coders independently developed an initial codebook through open coding of the data. After that, they collaborated to discuss individual codes, followed by a second round of independent coding using the emerging codebook. Finally, the coders reconvened to resolve disagreements, clarify coding details, and finalize the codebook. The coding process demonstrated high consistency and

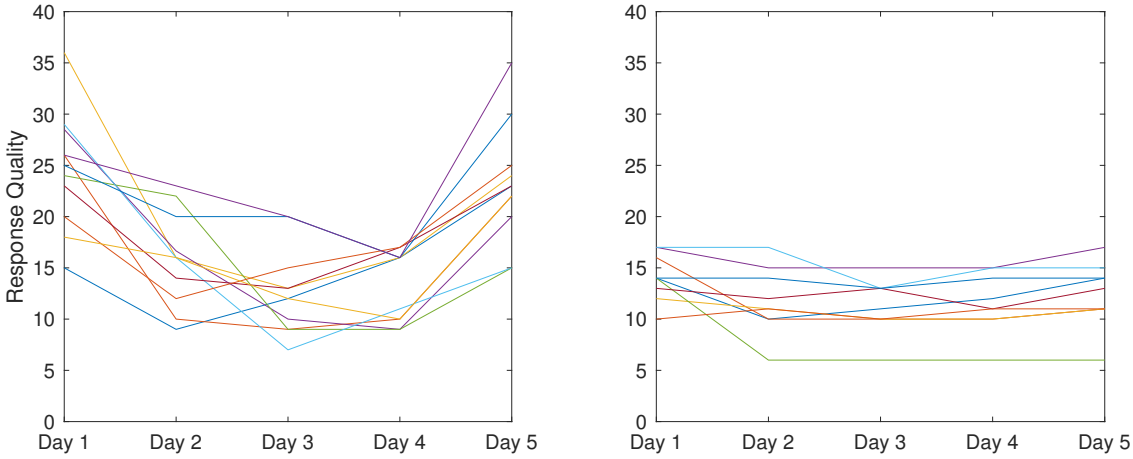


Figure 8: Observed patterns in response quality fluctuations over time. The Y-axis represents the total response quality of DAT. The first subplot illustrates a "U-shape" pattern, indicating an initial decline followed by recovery. The second subplot depicts a relatively steady pattern of response quality over five days.

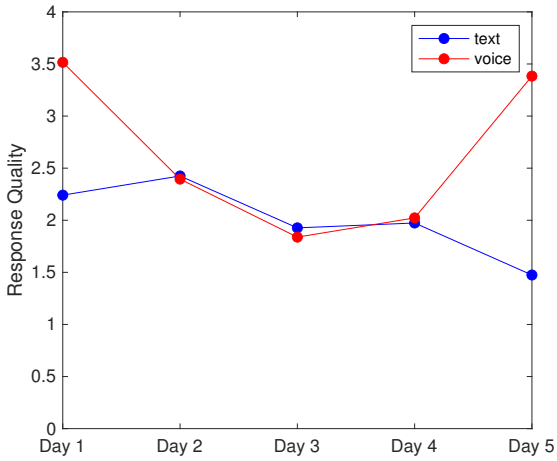


Figure 9: Patterns of response quality between input modalities. Blue line = text input; red line = voice input. Y-axis = mean response quality for each DAT.

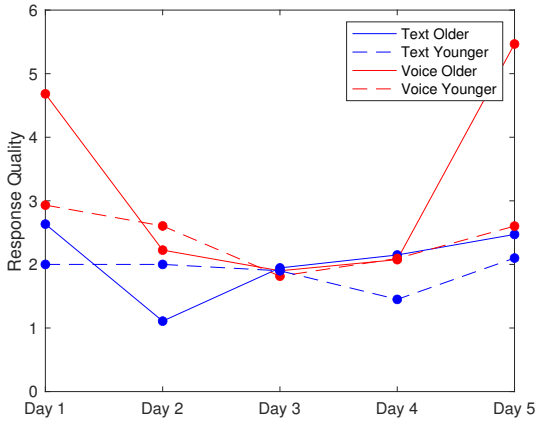


Figure 10: Effects of age groups (Older vs. Younger) and input modality (text vs. voice) on response quality over five days. X-axis = days of the study; Y-axis = average response quality.

accuracy, with an inner-rater reliability of 84%, as measured by Cohen's Kappa [60].

The findings were split into two high-level categories focusing on the positive and negative attitudes toward the sleep diary with multiple input methods.

(1) Positive attitude.

Advantages of voice input modality. A significant majority of the children, nine in total, expressed a clear preference for using the voice input modality in the sleep diary. The primary advantage of this modality was its ease of use, particularly when compared to text input. Many children found typing on a smartphone cumbersome due to the small size of on-screen keyboards, which often led to errors (C2, C3, C6, C8, C15). One child succinctly captured this frustration: *"I like talking because the keys on the phone are too small, sometimes I type wrong because of that."* (C3). Additionally, the voice input modality was appreciated for its speed and efficiency, with some children noting that speaking their responses were considerably faster than typing (C12). Beyond these practical benefits, the voice input modality also resonated with some children on a more personal level. They found the act of speaking more engaging than typing, which contributed to a more positive overall experience (C10, C8, C17). As one child said, *"I think talking is more fun. I don't*

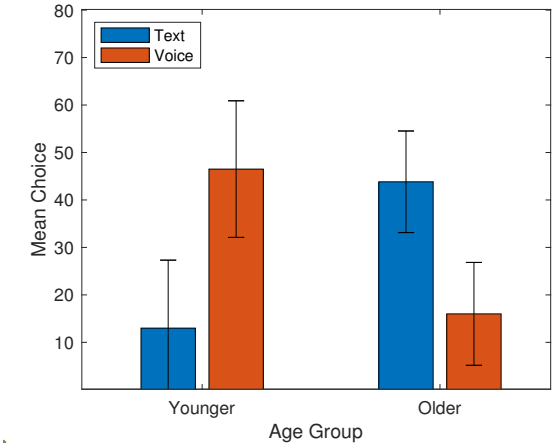


Figure 11: Overall Children's preference across input modalities. Error bars represent 95%CI, **p<0.001.

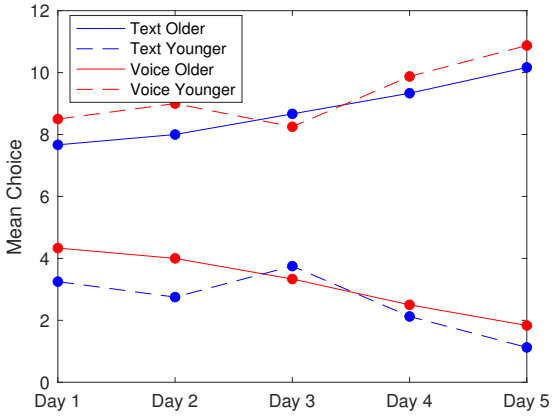


Figure 12: Children's choice over time. X-axis = days of the study; Y-axis = average modality choice per child.

like typing." (C8). This preference suggests that voice input not only simplifies the process but also makes it more interactive and less burdensome for children.

Advantages of text input modality. Conversely, five children favored the text input modality, highlighting different benefits that aligned more closely with their needs and preferences. One of the primary concerns with voice input was the accuracy of voice transcription, which some children found unreliable (C7, C9, C19). As one child explained, *"I prefer not talking aloud because it often goes wrong. I'm used to typing, so it's nicer to type."* (C7). This sentiment was echoed by another child who valued the precision that typing afforded, stating, *"I prefer typing because it is more accurate."*(C9). Another significant advantage of text input was its editability, which gave children greater control over their responses (C9, C11, C20). *"I can edit anything I want. But with talking, I can't modify what I've said."*(C11). This ability to refine their answers contributed to a sense of ownership and accuracy, which was important to some children. Moreover, the tactile and visual aspects of typing offered a more structured and controllable way for children to express their thoughts. One child noted, *"I found it easier to know where to press when typing than to know when I can talk and what to say."*(C14). This preference for text input reflects a desire for precision, control, and the ability to carefully craft responses, which some children found lacking in the voice input modality. Finally, the children preferring typing also stated that they preferred this modality as they are 'used to it' (C7). Therefore, children accustomed to text-based inputs might be more inclined to stick to this modality.

(2) Negative attitudes.

Repetitive content and decreased engagement. A recurring issue mentioned by nearly all the children was the monotony of answering the same set of questions in the sleep diary each day. This repetition gradually eroded their motivation and engagement over time, with one child comparing the experience to completing homework: *"There were always the same questions every day. It felt like homework."*(C3). This comparison highlights how the routine nature of the questions transformed the activity from a potentially engaging task into a mundane chore. To counteract this, some children incorporate more dynamic and entertaining content, such as jokes, to inject fun into the process and maintain their interest (C7, C8, C16). *"If possible, I hope the conversations can be more fun with jokes."* (C7). This feedback underscores the importance of variety and entertainment in maintaining the engagement of young users in daily tasks.

The conversational style. Several children also pointed out limitations in the conversational style of the sleep diary interfaces, which they felt could benefit from enhancements to make the interaction more appealing and motivating. Some children suggest enriching the conversational style in the interface to improve their motivation in daily self-reporting. For example, one child stated, *"You can use emojis, or use pictures to make it more fun."*(C1). This recommendation reflects a desire for a more visually stimulating and interactive experience that could transform routine data entry into a more enjoyable activity. Another child expressed dissatisfaction with the robotic nature of the voice used in the diary, describing it as lacking the warmth and relatability of human interaction: *"The voice sounds like a machine. I hope it is more like a human."*(C2). This feedback highlights the importance of creating a more natural and emotionally resonant user experience, particularly in tools designed for children, the sleep diary could better engage children and sustain their interest over time.

6 Discussion

Here we discuss the potential factors influencing children's sustained engagement and response quality, their preferences and attitudes toward input modalities (text and voice) over time, reflections of ethical considerations, and the limitations and future work of our study.

6.1 Towards Sustained Engagement

Firstly, we examined how the input modality influences on children's sustained engagement in the sleep diary, specifically using

response length as an indicator. Responses were, on average, 1.66 words longer with voice input than text input, suggesting that voice facilitate more detailed responses [76, 77]. This finding aligns with previous findings that input modality affects user behavior [34].

Voice input consistently elicited longer responses, particularly from younger children, while text input produced shorter responses across both age groups. While intriguing, this finding should be confirmed in future studies with more comprehensive data to ensure its robustness. Responses from older children were more variable, especially with text input, suggesting that factors like question content or familiarity with technology might play a role. These results imply that voice input may be more effective for eliciting detailed responses from younger children [76, 77], though these observations remain tentative due to the study's limitations. Further research is essential to confirm these patterns and understand the underlying mechanisms. For older children, additional strategies, such as incorporating interactive elements or mixed modalities, may enhance engagement and response detail.

The results of Giggle Gauge questionnaire further highlight the app's effectiveness in fostering both cognitive and emotional engagement. The combination of voice and text modalities accommodates individual preferences, enhancing the user experience and supporting sustained focus and positive emotions. These findings underscore the app's potential for broader applications in therapeutic contexts, where maintaining engagement is crucial.

6.2 Towards Higher Response Quality

Secondly, we examined the effect of input modalities on children's response quality in the sleep diary. Results revealed that voice input elicited higher-quality responses, with an average improvement of 0.54 units compared to text input. While this finding should be interpreted cautiously due to the limited sample size, it underscores the importance of understanding how input modality influences response quality – an area critical for e-health applications.

Further analysis of response quality patterns over five days identified two trajectories among the 20 participants. A "U-shaped" pattern suggested an initial decline in response quality, likely due to fatigue or adjustment challenges [8, 72], followed by recovery as children adapted to the task. In contrast, some children exhibited consistent response quality throughout, potentially indicating that these children were less affected by the factors causing the initial decline in the first group. These patterns highlight the importance of considering individual differences in task engagement and adaptation when designing interventions. Understanding these patterns can help tailor interventions or task designs to better support participants, particularly those who may require more time to adjust. However, given the short timeframe of study, it is important to acknowledge that these patterns might reflect short-term, context-dependent fluctuations rather than long-term trends. To establish the consistency of these patterns and to better understand their underlying factors, additional trials over a longer period are necessary.

Voice input demonstrated its potential for sustaining or enhancing response quality over time, especially as participants acclimated to the modality. Older children benefited the most, showing significant improvements with voice input, while their text input responses declined, possibly due to cognitive fatigue [8, 72]. Younger children, although actively engaged, showed smaller gains with voice input and relatively stable but lower-quality text responses. This suggests that younger children may require additional support or tailored approaches to fully leverage the advantages of voice input.

Combining these findings with sustained engagement analysis reveals a nuanced interaction between engagement and response quality. Younger children were more actively engaged and preferred voice input, while older children consistently provided higher-quality responses, likely influenced by cognitive and developmental differences. These results emphasize the potential of voice input to improve response quality, particularly in older children, while highlighting areas where additional support may help younger children sustain high-quality responses in self-reporting tasks.

6.3 Towards Preferred Input Modalities Over Time

Finally, we explored children's preferences and attitudes toward input modalities in the sleep diary over time.

The analysis revealed that age significantly influenced input modality preferences. Older children consistently preferred text input due to its familiarity and perceived ease of use, while younger children favored voice input but showed more variability, reflecting a context-dependent openness to both modalities. These findings challenge assumptions that younger children struggle with complex technologies, suggesting that they adapt based on situational factors. This aligns with previous research indicating that age impacts technology use and preference [52], while they also extending this understanding by highlighting younger children's willingness to explore both voice and text input. Furthermore, diverging preferences after Day 3 emphasize the need for adaptable user interfaces that cater to age-specific preferences in sustained engagement contexts. However, these observations are preliminary due to the study's short duration.

Children's attitudes toward input modalities also highlighted key insights. Many preferred voice input for its speed and ease, though transcription errors caused frustration, underscoring the need for improved accuracy. Conversely, children who favored text input appreciated its control and reliability, suggesting opportunities to enhance text-based interfaces. A hybrid system allowing seamless switching between modalities could address these diverse needs, balancing ease of use with precision. Additionally, children expressed a desire for more engaging content, such as jokes, emojis, and varied question sets, to counteract the monotony of repeated prompts. Incorporating gamification and rotating content could help maintain long-term motivation and engagement.

Overall, age strongly shapes input modality preferences, with younger children favoring voice for its fun and efficiency, and older children preferring text for accuracy. These insights provide a foundation for designing multimodal systems that accommodate diverse user needs while addressing challenges such as transcription errors and content monotony.

6.4 Ethical Considerations

Reflecting on the feedback of the chatbot-based sleep diary, our findings resonate with "agential realism" [5] in the HCI field. This theory posits that technologies are increasingly intertwined with human experiences, often appearing to exhibit a degree of agency and autonomy [83]. As artificial intelligence and interactive technologies evolve, the boundaries between human and machine agency become less distinct [87]. While this integration can enhance user experiences, it also raises ethical concerns, particularly for vulnerable populations such as children[54].

In our study, the chatbot appeared to act as an agent, potentially shaping children's emotional and social experiences, especially if it performed more human-like characteristics. While the chatbot may foster positive emotional connections, there is a risk unintended dependency, especially if children begin relying on it for emotional support or social interactions in ways that might hinder healthy development [55]. Previous studies on co-designing chatbot interactions with children offer valuable insights for our future work [32, 36, 91]. These studies emphasize the importance of careful system design to ensure that the chatbot provides appropriate emotional support while encouraging children to seek real-life connections and support from caregivers. Adopting this balanced approach can help mitigate risks associated with attachment and dependency [37].

6.5 Limitations and Future Work

We acknowledge five limitations in this study. *First*, the sample size, particularly with each age group, may limit the generalizability of the results. While the observed patterns are suggestive, a larger and more diverse sample could provide a more robust understanding of how age influences input modality preferences. Future research could benefit from a more nuanced categorization of age or a continuous approach to better capture age-related changes in modality preferences.

Second, the study was conducted over a relatively short period (five days), which may not be sufficient to capture longer-term preferences or learning effects that could further influence response quality. Future research could explore these factors over a longer period and with a more diverse participant pool to better understand the dynamics of response quality across different input modalities.

Third, the variability observed in the younger group's preferences may be influenced by unmeasured factors, such as individual differences in technology familiarity or specific daily task demands. Individual differences in participants' familiarity with voice or text input were not controlled for, which could have affected the results. Future studies should aim to control for these variables to better isolate the effect of age on input modality choice. Additionally, exploring other factors, such as task complexity and prior experience with technology, could provide deeper insights into how different user groups interact with various input modalities.

Fourth, the study's focus on only two input modalities (voice and text) may not fully capture the range of preferences that could emerge with other input options, such as gesture-based. Expanding the scope of input modalities in future research could provide a more comprehensive understanding of user preferences.

Fifth, while this study justifies the use of *response length* as a metric for behavioral engagement, it does not account for *ease of use* of input modalities as a potential mediator influencing children's behavioral engagement. Future studies will aim to investigate this effect in greater detail.

Finally, while sleep diaries are a valuable tool, particularly in the treatment of insomnia, the feedback obtained from ordinary children in this study may raise concerns among specialists regarding the reliability and applicability of the data in clinical contexts. The preferences and experiences of healthy children may differ slightly from those of patients suffering from insomnia, potentially leading to different patterns of diary usage and data accuracy. Therefore, it is crucial to approach these findings with caution when considering their application in therapeutic settings.

7 Conclusion

While voice-user interfaces are becoming increasingly popular, there is limited evidence on how this modality may support children's sustained self-reporting goals. This study addresses this gap by examining the effects of voice and text input modalities on children's sustained engagement and response quality in a sleep diary. Our findings suggest that while voice input helps younger children maintain engagement over five days, their response quality remains significantly lower than that of older children. Additionally, the identification of two distinct patterns of children's response quality over five days highlights the importance of considering individual differences in task responses over time. Moreover, age plays a crucial role in input modality preferences, with older children consistently favoring text input, while younger children show generally prefer voice input.

Given the study's limited sample size and short observation period, these findings should be interpreted as preliminary yet unique and valuable insights. Nonetheless, they underscore the potential benefits of incorporating voice input into traditional text-based sleep diaries could better meet the varied needs of children, potentially enhancing both sustained engagement and response quality. Further research with larger samples and longer observation periods is needed to confirm and expand upon these findings.

Acknowledgments

This work was supported by the China Scholarship Council. We are deeply thankful to the lovely children in the Netherlands who graciously shared their sleep stories with us.

References

[1] Tessa Aarts, Panos Markopoulos, Lars Giling, Tudor Vacaretu, and Sigrid Pillen. 2022. Snoozy: A Chatbot-Based Sleep Diary for Children Aged Eight to Twelve. In *Proceedings of the 21st Annual ACM Interaction Design and Children Conference* (Braga, Portugal) *(IDC '22)*. Association for Computing Machinery, New York, NY, USA, 297–307. doi:10.1145/3501712.3529718

[2] Christine Acebo, Avi Sadeh, Ronald Seifer, Orna Tzischinsky, Amy R Wolfson, Abigail Hafer, and Mary A Carskadon. 1999. Estimating sleep patterns with activity monitoring in children and adolescents: how many nights are necessary for reliable measures? *Sleep* 22, 1 (1999), 95–103. https://doi.org/10.1093/sleep/22.1.95

[3] Mehdi Alaimi, Edith Law, Kevin Daniel Pantasdo, Pierre-Yves Oudeyer, and Hélène Sauzeon. 2020. Pedagogical Agents for Fostering Question-Asking Skills in Children. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. doi:10.1145/3313831.3376776

[4] Daniel Avrahami, Kristin Williams, Matthew L. Lee, Nami Tokunaga, Yulius Tjahjadi, and Jennifer Marlow. 2020. Celebrating Everyday Success: Improving Engagement and Motivation using a System for Recording Daily Highlights. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. doi:10.1145/3313831.3376369

[5] Karen Barad. 2007. *Meeting the universe halfway: Quantum physics and the entanglement of matter and meaning* (1st ed.). Duke University Press Books, Durham, NC 27701 USA.

[6] Oscar A Barbarin and Barbara Hanna Wasik. 2011. *Handbook of child development and early education: Research to practice*. Guilford Press.

[7] Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2014. Fitting linear mixed-effects models using lme4. *Statistical Software* 67, 1 (2014), 1–48. 10.18637/jss.v067.i01

[8] Pazit Ben-Nun. 2008. Respondent Fatigue. In *Encyclopedia of Survey Research Methods*. Vol. 2. SAGE Publications, Inc., 742–743.

[9] Erin Beneteau, Olivia K. Richards, Mingrui Zhang, Julie A. Kientz, Jason Yip, and Alexis Hiniker. 2019. Communication Breakdowns Between Families and Alexa. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–13. doi:10.1145/3290605.3300473

[10] 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

[11] Richard R Bootzin and Perry M Nicassio. 1978. Behavioral treatments for insomnia. In *Progress in behavior modification*. Vol. 6. Elsevier, 1–45. https://doi.org/10.1016/B978-0-12-535606-0.50007-9

[12] Daniel J Buysse, Sonia Ancoli-Israel, Jack D Edinger, Kenneth L Lichstein, and Charles M Morin. 2006. Recommendations for a standard research assessment of insomnia. *Sleep* 29, 9 (2006), 1155–1173. https://doi.org/10.1093/sleep/29.9.1155

[13] 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).

[14] 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. https://doi.org/10.5665/sleep.1642

[15] Stella Chatzitheochari and Elena Mylona. 2021. Data quality in web and app diaries: A person-level comparison. *Journal of Time Use Research* 16, 1 (2021), 19–34. http://dx.doi.org/10.32797/jtur-2021-2

[16] Di (Laura) Chen, Dustin Freeman, and Ravin Balakrishnan. 2019. Integrating Multimedia Tools to Enrich Interactions in Live Streaming for Language Learning. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–14. doi:10.1145/3290605.3300668

[17] 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, Springer, Cham, 490–500. https://doi.org/10.1007/978-3-031-59717-6_32

[18] Yi Cheng, Kate Yen, Yeqi Chen, Sijin Chen, and Alexis Hiniker. 2018. Why doesn't it work? voice-driven interfaces and young children's communication repair strategies. In *Proceedings of the 17th ACM Conference on Interaction Design and Children* (Trondheim, Norway) *(IDC '18)*. Association for Computing Machinery, New York, NY, USA, 337–348. doi:10.1145/3202185.3202749

[19] Ronald D Chervin, James E Dillon, Kristen Hedger Archbold, and Deborah L Ruzicka. 2003. Conduct problems and symptoms of sleep disorders in children. *Journal of the American Academy of Child & Adolescent Psychiatry* 42, 2 (2003), 201–208. https://doi.org/10.1097/00004583-200302000-00014

[20] 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. https://doi.org/10.1097/00004583-200302000-00014

[21] Sudhansu Chokroverty. 2010. Overview of sleep & sleep disorders. *Indian Journal of Medical Research* 131, 2 (2010), 126–140.

[22] Dr. Nilong Vyas Danielle Pacheco. 2023. *Children and Sleep-An introduction to the importance of sleep in children and how to help them sleep better.* https://www.sleepfoundation.org/children-and-sleep

[23] Kailas Dayanandan and Brejesh Lall. 2024. Enabling Multi-modal Conversational Interface for Clinical Imaging. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI EA '24)*. Association for Computing Machinery, New York, NY, USA, Article 125, 13 pages. doi:10.1145/3613905.3650805

[24] Elisabeth Deutskens, Ko De Ruyter, Martin Wetzels, and Paul Oosterveld. 2004. Response rate and response quality of internet-based surveys: An experimental study. *Marketing letters* 15 (2004), 21–36. https://doi.org/10.1023/B:MARK.0000021968.86465.00

[25] 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* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 54, 15 pages. doi:10.1145/3411764.3445039

[26] Griffin Dietz, Zachary Pease, Brenna McNally, and Elizabeth Foss. 2020. Giggle gauge: a self-report instrument for evaluating children's engagement with technology. In *Proceedings of the Interaction Design and Children Conference* (London, United Kingdom) *(IDC '20)*. Association for Computing Machinery, New York, NY, USA, 614–623. doi:10.1145/3392063.3394393

[27] 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

[28] Joel E. Fischer, Stuart Reeves, Martin Porcheron, and Rein Ove Sikveland. 2019. Progressivity for voice interface design. In *Proceedings of the 1st International Conference on Conversational User Interfaces* (Dublin, Ireland) *(CUI '19)*. Association for Computing Machinery, New York, NY, USA, Article 26, 8 pages. doi:10.1145/3342775.3342788

[29] Andrzej Gałecki, Tomasz Burzykowski, Andrzej Gałecki, and Tomasz Burzykowski. 2013. *Linear mixed-effects model*. Springer.

[30] Mirta Galesic and Michael Bosnjak. 2009. Effects of questionnaire length on participation and indicators of response quality in a web survey. *Public opinion quarterly* 73, 2 (2009), 349–360. https://doi.org/10.1093/poq/nfp031

[31] Radhika Garg, Hua Cui, Spencer Seligson, Bo Zhang, Martin Porcheron, Leigh Clark, Benjamin R. Cowan, and Erin Beneteau. 2022. The Last Decade of HCI Research on Children and Voice-based Conversational Agents. 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 149, 19 pages. doi:10.1145/3491102.3502016

[32] Radhika Garg and Subhasree Sengupta. 2020. Conversational Technologies for In-home Learning: Using Co-Design to Understand Children's and Parents' Perspectives. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. doi:10.1145/3313831.3376631

[33] 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. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 4, 1, Article 11 (March 2020), 24 pages. doi:10.1145/3381002

[34] James J Gibson. 2014. *The ecological approach to visual perception: classic edition*. Psychology press.

[35] Herbert Paul Grice. 1975. Logic and conversation. *Syntax and semantics* 3 (1975), 43–58.

[36] Christine Grové. 2021. Co-developing a mental health and wellbeing chatbot with and for young people. *Frontiers in psychiatry* 11 (2021), 606041. https://doi.org/10.3389/fpsyt.2020.606041

[37] Ariel Han, Xiaofei Zhou, Zhenyao Cai, Shenshen Han, Richard Ko, Seth Corrigan, and Kylie A Peppler. 2024. Teachers, Parents, and Students' perspectives on Integrating Generative AI into Elementary Literacy Education. 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 678, 17 pages. doi:10.1145/3613904.3642438

[38] Xu Han, Michelle Zhou, Matthew J. Turner, and Tom Yeh. 2021. Designing Effective Interview Chatbots: Automatic Chatbot Profiling and Design Suggestion Generation for Chatbot Debugging. 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 389, 15 pages. doi:10.1145/3411764.3445569

[39] Alexis D. Henry. 2000. *Pediatric interest profiles: Surveys of play for children and adolescents, kid play profile, preteen play profile, adolescent leisure interest profiles*. Therapy Skill Builders.

[40] Emely Hoch, Yael Sidi, Rakefet Ackerman, Vincent Hoogerheide, and Katharina Scheiter. 2023. Comparing mental effort, difficulty, and confidence appraisals in problem-solving: A metacognitive perspective. *Educational Psychology Review* 35, 2 (2023), 61. https://doi.org/10.1007/s10648-023-09779-5

[41] Jan Karem Höhne, Konstantin Gavras, and Joshua Claassen. 2024. Typing or Speaking? Comparing Text and Voice Answers to Open Questions on Sensitive Topics in Smartphone Surveys. *Social Science Computer Review* (2024), 08944393231160961.

[42] Vanessa Ibáñez, Josep Silva, and Omar Cauli. 2018. A survey on sleep questionnaires and diaries. *Sleep medicine* 42 (2018), 90–96. https://doi.org/10.1016/j.sleep.2017.08.026

[43] Robert M Issenman and Iqbal H Jaffer. 2004. Use of voice recognition software in an outpatient pediatric specialty practice. *Pediatrics* 114, 3 (2004), e290–e293. https://doi.org/10.1542/peds.2003-0724-L

[44] Simon Jäger, Christopher Roth, Nina Roussille, and Benjamin Schoefer. 2024. Worker beliefs about outside options. *The Quarterly Journal of Economics* (2024),

qjae001. https://doi.org/10.1093/qje/qjae001
[45] Helene Joffe. 2011. Thematic analysis. *Qualitative research methods in mental health and psychotherapy: A guide for students and practitioners* (2011), 209–223. https://doi.org/10.1002/9781119973249.ch15
[46] Jeff Johnson. 2007. *GUI bloopers 2.0: common user interface design don'ts and dos*. Elsevier.
[47] Hyunhoon Jung, Hee Jae Kim, Seongeun So, Jinjoong Kim, and Changhoon Oh. 2019. TurtleTalk: An Educational Programming Game for Children with Voice User Interface. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI EA '19)*. Association for Computing Machinery, New York, NY, USA, 1–6. doi:10.1145/3290607.3312773
[48] Clare-Marie Karat, Christine Halverson, Daniel Horn, and John Karat. 1999. Patterns of entry and correction in large vocabulary continuous speech recognition systems. 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, 568–575. doi:10.1145/302979.303160
[49] 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
[50] Soomin Kim, Joonhwan Lee, and Gahgene Gweon. 2019. Comparing Data from Chatbot and Web Surveys: Effects of Platform and Conversational Style on Survey Response Quality. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. doi:10.1145/3290605.3300316
[51] Andreas M Klein, Jana Deutschländer, Kristina Kölln, Maria Rauschenberger, and Maria José Escalona. 2024. Exploring the context of use for voice user interfaces: Toward context-dependent user experience quality testing. *Journal of Software: Evolution and Process* 36, 7 (2024), e2618. https://doi.org/10.1002/smr.2618
[52] Radoslava Kraleva. 2017. Designing an Interface For a Mobile Application Based on Children's Opinion. *International Journal of Interactive Mobile Technologies* 11, 1 (2017). DOI:10.3991/ijim.v11i1.6099
[53] Alexander J Kull, Marisabel Romero, and Lisa Monahan. 2021. How may I help you? Driving brand engagement through the warmth of an initial chatbot message. *Journal of business research* 135 (2021), 840–850. https://doi.org/10.1016/j.jbusres.2021.03.005
[54] Adi Kuntsman. 2012. Introduction: Affective fabrics of digital cultures. In *Digital cultures and the politics of emotion: Feelings, affect and technological change*. Springer, 1–17. https://doi.org/10.1057/9780230391345_1
[55] Nomisha Kurian. 2024. 'No, Alexa, no!': designing child-safe AI and protecting children from the risks of the 'empathy gap'in large language models. *Learning, Media and Technology* (2024), 1–14. https://doi.org/10.1080/17439884.2024.2367052
[56] Tony CM Lam and Priscilla Bengo. 2003. A comparison of three retrospective self-reporting methods of measuring change in instructional practice. *The American Journal of Evaluation* 24, 1 (2003), 65–80. https://doi.org/10.1016/S1098-2140(02)00273-4
[57] Michal Luria, Rebecca Zheng, Bennett Huffman, Shuangni Huang, John Zimmerman, and Jodi Forlizzi. 2020. Social Boundaries for Personal Agents in the Interpersonal Space of the Home. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–12. doi:10.1145/3313831.3376311
[58] David C Mallinson, Maria E Kamenetsky, Erika W Hagen, and Paul E Peppard. 2019. Subjective sleep measurement: comparing sleep diary to questionnaire. *Nature and Science of Sleep* (2019), 197–206. https://doi.org/10.2147/NSS.S217867
[59] Stéphanie Mazza, Hélène Bastuji, and Amandine E Rey. 2020. Objective and subjective assessments of sleep in children: comparison of actigraphy, sleep diary completed by children and parents' estimation. *Frontiers in Psychiatry* 11 (2020), 495. https://doi.org/10.3389/fpsyt.2020.00495
[60] Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica* 22, 3 (2012), 276–282. https://hrcak.srce.hr/89395
[61] Jon F Miller and Robin S Chapman. 1981. The relation between age and mean length of utterance in morphemes. *Journal of Speech, Language, and Hearing Research* 24, 2 (1981), 154–161. https://doi.org/10.1044/jshr.2402.154
[62] G Milroy, Liam Dorris, and TM McMillan. 2008. Brief report: sleep disturbances following mild traumatic brain injury in childhood. *Journal of pediatric psychology* 33, 3 (2008), 242–247. https://doi.org/10.1093/jpepsy/jsm099
[63] Andreea Muresan and Henning Pohl. 2019. Chats with Bots: Balancing Imitation and Engagement. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI EA '19)*. Association for Computing Machinery, New York, NY, USA, 1–6. doi:10.1145/3290607.3313084
[64] Adrianna Murphy, Catherine McGowan, Martin McKee, Marc Suhrcke, and Kara Hanson. 2019. Coping with healthcare costs for chronic illness in low-income and middle-income countries: a systematic literature review. *BMJ global health* 4, 4 (2019), e001475. https://doi.org/10.1136/bmjgh-2019-001475
[65] Yoon Namkung and YouJin Kim. 2024. Learner engagement in collaborative writing: The effects of SCMC mode, interlocutor familiarity, L2 proficiency, and task repetition. *System* 121 (2024), 103251. https://doi.org/10.1016/j.system.2024.103251
[66] Shakked Noy and Whitney Zhang. 2023. Experimental evidence on the productivity effects of generative artificial intelligence. *Science* 381, 6654 (2023), 187–192. DOI:10.1126/science.adh2586
[67] Hyo-Jung Oh, Chung-Hee Lee, Hyeon-Jin Kim, and Myung-Gil Jang. 2005. Descriptive question answering in encyclopedia. In *Proceedings of the ACL 2005 on Interactive Poster and Demonstration Sessions* (Ann Arbor, Michigan) *(ACLdemo '05)*. Association for Computational Linguistics, USA, 21–24. doi:10.3115/1225753.1225759
[68] Luiza Superti Pantoja, Kyle Diederich, Liam Crawford, and Juan Pablo Hourcade. 2019. Voice Agents Supporting High-Quality Social Play. In *Proceedings of the 18th ACM International Conference on Interaction Design and Children* (Boise, ID, USA) *(IDC '19)*. Association for Computing Machinery, New York, NY, USA, 314–325. doi:10.1145/3311927.3323151
[69] Jean Piaget. 1976. *Piaget and his school: A reader in developmental psychology*. New York: Springer-Verlag.
[70] Laura Pina, Sang-Wha Sien, Clarissa Song, Teresa M. Ward, James Fogarty, Sean A. Munson, and Julie A. Kientz. 2020. DreamCatcher: Exploring How Parents and School-Age Children can Track and Review Sleep Information Together. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW1, Article 70 (May 2020), 25 pages. doi:10.1145/3392882
[71] Gary K Poock and EF Roland. 1982. *Voice recognition accuracy: What is acceptable?* Technical Report. Citeseer.
[72] Stephen R Porter, Michael E Whitcomb, and William H Weitzer. 2004. Multiple surveys of students and survey fatigue. *New directions for institutional research* 2004, 121 (2004), 63–73. https://doi.org/10.1002/ir.101
[73] Alexandros Potamianos and Shrikanth Narayanan. 2003. Robust recognition of children's speech. *IEEE Transactions on speech and audio processing* 11, 6 (2003), 603–616. doi:10.1109/TSA.2003.818026
[74] Kyrill Potapov, Asimina Vasalou, Victor Lee, and Paul Marshall. 2021. What do Teens Make of Personal Informatics? Young People's Responses to Self-Tracking Practices for Self-Determined Motives. 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 356, 10 pages. doi:10.1145/3411764.3445239
[75] Janet C Read, Stuart MacFarlane, and Chris Casey. 2002. Endurability, engagement and expectations: Measuring children's fun. In *Interaction design and children*, Vol. 2. Citeseer, 1–23. https://api.semanticscholar.org/CorpusID:16572839
[76] Melanie Revilla, Mick P Couper, Oriol J Bosch, and Marc Asensio. 2020. Testing the use of voice input in a smartphone web survey. *Social Science Computer Review* 38, 2 (2020), 207–224. https://doi.org/10.1177/0894439318810715
[77] Melanie Revilla, Mick P Couper, and Carlos Ochoa. 2018. Giving respondents voice? The feasibility of voice input for mobile web surveys. *Survey Practice* 11, 2 (2018). https://doi.org/10.29115/SP-2018-0007
[78] Sara E Rimm-Kaufman, Tim W Curby, Kevin J Grimm, Lori Nathanson, and Laura L Brock. 2009. The contribution of children's self-regulation and classroom quality to children's adaptive behaviors in the kindergarten classroom. *Developmental psychology* 45, 4 (2009), 958. https://doi.org/10.1037/a0015861
[79] Aki Rintala, Martien Wampers, Inez Myin-Germeys, and Wolfgang Viechtbauer. 2019. Response compliance and predictors thereof in studies using the experience sampling method. *Psychological assessment* 31, 2 (2019), 226. DOI:10.1037/pas0000662
[80] Kumar Rohit, Amit Shankar, Gagan Katiyar, Ankit Mehrotra, and Ebtesam Abdullah Alzeiby. 2024. Consumer engagement in chatbots and voicebots. A multiple-experiment approach in online retailing context. *Journal of Retailing and Consumer Services* 78 (2024), 103728. https://doi.org/10.1016/j.jretconser.2024.103728
[81] Sherry Ruan, Jacob O. Wobbrock, Kenny Liou, Andrew Ng, and James A. Landay. 2018. Comparing Speech and Keyboard Text Entry for Short Messages in Two Languages on Touchscreen Phones. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 4, Article 159 (Jan. 2018), 23 pages. doi:10.1145/3161187
[82] Marcel Ruoff, Brad A Myers, and Alexander Maedche. 2023. ONYX: Assisting Users in Teaching Natural Language Interfaces Through Multi-Modal Interactive Task Learning. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 417, 16 pages. doi:10.1145/3544548.3580964
[83] Pedro Sanches, Noura Howell, Vasiliki Tsaknaki, Tom Jenkins, and Karey Helms. 2022. Diffraction-in-action: Designerly Explorations of Agential Realism Through Lived Data. 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 540, 18 pages. doi:10.1145/3491102.3502029
[84] Franklin E Satterthwaite. 1946. An approximate distribution of estimates of variance components. *Biometrics bulletin* 2, 6 (1946), 110–114. https://doi.org/10.2307/3002019

[85] Holger Schielzeth, Niels J Dingemanse, Shinichi Nakagawa, David F Westneat, Hassen Allegue, Céline Teplitsky, Denis Réale, Ned A Dochtermann, László Zsolt Garamszegi, and Yimen G Araya-Ajoy. 2020. Robustness of linear mixed-effects models to violations of distributional assumptions. *Methods in ecology and evolution* 11, 9 (2020), 1141–1152. https://doi.org/10.1111/2041-210X.13434

[86] Hannah Schwarz, Melanie Revilla, and Wiebke Weber. 2020. Memory effects in repeated survey questions: Reviving the empirical investigation of the independent measurements assumption. *Survey Research Methods. 2020; 14 (3): 325-44.* (2020). DOI:10.18148/SRM/2020.V14I3.7579

[87] Woosuk Seo, Chanmo Yang, and Young-Ho Kim. 2024. ChaCha: Leveraging Large Language Models to Prompt Children to Share Their Emotions about Personal Events. 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 903, 20 pages. doi:10.1145/3613904.3642152

[88] Mohammad Shaharyar Shaukat, Mohammed Tanzeem, Tameem Ahmad, and Nesar Ahmad. 2021. Semantic similarity–based descriptive answer evaluation. In *Web semantics*. Elsevier, 221–231. https://doi.org/10.1016/B978-0-12-822468-7.00014-6

[89] Dong-Hee Shin and Kyung-mi Chung. 2017. The effects of input modality and story-based knowledge on users' game experience. *Computers in Human Behavior* 68 (2017), 180–189. https://doi.org/10.1016/j.chb.2016.11.030

[90] Michelle A Short, Teresa Arora, Michael Gradisar, Shahrad Taheri, and Mary A Carskadon. 2017. How many sleep diary entries are needed to reliably estimate adolescent sleep? *Sleep* 40, 3 (2017), zsx006. https://doi.org/10.1093/sleep/zsx006

[91] Lucas M. Silva, Franceli L. Cibrian, Clarisse Bonang, Arpita Bhattacharya, Aehong Min, Elissa M Monteiro, Jesus Armando Beltran, Sabrina Schuck, Kimberley D Lakes, Gillian R. Hayes, and Daniel A. Epstein. 2024. Co-Designing Situated Displays for Family Co-Regulation with ADHD Children. 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 124, 19 pages. doi:10.1145/3613904.3642745

[92] Amanda L Smith and Barbara S Chaparro. 2015. Smartphone text input method performance, usability, and preference with younger and older adults. *Human factors* 57, 6 (2015), 1015–1028. https://doi.org/10.1177/0018720815575644

[93] Catherine E Snow. 1996. Issues in the study of input: Finetuning, universality, individual and developmental differences, and necessary causes. *The handbook of child language* (1996), 179–193. https://doi.org/10.1111/b.9780631203124.1996.00007.x

[94] Christine J So, Matthew W Gallagher, Cara A Palmer, and Candice A Alfano. 2021. Prospective associations between pre-sleep electronics use and same-night sleep in healthy school-aged children. *Children's Health Care* 50, 3 (2021), 293–310. https://doi.org/10.1080/02739615.2021.1890078

[95] Tobias Sonne, Jörg Müller, Paul Marshall, Carsten Obel, and Kaj Grønbæk. 2016. Changing Family Practices with Assistive Technology: MOBERO Improves Morning and Bedtime Routines for Children with ADHD. 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, 152–164. doi:10.1145/2858036.2858157

[96] Micol Spitale, Silvia Silleresi, Giulia Cosentino, Francesca Panzeri, and Franca Garzotto. 2020. "Whom would you like to talk with?": exploring conversational agents for children's linguistic assessment. In *Proceedings of the Interaction Design and Children Conference* (London, United Kingdom) *(IDC '20)*. Association for Computing Machinery, New York, NY, USA, 262–272. doi:10.1145/3392063.3394421

[97] Anuj Tewari and John Canny. 2014. What did spot hide? a question-answering game for preschool children. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Toronto, Ontario, Canada) *(CHI '14)*. Association for Computing Machinery, New York, NY, USA, 1807–1816. doi:10.1145/2556288.2557205

[98] V Ya Tsvetkov. 2009. Information objects and information Units. *European Journal of Natural History* 5, 2 (2009), 99.

[99] 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. https://doi.org/10.5664/jcsm.10246

[100] Robert A Virzi. 1992. Refining the test phase of usability evaluation: How many subjects is enough? *Human factors* 34, 4 (1992), 457–468. https://doi.org/10.1177/001872089203400407

[101] Kathleen D Vohs and Roy F Baumeister. 2016. *Handbook of self-regulation: Research, theory, and applications.* Guilford Publications.

[102] Tudor Văcăreţu, Nikolaos Batalas, Begum Erten-Uyumaz, Merel Van Gilst, Sebastiaan Overeem, and Panos Markopoulos. 2019. Subjective sleep quality monitoring with the hypnos digital sleep diary: Evaluation of usability and user experience. In *12th International Conference on Health Informatics, HEALTHINF 2019-Part of 12th International Joint Conference on Biomedical Engineering Systems and Technologies, BIOSTEC 2019*. SciTePress Digital Library, 113–122.

[103] 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. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 506 (Nov. 2022), 27 pages. doi:10.1145/3555619

[104] Can Wang, Tao Bo, Yun Wei Zhao, Chi-Hung Chi, Kwok-Yan Lam, Sen Wang, and Min Shu. 2018. Behavior-Interior-Aware User Preference Analysis Based on Social Networks. *Complexity* 2018, 1 (2018), 7371209. https://doi.org/10.1155/2018/7371209

[105] Misuhiro Watanabe, Kazuki Mitsui, Kazunori Sato, Seiko Nakano, Yasuhisa Koide, et al. 2023. A 5-month comparative study of Japanese input speed by keyboard of elementary school children learning with 1: 1 devices for the first time. *International Journal of Learning Technologies and Learning Environments* 6, 1 (2023). https://doi.org/10.52731/ijltle.v6.i1.723

[106] 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

[107] Jay G Wilpon and Claus N Jacobsen. 1996. A study of speech recognition for children and the elderly. In *1996 IEEE international conference on acoustics, speech, and signal processing conference proceedings*, Vol. 1. IEEE, 349–352. https://api.semanticscholar.org/CorpusID:28648133

[108] Chi-Hsuan Wu, Shih-yang Liu, Xijie Huang, Xingbo Wang, Rong Zhang, Luca Minciullo, Wong Kai Yiu, Kenny Kwan, and Kwang-Ting Cheng. 2024. CMOSE: Comprehensive Multi-Modality Online Student Engagement Dataset with High-Quality Labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4636–4645.

[109] Meng-Hsin Wu, Su-Fang Yeh, XiJing Chang, and Yung-Ju Chang. 2021. Exploring Users' Preferences for Chatbot's Guidance Type and Timing. In *Companion Publication of the 2021 Conference on Computer Supported Cooperative Work and Social Computing* (Virtual Event, USA) *(CSCW '21 Companion)*. Association for Computing Machinery, New York, NY, USA, 191–194. doi:10.1145/3462204.3481756

[110] 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

[111] Zhemin Zhu, Delphine Bernhard, and Iryna Gurevych. 2009. A multi-dimensional model for assessing the quality of answers in social Q&A sites. In *ICIQ*. 264–265.

[112] David Zyngier. 2008. (Re) conceptualising student engagement: Doing education not doing time. *Teaching and teacher education* 24, 7 (2008), 1765–1776. https://doi.org/10.1016/j.tate.2007.09.004