BetaDAPR: An AI-based Expert Support System for Art Therapists with Qualitative and Quantitative Assistance

MIGYEONG YANG*, Sungkyunkwan University, Republic of Korea
CHAEHEE PARK, Sungkyunkwan University, Republic of Korea
HYUNSEON WON, Sungkyunkwan University, Republic of Korea
TAEEUN KIM, Seoul Women's University, Republic of Korea
HAYEON SONG, Sungkyunkwan University, Republic of Korea
JINYOUNG HAN†, Sungkyunkwan University, Republic of Korea

Sketch-based drawing assessments are valuable for understanding cognitive and psychological states but rely on therapists' expertise, making them labor-intensive. While a few automated methods have been proposed, their capability is limited to supporting quantitative analysis, which falls short of qualitative needs in art therapy. To overcome this limitation, we introduce *BetaDAPR*, an expert support system offering both qualitative and quantitative assistance for large-scale drawing assessments. To investigate how the presence or absence of our system's qualitative assistance is perceived depending on the art therapist's level of expertise, we conducted a 2×2 factorial design experiment. Results showed that qualitative support improved therapists' intention to use the system and increased trust, especially for senior therapists, enhancing acceptance of its physical presence. Interviews revealed that AI assistance is differently perceived by the therapists' experience level, affecting system evaluations. We believe our findings shed light on developing AI-based expert support systems for art therapy.

CCS Concepts: • **Human-centered computing → Empirical studies in HCI**; • **Applied computing** → *Arts and humanities*; • **Information systems → Expert systems**; • **Computing methodologies → Artificial intelligence**.

Additional Key Words and Phrases: Art therapy, Drawing assessments, Expert support system, AI, AI perception, User experience, Qualitative support, Professional experience

ACM Reference Format:
Migyeong Yang, Chaehee Park, Hyunseon Won, Taeeun Kim, Hayeon Song, and Jinyoung Han. 2026. BetaDAPR: An AI-based Expert Support System for Art Therapists with Qualitative and Quantitative Assistance. *Proc. ACM Hum.-Comput. Interact.* 10, 6, Article CSCW055 (October 2026), 34 pages. https://doi.org/10.1145/3816903

1 INTRODUCTION

Sketches are a valuable tool for analyzing psychological and cognitive states. For instance, a person's sketch can reveal various mental conditions, such as stress and burnout [49, 95], cognitive development [71], and depression [32]. Unlike self-report surveys, drawing assessments offer

*Currently at NAVER AI Lab (migyeong.yang@navercorp.com).
†Corresponding author.

Authors' addresses: Migyeong Yang, Sungkyunkwan University, Seoul, Republic of Korea, mgyang@g.skku.edu; Chaehee Park, Sungkyunkwan University, Seoul, Republic of Korea, chaeheepark@g.skku.edu; Hyunseon Won, Sungkyunkwan University, Seoul, Republic of Korea, dprth1014@g.skku.edu; Taeeun Kim, Seoul Women's University, Seoul, Republic of Korea, tekim@swu.ac.kr; Hayeon Song, Sungkyunkwan University, Seoul, Republic of Korea, songhy@skku.edu; Jinyoung Han, Sungkyunkwan University, Seoul, Republic of Korea, jinyounghan@skku.edu.

ACM 2573-0142/2026/10-ARTCSCW055
https://doi.org/10.1145/3816903

unique insights into an individual's psychological state, uncovering preconscious or unconscious aspects that verbal reports might not capture [5, 35, 61]. Consequently, tools such as Draw-a-Person, Draw-A-Person-in-the-Rain, and House-Tree-Person are commonly used as supplementary methods for initial psychological evaluations in various fields, including education [55, 71] and clinical settings [49, 95]. Additionally, the simplicity and accessibility of sketching make these assessments particularly useful for large groups, such as in educational institutions [21, 31, 39] and professional environments [49].

However, interpreting drawing assessments requires extensive expert evaluation to identify psychological indicators according to established scoring criteria, e.g., how individuals represent human figures and their environments in sketches. This manual scoring process is both time-consuming and resource-intensive. Additionally, the results often depend on the evaluators' expertise, necessitating expensive inter-rater reliability assessments among multiple experts [95]. These challenges are particularly acute in real-world contexts where art therapists conduct large-scale assessments. As art therapy traditionally operates through individual, face-to-face sessions where therapists maintain close attention to each client's responses [34, 65], institutional demands which increasingly require assessments at scale are burdensome. For instance, in educational institutions, therapists evaluate 20–30 students simultaneously for psychological screening [21, 31], and corporate wellness programs employ drawing assessments to monitor employee stress levels across entire organizations [49]. In these group settings—which constitute 'large-scale' relative to the traditional one-on-one therapeutic context—therapists face competing demands: the need for timely feedback to inform interventions, budget constraints that limit the number of expert hours available, and the documentation requirements for each participant [80, 91]. These practical constraints underscore the need for technological support that can assist practitioners without compromising the depth essential to drawing assessment interpretation.

To address this issue, automated methods for analyzing drawing assessments have been developed to standardize the evaluation process, reduce expert biases, and assist human evaluators by cutting down on time and costs. Recent advancements have focused on deep learning techniques aiming at evaluating drawing-based cognitive assessments. For instance, Salar et al. [77] focused on object detection (e.g., people) as fundamental elements for assessment. Additionally, some studies have investigated various scoring criteria, such as head size [94] and emotional expressions [25]. In the realm of emotion analysis, facial features in sketches are first detected, and then emotions like happiness and sadness are predicted [25]. Taking a step further, more recent work by Kim et al. [46] and Kang et al. [43] proposed an expert support system for analyzing drawing assessments based on an object detection model.

While previous studies have provided valuable insights, they have primarily focused only on the *quantitative* capability, i.e., automatic scoring. This may pose a significant limitation in drawing assessments where *qualitative* analysis is crucial. For example, post-drawing inquiry (PDI) plays an important role in obtaining client information in drawing assessments [65]. However, in large-scale settings, asking questions to a client in person is not feasible, which may result in inaccurate judgments due to the lack of information about the client. Also, the drawing assessment report for each client is another crucial qualitative analysis outcome [76]. This documentation allows art therapists to maintain their focus on their clients' status, ultimately optimizing therapeutic outcomes [34]. However, in large-scale settings, the task of writing individual reports can be a substantial workload, which in turn often leads to procrastination [37] or reliance on anecdotal reports [34].

To overcome these challenges, we propose *BetaDAPR*, an expert support system capable of providing both qualitative and quantitative assistance during large-scale drawing assessments.

Specifically, our system provides qualitative support through AI-generated personalized post-drawing inquiries and automated report drafting, alongside quantitative support via automated DAPR scoring. Among the widely used drawing assessments, we selected the Draw-A-Person-in-the-Rain (DAPR) assessment, an advanced version of the Draw-A-Person (DAP) assessment. The DAPR assessment builds on the insights offered by the DAP assessment and additionally provides important information about participants' stress levels and coping mechanisms [55, 95]. The proposed system consists of two components: (i) a drawing assessment system and (ii) a drawing assessment analysis system. First, the drawing assessment system is a tablet PC-based tool that clients interact with. Within the system, clients input their information, draw a sketch according to the provided instructions, and answer post-drawing inquiries (PDIs). These inquiries consist of five questions commonly used in art therapy and two questions generated by the AI tailored to the individual client. Afterward, the client answers his/her anxiety scale. Through this process, client data can be comprehensively collected, even in large-scale settings. Additionally, by conducting personalized PDIs for each client, art therapists can obtain detailed information about each client. Based on the collected client data, the art therapist utilizes the drawing assessment analysis system. The system presents client information, DAPR analysis results, and the distribution of clients. In this system, art therapists can review both qualitative results (i.e., PDI responses and summary reports) and quantitative results (i.e., DAPR scores and distribution of DAPR scores), which assists the art therapists in making their final assessment.

In the proposed *BetaDAPR*, we mainly developed three following modules: (i) a user-specific PDI generation module, (ii) a summary report generation module, and (iii) a DAPR scoring model. The user-specific PDI generation module, a function integrated into the drawing assessment system, aims to elicit further information from the client by generating personalized questions. As asking proper questions to the client is crucial for obtaining valuable insights [65], we developed a module that generates personalized PDIs resembling those posed by actual art therapists, utilizing questions formulated by art therapists for each image. Experiments involving 16 art therapists demonstrated that these PDIs are well-suited to the client's drawings, indicating their potential usefulness in real-world, large-scale settings. Based on the client's responses to the generated PDIs and their drawings, the summary report generation module concisely documents the analysis results in terms of DAPR, generating a report. This aims to alleviate the workload of art therapists by generating a report draft. To achieve this, we utilized actual client reports written by art therapists. Experiments with 16 art therapists showed that the generated reports are of sufficient quality to serve as drafts. In addition to these qualitative analysis supports, we developed the DAPR scoring module to provide assistance through more accurate information in quantitative analysis as well. In calculating the score, previous models primarily focused on object detection [46, 77, 94], hindering in-depth analysis. To address this limitation, we proposed to combine an MLLM (Multimodal Large Language Model) with object detection methods for comprehensive evaluation. Consequently, we confirmed that our module's results closely align with the evaluations of art therapists, demonstrating significantly superior performance compared to existing models.

To evaluate our proposed system, we first collected data from 37 individuals aged 6 to 61 using the drawing assessment system. 60 art therapists then used the drawing assessment analysis system and participated in surveys and interviews. To ascertain whether our qualitative assistance, in addition to quantitative support, is beneficial to art therapists, we conducted a comparative study with the existing system, *AlphaDAPR* [46], which only provides quantitative support. In addition, in line with previous research suggesting that the perception of AI support varies depending on experience levels [46, 53], we conducted a 2×2 factorial design. Through this, we aim to investigate the effects of two independent variables, each with two levels (*AlphaDAPR/BetaDAPR* and *Junior/Senior*), resulting in four distinct experimental conditions. The results with 60 art therapists demonstrated that both

Junior and *Senior* groups exhibited a significantly higher intention to use *BetaDAPR* compared to *AlphaDAPR*, which highlights the effect of qualitative assistance. Regarding professional experience, *Junior* group tended to show higher AI usage and perceive AI as an assistant therapist compared to *Senior* in the interview study. However, contrary to this perception, the *Senior* group, particularly those who used *BetaDAPR*, showed high levels of trust and physical presence compared to other groups. A mediation analysis revealed that the high quality and comprehensive support (i.e., both quantitative and qualitative assistance) led to increased trust, which in turn made participants feel that they were actually in the DAPR analysis environment. Likewise, different results were observed depending on the two independent variables, suggesting the importance of considering these factors when proposing expert support systems in art therapy. The contribution of this work is threefold:

- To the best of our knowledge, we proposed *BetaDAPR*, the first expert support system that integrates quantitative scoring with qualitative interpretive assistance for large-scale Draw-A-Person-in-the-Rain (DAPR) assessments.
- We develop three novel AI modules: a user-specific PDI generation module that produces therapist-like follow-up questions, a summary report generation module that drafts concise DAPR interpretations, and a DAPR scoring module that outperforms a prior scoring system and aligns closely with expert therapist judgments.
- A 2 × 2 between subjects study with 60 art therapists (*AlphaDAPR/BetaDAPR* and *Junior/Senior*) demonstrates that adding qualitative assistance significantly increases intention to use, satisfaction and perceived usefulness. The study further reveals experience-dependent differences in the perceived value and usage patterns of the system.

2 BACKGROUND

2.1 Draw-A-Person-in-the-Rain Assessment

The Draw-A-Person-in-the-Rain (DAPR) assessment is a prominent tool in psychotherapeutic practices, used for monitoring progress or conducting group screenings [55, 95]. This assessment allows experts to evaluate how individuals deal with and respond to stress [39, 49, 55, 95]. Participants are instructed to draw a person or people in the rain, and these drawings are then reviewed by a human expert, usually an art therapist, using the DAPR assessment scale [51, 89]. Within this framework, elements such as rain and puddles indicate perceived stress levels, while resources like umbrellas represent coping mechanisms. However, the manual scoring process is labor-intensive and time-consuming, particularly with large groups like school populations, as it involves detailed analysis of stress-related and resource-related elements and their interactions. Integrating deep learning techniques as supplementary tools could streamline this process, facilitating faster preliminary analysis and allowing experts to focus on more in-depth decision-making.

2.2 AI Technology in Art Therapy

With the advancement of AI technologies in interpreting images and sketches, AI-based methods have been increasingly adopted to analyze drawing assessments in art therapy [23, 71]. Deep learning approaches, such as VGG-16, have been utilized to evaluate dementia severity by analyzing clock drawings [23]. Kim et al. [47] utilized deep neural networks (DNN) to classify individuals' psychological states and estimate their numerical values through the interpretation of drawings. Rakhmanov et al. [71] curated a Draw-A-Person assessment dataset from elementary students and employed traditional machine learning algorithms alongside convolutional neural network models to classify children's mental development stages. Additionally, a model was developed to effectively detect objects in drawings for assessment purposes [77]. Similarly, various scoring criteria in drawing assessments, such as head size [94] and emotional expression [25], were also analyzed. Recently, taking a step further, expert support systems that support professionals with the scoring result of drawing assessments have been proposed [43, 46].

Although existing studies can be helpful to art therapists, they have limitations in that they only focused on the quantitative capability, i.e., calculating scores. In practice, qualitative aspects, e.g., post-drawing inquiry (PDI), were extensively analyzed by art therapists, which make more accurate judgments [65], and it is an area where they actually face difficulties [34, 37]. Therefore, we propose *BetaDAPR*, which provides not only quantitative supports but also qualitative assistance. Furthermore, while prior work mostly relied on existing object detection methods, which have a limitation in capturing the complex characteristics present in drawing assessment analysis scales (i.e., the style of rain), *BetaDAPR* can capture a context information of a given sketch, which is important in comprehensive drawing analysis.

2.3 Perceptions of AI among Domain Experts

Therapy is a complex cooperative process sustained by extensive articulation work, the ongoing coordination of tasks, roles, and information across people, tools, and records [9, 75]. As Schmidt and Bannon [79] originally conceptualized, articulation work encompasses the effort required to manage the distributed nature of cooperative work, making activities of multiple actors coherent. This coordination work often intensifies as invisible labor in remote and hybrid care [26, 60]. With the introduction of AI systems into therapeutic workflows, new forms of articulation work emerge, requiring practitioners to coordinate not only with human colleagues but also with algorithmic actors [56, 99]. Recent work highlights that AI integration fundamentally reshapes these coordination practices, demanding explicit attention to how human expertise and machine capabilities are distributed and managed within collaborative work arrangements [24, 30]. In this setting, AI has been increasingly integrated into medical and therapeutic practice [38, 86, 88] and is valued for efficiency and consistency [18, 29, 69]. Importantly, many professionals view it as an assistive rather than substitutive technology, acknowledging that while it cannot replicate essential human competencies [18], it can take over selected tasks such as drafting initial notes or recommendations for expert review [28, 90]. This reframes the primary design goal: not to create an engine that replaces professional judgment, but to develop workflow-aware tools that support the complex articulation work among actors [45, 70, 87].

Despite this potential, experts' perceptions of AI are often ambivalent. While acknowledging its role as a helpful assistant [18, 28, 29], many raise significant concerns about displacement and the risks of error in therapeutic domains, such as patient safety [90], medico-legal exposure [69], and the emotional burden of false positives [15]. Addressing this ambivalence requires cultivating professional trust, a precondition for adopting expert support systems [12, 81]. This trust is primarily built on explainability, as evidence shows that professionals are more willing to use AI when they can understand and verify its reasoning via visual aids or textual explanations [1, 38, 66, 88]. It deepens when the system is designed for collaboration, which means clarifying the respective roles of the AI and the user [54, 63] while ensuring ultimate accountability remains with the human expert [28, 90]. Consequently, successful adoption hinges on an unobtrusive fit into existing procedures through interfaces that surface explanations and uncertainties [6, 13, 98], helping to position AI as a reliable partner in therapeutic practice [20, 40, 100]. However, these safeguards may not fully resolve concerns when AI touches qualitative and interpretive labor.

And yet, even when systems follow these trust-building principles, recent work suggests that AI assistance in such subjective tasks can paradoxically heighten concerns about job replacement. This anxiety appears particularly pronounced among junior experts, as they perceived that the AI begins to handle tasks once considered core to human expertise [46, 53]. To examine this tension, we employed a 2 × 2 factorial design to investigate the impact of these factors on adopting AI-based supporting systems in art therapy: qualitative support in the system (two levels: absence (*AlphaDAPR*) and presence (*BetaDAPR*)) and experience of art therapists (two levels: *Juniors* and *Seniors*).

3 BETADAPR

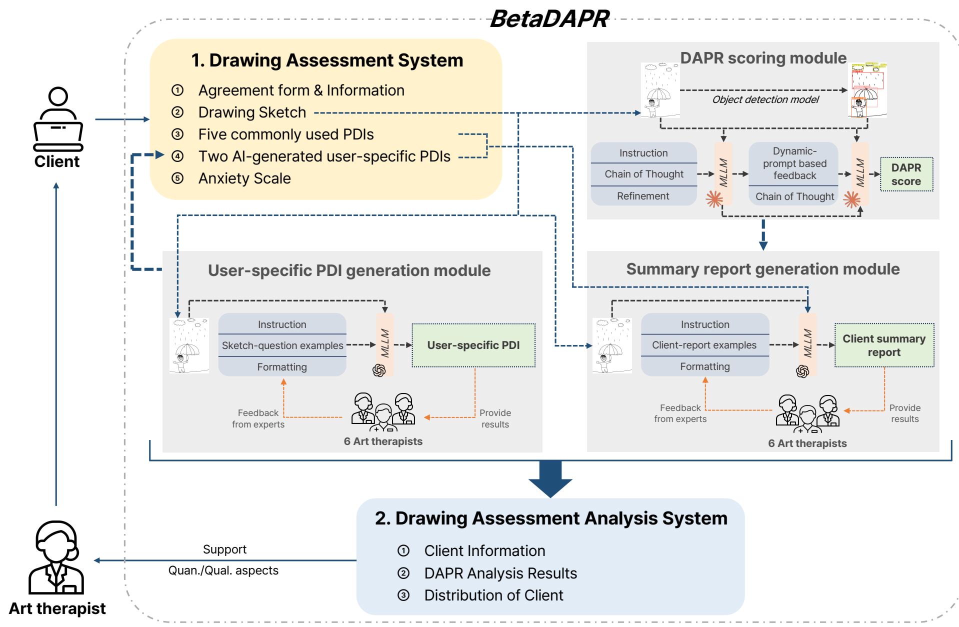
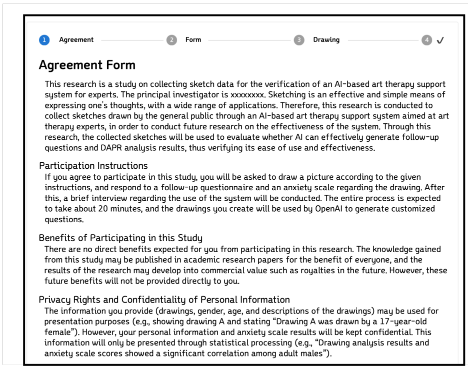


Fig. 1. Overall system process of *BetaDAPR*.

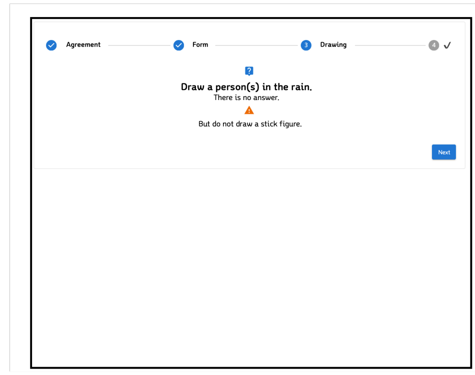
BetaDAPR is designed to support art therapists in large-scale assessment contexts where traditional one-on-one observation are impractical. As illustrated in Figure 1, the system consists of two components that map onto distinct phases of the clinical workflow: (i) a drawing assessment system used during sessions to collect client data, and (ii) a drawing assessment analysis system used after sessions to support therapists' interpretation and decision-making. While we recognize the value of extensive observation in traditional therapeutic settings, *BetaDAPR* specifically addresses scenarios where such individualized attention is not feasible due to scale constraints.

3.1 Drawing Assessment System

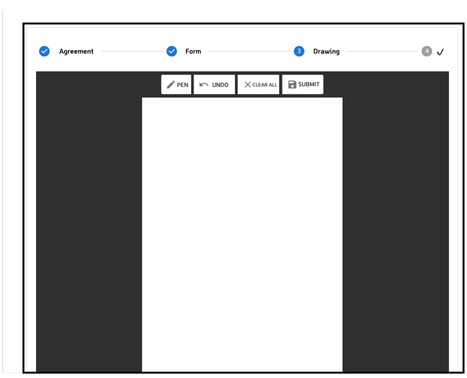
As shown in Figure 2, a drawing assessment system was developed that can be used on tablet PCs to collect clients' information and sketches. The system begins by explaining the purpose of drawing assessment, and an agreement form is provided to secure users' voluntary consent (see Figure 2(a)). Following this, users are asked to provide basic information such as their names, ages, and genders. Next, as depicted in Figure 2(b), users are given a maximum of 15 minutes to complete their drawings with the following instruction: *"It is raining. Please draw a person in the rain. You are free to choose the content and direction of the drawing, but ensure the figure is fully formed, not just a simple stick figure."* Users then proceed to create their drawings freely, as shown in Figure 2(c). After completing the drawing, participants respond to five predefined post-drawing inquiries (PDIs),



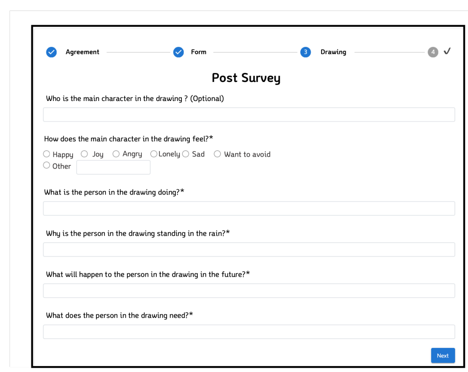
(a) Agreement Form



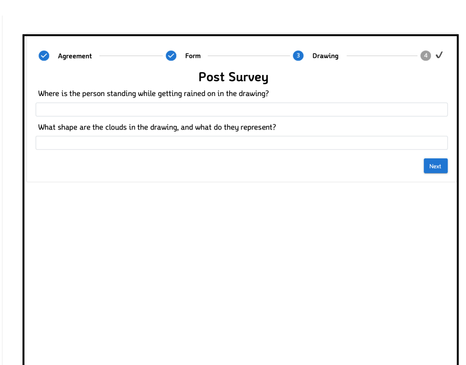
(b) Drawing Instruction



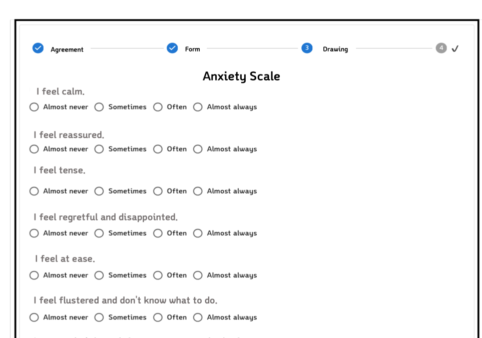
(c) Canvas



(d) Five PDIs: Pre-defined questions on the drawing



(e) Two PDIs: AI-generated questions on the drawing



(f) Anxiety Scale

Fig. 2. User interface of a drawing assessment system.

which are commonly used in traditional art therapy processes (see Figure 2(d)). Note that these questions are designed by clinical art therapists whose careers are at least 23 years. Then, users answer two AI-generated PDIs tailored to their drawings (see Figure 2(e)). After responding to the questions, users complete an age-appropriate anxiety scale; the scale for child&adolescents [74] and adults [85] (See Figure 2(f)). Upon completing the scale, users can review their drawings, drawing duration, and responses to the common and AI-customed PDIs.

In the drawing assessment system, a user-specific PDI generation module is designed and developed to ask two tailored questions about the user's drawing. Asking proper questions about the sketch is very important in drawing assessment as it can foster clients to provide more in-depth information about the context and their intention, which can help in more accurate evaluation [65]. Despite this importance, art therapists cannot provide personalized questions in large-scale settings. To address this gap, we developed a user-specific PDI generation module.

The overall process of the proposed module is illustrated in the 'User-specific PDI generation module' depicted in Figure 1. Here, we utilized the Multimodal Large Language Model (MLLM), specifically GPT-4V (`gpt-4-vision-preview`) provided by OpenAI. In the MLLM, we applied the following three prompting components: 'Instruction,' 'Sketch-question examples,' and 'Formatting.' First, 'Instruction' offers clear task guidelines to ensure the consistency and relevance of the generated questions, ensuring no overlap with predefined common PDIs. The used instruction is described as follows: *'You are an artificial intelligence assisting an art therapist. Your task is to generate questions for the Draw-A-Person-in-the-Rain assessment, similar to what an art therapist would ask.'* Second, 'Sketch-question examples' enable the module to learn from sample interactions consisting of image-PDI pairs. The images were sourced from the SceneDAPR dataset [43], which specifically targets psychological drawing for the DAPR assessment. For each image, three art therapists provided two unique PDIs, resulting in a total of six questions per image. Note that we recruited art therapists whose clinical experience is over nine years, ensuring that the generated questions align with professional standards. Lastly, through 'Formatting,' the generated questions follow a consistent format. For example, each question must include only one question mark, avoid references to design or color, and use simple language suitable for a young audience. Speculative expressions are excluded, and specific predefined questions—such as *"What is the person in the drawing doing?"* or *"Why is the person in the rain?"*—are omitted. Together, these components form a few-shot prompt utilizing four images with six PDIs for each image that enables the module to generate contextually appropriate and professionally relevant PDIs. Note that the prompts underwent an iterative refinement process using a human-in-the-loop approach to enhance the quality and accuracy of the final outputs. A total of 10 prompts were iteratively refined based on feedback from 6 art therapists, who assessed the quality of PDIs on a 1–5 point scale for five randomly selected drawings at each iteration. Additionally, the art therapists provided specific comments on areas needing improvement for each PDI. The initial evaluation resulted in a low average score of 1.63, with common suggestions for revisions focusing on user references and emotional expressions. As feedback was incorporated and subsequent iterations were conducted, the scores gradually increased, eventually exceeding an average of 4 points in the final evaluation. This human-in-the-loop process ensured that the PDIs generated by our module reached an acceptable standard, as detailed in Section 5.1.1.

3.2 Drawing Assessment Analysis System

To support decision-making for art therapists, a drawing assessment analysis system was developed alongside the drawing assessment system. As shown in Figures 3 and 4, the system is organized into three sequential sections following the natural workflow of DAPR assessment analysis: (i) client information for initial review, (ii) DAPR analysis results for detailed interpretation, and (iii) distribution of clients for contextualizing individual cases within the organization. This structure mirrors therapists' existing analysis process, minimizing navigation complexity while providing comprehensive information access.

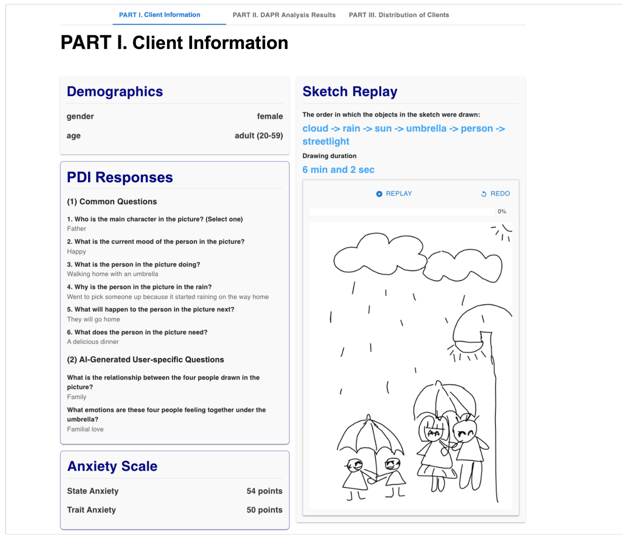


Fig. 3. User interface of client information in a drawing assessment analysis system, including demographics, PDI responses, anxiety scale, and sketch replay and drawing duration.

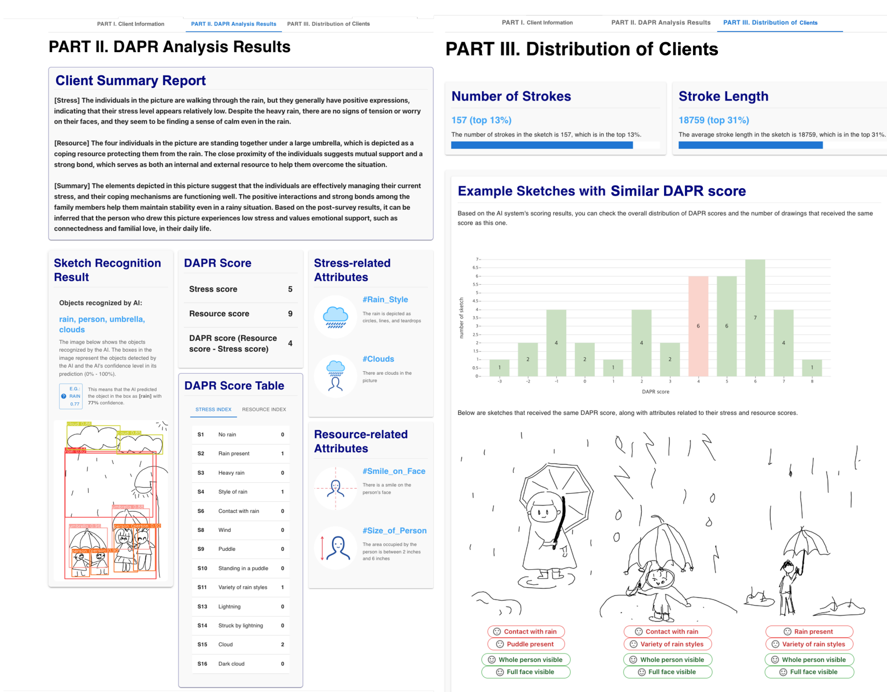


Fig. 4. Left: User interface of the DAPR analysis results in a drawing assessment analysis system, including the client summary report, sketch recognition results, DAPR score and score table, and stress- and resource-related attributes. Right: User interface showing the distribution of clients within the organization, including the number of strokes, stroke length, and example sketches with similar DAPR scores.

3.2.1 Client Information.

- **Demographic information**: Given the potential variability in art therapy processes and outcomes based on gender and age [44], the system displays demographic information, categorizing participants into child to adults (see Figure 3).
- **PDI responses**: As analyzing responses to PDIs is crucial in art therapy [5, 61], we present the client's responses to five pre-defined common PDIs and two user-specific AI-generated PDIs as explained in Section 3.1.
- **Anxiety scale**: The scores from the anxiety scale, which was obtained at the final stage of the drawing assessment system, are provided to explore potential correlations with DAPR scores [36, 97]. The adult anxiety was measured using the STAI-X scale [84], which is divided into state and trait anxiety. On the other hand, in the case of child & adolescent, the anxiety was measured based on RCMAS-2 [73], which is represented by a single score.
- **Sketch replay and drawing duration**: The ability to replay the sketch sequence is especially useful in remote settings where direct observation of participants is not possible, making it an essential tool for art therapists [46]. The system provides a video of the drawing sequence for review. Additionally, displaying the drawing duration can reduce the need for art therapists to manually measure the drawing process [46].

3.2.2 DAPR Analysis Results.

- **Client summary report**: During the art therapy process, reports are typically generated based on observations made during consultations on clients' drawings [82]. To streamline this process, we developed a summary report generation module, utilizing OpenAI's MLLM GPT-4V (`gpt-4-vision-preview`) to automatically generate reports based on clients' drawings and the response to PDIs. The module is guided by three key prompting methods: 'Instruction,' 'Client-report examples,' and 'Formatting.' 'Instruction' provides guidance on the task, ensuring consistency in report generation. For example, the prompt instructs: *'You are an artificial intelligence assisting an art therapist. Your task is to generate reports for the Draw-A-Person-in-the-Rain assessment, similar to those written by art therapists.'* Next, 'Client-report examples' help the module learn from client information-report pairs. The client information includes a client's drawing and his/her response to PDIs, and based on it, two art therapists wrote the report. These examples assist the module in creating contextually appropriate reports related to stress, resources, and summary sections for each image. Note that the clinical experience of two art therapists is over nine years, ensuring that the generated reports align with professional standards. Lastly, 'Formatting' ensures that the reports follow a structured format. For instance, the module is guided to structure reports into three sections: [Stress], [Resource], and [Summary]. It also follows guidelines such as avoiding speculative language, simplifying complex descriptions, and focusing on elements present in the drawing. This structured approach, combined with the few-shot learning method utilizing two images, their corresponding PDI responses, and reports, ensures that the module generates high-quality and reliable outputs. Similar to the human-in-the-loop approach used in the user-specific PDI generation module, the prompt for this module also underwent an iterative refinement process, as illustrated in Figure 1. An initial evaluation by six art therapists resulted in a score of 3.6, with feedback primarily recommending the removal of speculative or presumptive elements. Following four rounds of iterative refinement, the prompt's score improved to 4, leading to its adoption as the final version. This iterative process ensured that the final module produced accurate and professional reports based on the input sketches and associated PDIs.
- **Sketch recognition results**: A sketch object detection module is employed to help art therapists understand the AI analysis. The module is fine-tuned using YOLO-v8 [41] based

on the SceneDAPR dataset [43]. This module predicts and displays the elements recognized in six primary objects (person, rain, umbrella, cloud, puddle, and lightning) in participants' drawings, as depicted on the left side of Figure 4.

- **DAPR scores and score table**: DAPR scores are calculated by subtracting stress scores from resource scores [50]. Instead of merely presenting stress and resource scores, we demonstrate a detailed DAPR score table containing the specific index for stress and resource items and the final DAPR score. The DAPR scoring module leverages the capabilities of the MLLM Claude-3.5 (`claude-3-5-sonnet-20240620`) by Anthropic to enhance visual understanding and improve the accuracy of DAPR assessments with a two-stage strategy. In the first stage, a pre-trained object detection model, fine-tuned using the SceneDAPR dataset [43], is employed. This module utilizes YOLO-v8 [41] to identify six primary objects from the original sketches, as described in the *'Sketch recognition results'*. Then, MLLM refines these detection results by comparing the original sketch with the object-detected image. Three prompting methods—'Instruction', 'Multimodal Chain of Thought (COT)', and 'Refinement'—are applied during this process to guide the MLLM effectively. 'Instruction' provides explicit task guidelines to ensure consistent output, while 'Multimodal COT' facilitates step-by-step reasoning for more accurate analysis. Finally, 'Refinement' helps further improve detection results. Through this approach, a feedback report is generated, containing the detection results of six primary objects with detailed explanations when comparing original images and object-detected images. This report is utilized in the second stage, where the DAPR score is calculated. The module scores stress-related items (negative factors) and resource-related items (coping mechanisms) based on the DAPR assessment scale [51, 89] using two prompting methods: 'Dynamic prompt-based feedback' and 'Multimodal Chain of Thought (COT).' 'Dynamic prompt-based feedback' helps guide the module's evaluation by incorporating feedback from the previous stage, while 'Multimodal COT' enables a step-by-step reasoning process to ensure accurate scoring. The final DAPR score is determined by subtracting the stress score from the resource score. By refining object detection results and applying a structured approach to assessment, our module ensures accurate and reliable DAPR scores.
- **Stress- and resource-related attributes**: To facilitate the intuitive interpretation of sketches and DAPR scores, keywords representing stress-related and resource-related attributes are displayed, reflecting the expressions within the sketches [46].

3.2.3 *Distribution of Clients within the Organization.*

- **Numbers of Strokes and length**: Since line characteristics are crucial in the DAPR assessments, providing them can enhance experts' understanding [21]. The number of lines indicates how many lines the participant has drawn, and the length of the lines represents the calculated average of these lengths.
- **Example sketches with similar DAPR score**: As illustrated on the right side of Figure 4, the system shows the distribution of DAPR scores, allowing for the assessment of where a specific client falls within the overall group. Additionally, three sketches with similar DAPR scores are displayed along with relevant keywords, enhancing the understanding of interpretation consistencies across similar scores [46].

4 RESEARCH METHOD

4.1 Study Setup

To provide the drawing assessment analysis system with art therapists, collecting client data (i.e., drawings and PDI responses) should be preceded to be included in the system. Therefore, we first collected data using the drawing assessment system within *BetaDAPR*. We recruited 37 participants aged between 6 and 61 using a snowball sampling method. They used tablet PCs to fill out personal information, draw pictures according to instructions, answer PDIs, and respond to an anxiety scale. After completing all tasks, a brief interview was conducted to inquire whether they noticed any differences between the first five common PDIs and the subsequent two AI-generated user-specific PDIs. The drawing assessment experiment lasted approximately 15-20 minutes, and they received a compensation of approximately $4. The drawing assessment experiment was conducted with Institute Review Board approval.

Subsequently, the drawing assessment analysis system within *BetaDAPR* was configured using the collected dataset. Five images were selected from the 37 images to ensure diversity in client demographics and assessment outcomes, following a similar approach to prior work [46] that selected images based on quartile distributions. Specifically, we selected images representing varied age groups (14–55 years) and DAPR score ranges (-1 to 6 points) to allow participants to evaluate the system across different client profiles. The number of images was determined considering the experimental session duration; reviewing five images took approximately 20 minutes, which we considered reasonable for maintaining participant engagement and response quality. After configuring *BetaDAPR*, the experimental setup for the comparison system, *AlphaDAPR* [46], was conducted. For a fair comparison, all other conditions (user interface design, experimental methods, etc.) were kept identical, but only the quantitative elements presented in prior work [46] were included. Note that the DAPR scores in *AlphaDAPR* were generated using *BetaDAPR*'s scoring module to isolate the effect of qualitative assistance, ensuring that any observed differences reflect the presence or absence of qualitative support (i.e., PDI responses and reports) rather than technical differences. For more details, please see Appendix A.1.

4.2 Procedure

To examine the effectiveness of the proposed system, we recruited 60 art therapy experts through two channels: snowball sampling and online recruitment via website announcements. To compare the perceptions of *Juniors* and *Seniors*, 30 junior and 30 senior art therapists were recruited. Subsequently, each group was randomly assigned to one of the following two conditions, with half of each group allocated to each condition:

- *AlphaDAPR*: Participants in this group first received an image-based system notice, which includes a clear explanation of the roles of both AI and users (See Appendix A.2 for more details.) to make it easy to understand. It emphasized that *AlphaDAPR* supports quantitative analysis of large-scale groups, while art therapists complete the final analysis considering the analysis results provided by the AI system as supplemental information. After that, the participants examined *AlphaDAPR* results of the five sketches. Participants explored each analytical item (i.e., DAPR score and distribution of participants) on the system.
- *BetaDAPR*: Participants in this group also received a system notice based on an image-based explanation (See Appendix A.2 for more details.). It highlighted AI's quantitative and qualitative assistance and the importance of final decisions in art therapists. Then, the participants examined *BetaDAPR* results of the five sketches, evaluating each element (e.g., client summary report and DAPR score table).

Confirming that all three sketches had been examined, all participants completed the questionnaires. Of the 60 participants, 20 participated in interviews on their experience using the system. The entire experimental session lasted approximately 35 minutes per participant. This included reviewing the system notice (approximately 5 minutes), examining five sketches using the drawing assessment analysis system (approximately 20 minutes), completing questionnaires (approximately

Demographic Variables	Frequency	Percent (%)	Demographic Variables	Frequency	Percent (%)
Sample size	15	100.0	Sample size	15	100.0
AlphaDAPR-Junior			***BetaDAPR-Junior***		
Age			*Age*		
21-30 years	4	26.67	21-30 years	4	26.67
31-40 years	8	53.33	31-40 years	9	60.00
41-50 years	3	20.00	41-50 years	2	13.33
Education			*Education*		
Master Student	6	40.00	Master Student	6	40.00
Master Graduate	7	46.67	Master Graduate	6	40.00
PhD Student	2	13.33	PhD Student	3	20.00
Art Therapist Work Experience			*Art Therapist Work Experience*		
1-2 years	6	40.00	1-2 years	5	33.33
3-5 years	9	60.00	3-5 years	10	66.67
AlphaDAPR-Senior			***BetaDAPR-Senior***		
Age			*Age*		
21-30 years	1	6.67	21-30 years	1	6.67
31-40 years	7	46.67	31-40 years	6	40.00
41-50 years	6	40.00	41-50 years	6	40.00
51-60 years	1	6.67	51-60 years	2	13.33
Education			*Education*		
PhD Candidate	6	40.00	PhD Candidate	5	33.33
PhD Graduate	9	60.00	PhD Graduate	10	66.67
Art Therapist Work Experience			*Art Therapist Work Experience*		
5-10 years	9	60.00	5-10 years	9	60.00
11-20 years	5	33.33	11-20 years	5	33.33
21- years	1	6.67	21- years	1	6.67

Table 1. Demographic information of the participants.

10 minutes), and for the 20 interview participants, an additional semi-structured interview (approximately 30 minutes). All participants received about \$7.5 upon completion, and those also involved in the interviews received an extra \$4. Note that since *AlphaDAPR* does not have a drawing assessment system, this experiment was conducted using only the drawing assessment analysis system that art therapists would see. This study was approved by the Institute Review Board.

4.2.1 Participants. We recruited experts with art therapy degrees. A bachelor's degree was the minimum educational background required to qualify for participation in the experiment. As we defined art therapists with a PhD candidate or higher as *Senior* and those with less education as *Junior*, we finally recruited 30 participants for each group. Subsequently, both the groups were randomly divided into *AlphaDAPR* and *BetaDAPR* groups. Ultimately, we had four groups: *AlphaDAPR-Junior, AlphaDAPR-Senior, BetaDAPR-Junior, and BetaDAPR-Senior*, each with 15 art therapists participating in the experiment.

Table 1 summarizes the detailed information of the participants. We observed that the *AlphaDAPR* and *BetaDAPR* groups have an even distribution of *Junior* and *Senior* participants. As expected, the *Senior* group, with their higher education levels, exhibit higher ages and work experience. The average age of *Junior* in *AlphaDAPR* and *BetaDAPR* was 32.73 years (*SD*=4.93) and 32.33 years (*SD*=4.27), respectively, while the average age of *Senior* was 39.73 (*SD*=5.66) and 40.13 (*SD*=6.15), respectively, indicating a difference of about 7-8 years. Similarly, the average work experience of *Juniors* in *AlphaDAPR* and *BetaDAPR* was 2.93 years (*SD*=1.53) and 2.87 years (*SD*=1.46), respectively, while the average for *Senior* was 10.13 years (*SD*=5.03) and 10.27 years (*SD*=4.89), respectively, showing a difference of about 7 years.

4.2.2 Measures. We measure quantitative data with questionnaires to examine the effectiveness of using *BetaDAPR*. All questions were rated on a 5-point agreement scale ranging from strongly disagree to strongly agree. Similar to Guo et al. [33], two attention check questions like "Please select 'neutral' " were inserted among the user experience questions to ensure the quality of the participant's responses.

User-specific PDI			Client Summary Report		
	M	*SD*		*M*	*SD*
Helpfulness	4.21	0.82	Helpfulness	4.33	0.80
Fit	3.97	0.91	Fit	4.09	0.91
Fluency	4.02	0.92	Fluency	4.34	0.75
Level	3.93	0.93	Level	4.08	0.93

Table 2. Performance of expert verification for the user-specific PDI generation module and the summary report generation module.

We first measured *Understandability* using the three items in [62] (Cronbach's α = 0.724). *Explainability* was measured with two items reported in [92] (Cronbach's α = 0.811). *Perceived Ease of Use* and *Perceived Usefulness* were measured with two and five statements, respectively, developed by [27] (Cronbach's α = 0.857 and 0.882, respectively). *Satisfaction* was measured with three items from [96] (Cronbach's α = 0.866). *Trust* was measured with three items used in [68] (Cronbach's α = 0.847). *Technology Acceptance* was measured with two items from [93] (Cronbach's α = 0.853). *AI Familiarity* was measured with two items in [46] (Cronbach's α = 0.708). To measure the *Co-presence* of AI in the system, two items from [58] were utilized (Cronbach's α = 0.752). *Physical presence*, which means the extent to which art therapists felt they were genuinely engaged in a real DAPR analysis session while using the system, as opposed to simply reviewing information on a screen [14], was measured with two statements developed by [52] (Cronbach's α = 0.704). Finally, to confirm that the participants were willing to continue using the system, their *Intention to Use* was measured using two statements derived from [59] and [67] (Cronbach's α = 0.743). See Appendix A.3 for more detailed survey items.

5 RESULTS

5.1 Performance on *BetaDAPR*

Before examining art therapists' perceptions of the system, we first validated the results of the three proposed modules: (i) a user-specific PDI generation module, (ii) a summary report generation module, and (iii) a DAPR scoring module, incorporated into *BetaDAPR*.

5.1.1 User-specific PDI generation module. To investigate the performance of the proposed user-specific PDI generation module, we conducted separate experiments with 37 users who utilized our drawing assessment system and 16 art therapists with over 7 years of clinical experience. To the users, we specifically asked whether they noticed any differences between five common PDIs and AI-based user-specific PDIs as follows: "After drawing the sketch, you first answered five questions, and then two more questions appeared. Did you notice any differences between the first five questions and the latter two questions?". The majority, 64.86%, reported that they did not notice any difference between the PDIs. However, 21.62% of users mentioned that the user-specific PDIs felt more detailed, with some even noting that the questions appeared to consider their drawings and provided personalized insights. After completing the response, we informed the users that the latter two questions were generated by the AI based on their drawing. Then, 27.03% of participants commented that the user-specific PDIs felt natural, expressing surprise at the advancement of AI in generating such questions. In summary, no users expressed any sense of incongruity with the AI-generated PDIs, and most users considered them to be similar to the five commonly used art therapy PDIs. Furthermore, several users were surprised and expressed that they felt the questions were personalized for them. This implies that, from the user's perspective, the result of the user-specific PDI generation module was natural and contextually appropriate.

	Group A	Group B
Experts	0.99 [0.99, 1.]	0.99 [0.98, 0.99]
+ AlphaDAPR	0.75 [0.65, 0.83]	0.77 [0.67, 0.84]
+ BetaDAPR	0.93 [0.88 0.96]	0.90 [0.84 0.95]

Table 3. Intraclass correlation coefficient (ICC) in groups A and B. 'Expert' means three experts per group. '+ AlphaDAPR' and '+ BetaDAPR' means three experts with the scoring model/module result of *AlphaDAPR* and *BetaDAPR*, respectively.

Although the actual users evaluated the result of the user-specific PDI generation module positively, it is crucial to verify from an art therapy perspective whether the questions generated by our module are suitable for eliciting important information from clients. Therefore, we conducted a module validation experiment with art therapists. The experts were provided with 20 cases that obtained information about the client (age, gender), the drawing they created, and the PDIs generated by our module. They then responded to the following four items regarding the PDIs: "Would it be beneficial to have an AI that generates these types of questions and receives responses from the client?" (Helpfulness), "Are these questions well-suited to the drawings created by the client?" (Fit), "Do these questions read naturally and smoothly?" (Fluency), and "From the perspective of an art therapist, please evaluate the quality of these questions." (Level). The results are presented in the Table 2. Experts reported that the user-specific PDIs were helpful (M = 4.21, SD = 0.82), relevant (M = 3.97, SD = 0.91), fluent (M = 4.02, SD = 0.92), and demonstrated a decent level of quality (M = 3.93, SD = 0.93). This indicates that our user-specific PDI generation module, from the perspective of experts, demonstrates a performance that is more than capable of playing a role in obtaining key information from clients in large-scale settings.

5.1.2 Summary report generation module. To evaluate the performance of the proposed summary report generation module, we conducted a similar evaluation process as we did for the user-specific PDI generation module in terms of art therapists. Here, experts were provided with 10 cases, each including client information (age, gender), the client's drawing, the response to PDIs (five common PDIs and two user-specific PDIs), and the summary report generated by our module. They evaluated the summary report based on four criteria: "Would it be beneficial to have an AI generate drafts of reports like this?" (Helpfulness), "Considering the drawing and the responses to PDIs, is this report well-suited to the client?" (Fit), "Does this report read naturally and smoothly?" (Fluency), and "From an art therapist's perspective, please evaluate the quality of this report." (Level). As summarized in Table 2, the result showed that art therapists rated the client summary reports high, with all categories averaging a score of four or above: helpfulness (M = 4.33, SD = 0.80), relevance (M = 4.09, SD = 0.91), fluency (M = 4.34, SD = 0.75), and level (M = 4.08, SD = 0.93). This suggests that our module has generated reports that are satisfactory to experts, leading to the potential to significantly assist art therapists in report writing when conducting large-scale drawing assessments.

5.1.3 DAPR scoring module. To validate the performance of the quantitative assistant, the DAPR scoring module, we compared the DAPR scores obtained from the module with those calculated by human experts. For the comparison, we randomly selected 200 drawings from the test set of SceneDAPR dataset [43], and six clinical art therapists were required to examine the drawings. All the art therapists show at least six years of clinical experience and hold a Master's degree in clinical art therapy. The experts were divided into two groups, Groups A and B, based on their careers, resulting in the average clinical experience of groups A and B being 13 and 9.6 years, respectively. Then, 100 different sketches were assigned to each group. Table 3 demonstrates the inter-rater

Alpha-Junior (AJ)			***Beta-Junior (BJ)***		
	M	*SD*		*M*	*SD*
Understandability	4.29	0.49	Understandability	4.36	0.57
Explainability	3.93	0.50	Explainability	4.10	0.60
Perceived Ease of Use	4.17	0.49	Perceived Ease of Use	4.40	0.57
Perceived Usefulness	3.96	0.45	Perceived Usefulness	4.33	0.78
Satisfaction	3.69	0.39	Satisfaction	4.07	0.70
Trust	3.53	0.39	Trust	3.51	0.55
Technology Acceptance	3.30	0.70	Technology Acceptance	3.43	1.02
AI Familarity	3.33	0.43	AI Familarity	3.32	0.60
Co-Presence	2.73	0.46	Co-Presence	2.67	0.49
Physical Presence	3.10	0.66	Physical Presence	2.80	0.86
Intention to Use	4.07	0.75	Intention to Use	4.50	0.82
Alpha-Senior (AS)			***Beta-Senior (BS)***		
	M	*SD*		*M*	*SD*
Understandability	4.22	0.56	Understandability	4.51	0.52
Explainability	4.10	0.43	Explainability	4.33	0.56
Perceived Ease of Use	4.17	0.52	Perceived Ease of Use	4.53	0.48
Perceived Usefulness	4.03	0.52	Perceived Usefulness	4.40	0.52
Satisfaction	3.82	0.62	Satisfaction	4.16	0.68
Trust	3.56	0.48	Trust	4.02	0.57
Technology Acceptance	3.30	1.00	Technology Acceptance	3.22	0.56
AI Familarity	3.28	0.48	AI Familarity	3.27	0.47
Co-Presence	2.62	0.50	Co-Presence	2.76	0.51
Physical Presence	2.57	0.73	Physical Presence	3.20	0.68
Intention to Use	3.97	0.48	Intention to Use	4.40	0.63

Table 4. Mean and Standard Deviation of survey results in each group.

reliability of the DAPR scores within a group by calculating the intraclass correlation coefficient (ICC) with a 95% confidence interval. The ICC values for each expert group A and B were 0.99 [0.90, 1.] and 0.99 [0.98, 0.99], respectively, indicating a high agreement (≥0.90) among the experts [48]. We then compare the calculated score results obtained from the DAPR scoring modules with the experts. In terms of *AlphaDAPR*'s scoring model, the ICC values were 0.75 [0.65-0.83] and 0.77 [0.67-0.84] for group A and B, respectively, which suggests a good agreement (≥ 0.75) between the annotators [48]. On the other hand, when compared to our DAPR scoring module, while the ICC slightly decreases if it is incorporated, overall, the ICC values with Group A (0.93 [0.88, 0.96]) and B (0.90 [0.84, 0.95]) tend to show a high agreement (≥0.90) between the annotators [48]. This demonstrates a significant level of agreement between the human evaluation results and our module-based results, which highlights the potential usage of *BetaDAPR* in assessing DAPR drawings. Through the result, we could confirm that *BetaDAPR* demonstrates superior performance compared to *AlphaDAPR* in terms of quantitative support as well.

5.2 Survey Results

Table 4 summarizes the descriptive statistics for each experimental condition. Overall, *BetaDAPR* groups showed higher scores than *AlphaDAPR* groups across some measures. To examine the statistical significance of these differences and analyze the effects of two independent factors—qualitative support and experience level—we conducted a two-way analysis of variance (ANOVA).

(1) Overall results, as depicted in Figure 5, revealed several significant effects across different measurement categories. For 'Satisfaction,' the ANOVA showed a significant main effect of qualitative support between the *AlphaDAPR* and *BetaDAPR* systems ($F(1, 56) = 5.12, p < .05$). Participants reported higher satisfaction with the *BetaDAPR* system, highlighting the impact of qualitative assistance on user satisfaction. Regarding 'Perceived Ease of Use,' there was also a significant main effect of qualitative support ($F(1, 56) = 5.04, p < .05$), with the *BetaDAPR* being

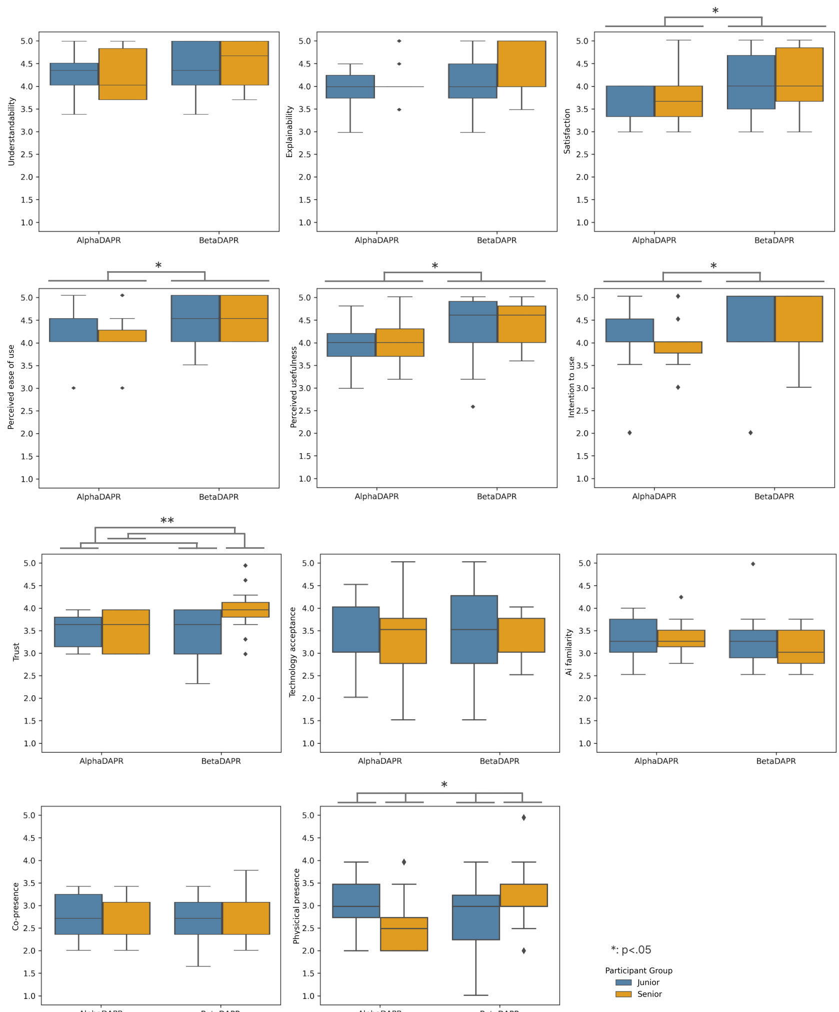


Fig. 5. Box plots of scores for measurement questionnaire items for *AlphaDAPR* and *BetaDAPR* according to participants' experience level (*Juniors* and *Seniors*).

perceived as easier to use compared to the *AlphaDAPR*. For 'Usefulness,' the absence and presence of qualitative support also showed a significant main effect ($F(1, 56) = 6.16, p < .05$), with participants rating the *BetaDAPR* as more useful than the *AlphaDAPR*. In terms of 'Intention to Use,' the results indicated a significant main effect of qualitative assistance ($F(1, 56) = 6.01, p < .05$), suggesting that participants were more inclined to use the *BetaDAPR* over the *AlphaDAPR*. For 'Trust,' a significant main effect was observed for the experience level ($F(1, 56) = 4.21, p < .05$), with *Seniors* expressing greater trust in the system compared to *Juniors*. Lastly, for 'Physical Presence,' there was a significant interaction between qualitative support and experience level ($F(2, 56) = 6.03, p < .05$). This interaction indicated that both factors significantly influenced the participants' perceived physical presence during the experiment. These findings underscore that qualitative support and the art therapist's experience level significantly affect the evaluation of expert support systems.

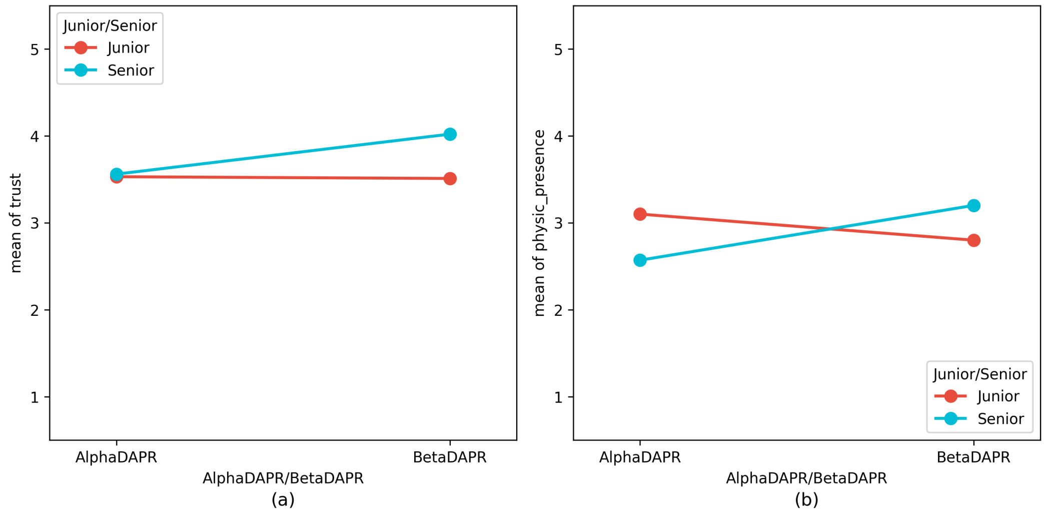


Fig. 6. Trust and physical presence trend according to two factors (qualitative support presence/absence and experience level).

Trust		df	F	p-value	Physical Presence		df	F	p-value
AlphaDAPR	Junior/Senior	1	0.02	0.891	AlphaDAPR	Junior/Senior	1	4.41	0.045*
	Residual	28				Residual	28		
BetaDAPR	Junior/Senior	1	6.28	0.018*	BetaDAPR	Junior/Senior	1	2.00	0.168
	Residual	28				Residual	28		
Junior	AlphaDAPR/BetaDAPR	1	0.02	0.899	Junior	AlphaDAPR/BetaDAPR	1	1.15	0.294
	Residual	28				Residual	28		
Senior	AlphaDAPR/BetaDAPR	1	5.86	0.022*	Senior	AlphaDAPR/BetaDAPR	1	6.09	0.020*
	Residual	28				Residual	28		

Table 5. Performance result of simple effect tests.

The presence of qualitative support notably impacted key elements such as satisfaction, ease of use, usefulness, and intention to use. The detailed ANOVA results are provided in Appendix B.1.

(2) Trust & Physical presence presented noteworthy results. While it is certainly significant that participants who used *BetaDAPR* were more satisfied, found it more useful, and expressed a higher intention to continue using it compared to *AlphaDAPR*, the high trust and physical presence scores from *Seniors* who used *BetaDAPR* were unexpected and novel findings. Therefore, we conducted further analysis to investigate this phenomenon. First, we examined the trends in trust and physical presence according to the two conditions (see Figure 6). In terms of trust (see Figure 6(a)), *Juniors* showed slightly lower trust in *BetaDAPR* compared to *AlphaDAPR*, while *Seniors* demonstrated higher trust in *BetaDAPR*. A similar trend was observed for physical presence (see Figure 6(b)), where *Juniors* felt less like they were in a DAPR analysis environment in *BetaDAPR*, but *Seniors* felt more so. To analyze this in more detail, we performed a simple effects test, and the results are shown in Table 5. In terms of trust, there was a significant difference in values between *Junior* and *Senior* participants within the *BetaDAPR* group ($F(1, 28) = 6.28, p < .05$). Conversely, for physical presence, a significant difference was observed between *Junior* and *Senior* participants within the *AlphaDAPR* group ($F(1, 28) = 4.41, p < .05$). When considering the *Junior/Senior* factor, *AlphaDAPR* and *BetaDAPR* groups showed significant value differences when participants were *Seniors* in both trust ($F(1, 28) = 5.86, p < .05$) and physical presence ($F(1, 28) = 6.09, p < .05$). Based

System Components	% of Participants (n = 15)			
	Alpha-Junior	Alpha-Senior	Beta-Junior	Beta-Senior
Demographics	60.00	66.67	33.33	80.00
PDI Responses	-	-	73.33	86.67
Anxiety Scale	-	-	40.00	60.00
Sketch Replay and Drawing Duration	80.00	93.33	100.00	86.67
Sketch Recognition Results	53.33	60.00	46.67	66.67
DAPR Score	60.00	46.67	33.33	80.00
DAPR Score Table	-	-	66.67	60.00
Client Summary Report	-	-	40.00	80.00
Stress-related Attributes	46.67	46.67	20.00	60.00
Resource-related Attributes	53.33	40.00	46.67	60.00
Number of Strokes and Length	60.00	60.00	33.33	33.33
Example Sketches with Similar DAPR Score	60.00	53.33	26.67	26.67

Table 6. The components of the system that the participants recognized as useful (multiple choice).

Theme	Sub-theme	% of Participants				
		All (n = 20)	Alpha-Junior (AJ) (n = 7)	Alpha-Senior (AS) (n = 5)	Beta-Junior (BJ) (n = 4)	Beta-Senior (BS) (n = 4)
Quantitative Analysis	Helpfulness	100.00	100.00	100.00	100.00	100.00
Qualitative Analysis	Prospective Benefits	45.00	71.43	80.00	-	-
	Helpfulness	35.00	-	-	100.00	75.00
AI Presence	Human Assistant	25.00	28.57	00.00	50.00	0.00
	Assistive Tool	65.00	71.43	100.00	25.00	75.00
	Both	10.00	0.00	0.00	25.00	25.00
Trust	Trustworthiness	10.00	0.00	0.00	0.00	50.00
Concerns Regarding Adoption	Authority of Art Therapists	15.00	42.86	0.00	0.00	0.00
	Over-reliance	20.00	0.00	80.00	0.00	0.00
	Privacy Issues	20.00	0.00	0.00	50.00	50.00
Intention to Use	Future Usage	85.00	85.71	60.00	100.00	100.00

Table 7. Themes and sub-themes from the participants' interview data representing the experience of using the system (*AlphaDAPR* or *BetaDAPR*).

on the result, we finally examine the mediation effect between physical presence and trust through a Sobel test [83]. In the *Senior* group, the Sobel test revealed a significant indirect effect of the qualitative support (*AlphaDAPR/BetaDAPR*) on physical presence through trust ($Z = 2.13$, $p < .05$). This suggests that *BetaDAPR*, which provides qualitative information in addition to quantitative data, enhances users' trust, making them feel as if they are analyzing DAPR in a real-world setting.

5.3 Interview Results

In this section, we investigate how participants experience and perceive the system (*AlphaDAPR* or *BetaDAPR*). To this end, we first focused on how the participants evaluated the effectiveness of each component in the drawing assessment analysis system. Therefore, we asked the extent to which the participants perceived each system component useful. The result is demonstrated in Table 6. All groups of art therapists felt that 'sketch replay and drawing duration' were helpful, which aligns with the previous studies [46]. Examining the components exclusively included in *BetaDAPR*, we observed that both junior and senior therapists found them useful, with a notably higher preference among *Seniors*. The components aimed at assisting qualitative analysis, such as 'PDI responses' and 'client summary report,' garnered an overwhelming approval rate of over 80% (86.67% and 80%, respectively). This might explain why participants who used *BetaDAPR* did not find the 'number of strokes and length' and 'example sketches with similar DAPR score' as relatively useful, which were highly rated by *AlphaDAPR* participants.

Then, we conducted semi-structured interviews with participants in each group to explore how they experience and perceive the system. The interview questions were organized based on participants' rates on the questionnaire results. The questions consist of (i) participants' prior experience with AI, (ii) which parts of the system the participants liked and disliked, (iii) how they felt about AI-based functions with quantitative/qualitative assistance, and (iv) potential for real-world application and concerns regarding adoption. Note that we replaced the third question about qualitative assistance with "how they feel if AI-based functions with qualitative assistance were added" for *AlphaDAPR* participants. All interview sessions are recorded. Two authors independently coded interviews that were recorded and transcribed. Then, the authors organized the frequently observed patterns and themes with thematic analysis [17], as summarized in Table 7. Detailed findings from each interview are presented in the six subsections below.

5.3.1 Quantitative Support. The system (*AlphaDAPR* and *BetaDAPR*) provided quantitative analysis such as sketch replay, drawing duration, DAPR score, and example sketches with similar DAPR scores. All participants stated that they found the quantitative support was useful. One participant remarked, "What I appreciated most was the system's ability to quickly interpret the art assessments and provide detailed insights, including the scores and attributes within the analysis." (AJ2). Another participant noted, "The best part was showing the process of drawing. I mean, when we were checking the DAPR scores, subjective concepts kept getting involved. But the fact that it came out clearly and quantified as a score was really helpful." (AS2). Similarly, BJ2 stated, "The order of the drawings was helpful. Also, when it suggested drawings with similar DAPR scores, the understanding was much quicker." Furthermore, another participant mentioned that as more data accumulates, AI might outperform humans in quantitative analysis: "When it comes to objective evaluation, if more data is collected and averaged, I think AI might even perform better than humans in this regard." (BS1). Through these statements, we observed that art therapists perceived the quantitative analysis assistance of the system positively regardless of the two conditions (*AlphaDAPR/BetaDAPR* and *Junior/Senior*).

However, *AlphaDAPR* users also noted limitations of quantitative-only support. AJ7 remarked, "The numerical analysis was convenient, but detailed aspects like facial expressions or whether the person was drawn from behind were missing. I felt it was a bit lacking that only numerical things were analyzed." This feedback suggests that while quantitative support provides valuable efficiency gains, therapists perceived a gap between numerical outputs and the interpretive depth required for comprehensive assessment.

5.3.2 Qualitative Support. On the contrary to *AlphaDAPR*, *BetaDAPR* also assists qualitative aspects, including user-specific PDI generation and summary report generation. Of the participants who used *BetaDAPR*, 87.5% were satisfied with their experience where the AI-generated qualitative analysis is a helpful support tool. One junior art therapist commented on the client summary report, "It was good. It covers the basic details, which can be referenced, making it very helpful." (BJ2). A senior art therapist noted the usefulness of PDI responses: "I thought that was great. When working one-on-one, we can ask follow-up questions based on the drawing process, but in large-scale settings, those details were missed." (BS4). Additionally, 75% of participants who used *AlphaDAPR* expressed interest in having AI-generated qualitative analysis and believed it would be beneficial. When asked whether qualitative analysis would be helpful, one participant noted, "I think it would be good. If AI could ask certain questions to those who drew specific elements, it would add more meaning." (AS1). Taken together, these responses indicate that qualitative support is valued not merely as an efficiency tool, but as a means of preserving the depth of clinical insight that large-scale settings would otherwise compromise. However, AJ1 cautioned that it should be used selectively: “It might need to be applied to more limited aspects. PDIs are often based on the client’s attitude, which is observed during the process, not just from the drawings.”

5.3.3 AI Presence. Participants were asked whether the AI in the system felt more like a human assistant or simply an assistive tool. 88.89% of *Seniors* stated that the system felt like an assistive tool rather than a human assistant. BS2 remarked, “It felt more like receiving help from an automated analysis program rather than from someone assisting me directly.” Similarly, AS5 described the system as “not an assistant therapist, but more like a recording machine,” emphasizing that its main function for them was to document the drawings. However, *Juniors* had mixed opinions. Among them, 36.36% perceived the system as a human assistant, 54.55% viewed it as an assistive tool, and 9.09% saw it as both. BJ2 mentioned that “It felt like an assistant therapist. Since even human therapists are not perfect, AI seems like an imperfect human in that sense.” Another participant, while acknowledging that it felt more like an assistive tool, had a nuanced perspective: “It is somewhat ambiguous. In group settings, it plays the role of an assistant therapist by providing information that a human therapist might miss. But intuitively, I still perceive it as just a tool.” (AJ3). Likewise, a distinct difference in AI presence perception was observed between *Juniors* and *Seniors*. In the *Senior* group, all except one participant viewed AI within the system as an assistive tool. Conversely, a considerable number of *Juniors* perceived AI as a co-therapist. This discrepancy appears to be influenced by their prior experience with AI. When asked about prior experience with AI, only 18.18% of *Seniors* reported frequent use of recent tools like ChatGPT, whereas 54.55% of *Juniors* did. This suggests that prior exposure to and utilization of AI significantly impacts how individuals perceive its presence, either as more proximate or distant.

5.3.4 Trust. Although no specific questions about trustworthiness were asked, half of the senior art therapists who used *BetaDAPR* mentioned that the detailed information provided by the analysis results made the system feel more trustworthy. BS2 commented, “Having features like the DAPR score table increases my confidence when reviewing the results. Seeing how the interpretations are reflected in the scores makes the system feel more reliable.” Similarly, BS4 stated, “The numerical aspect gives it a bit more credibility, making the results feel more grounded. It provided a basis for areas that might otherwise feel ambiguous, which made me feel more at ease.” Overall, the detailed DAPR scores provided by *BetaDAPR* were perceived as helpful by senior art therapists. This observation aligns with the high level of trust exhibited by senior therapists who used *BetaDAPR*, as shown in Table 4. On the other hand, some junior art therapists expressed distrust in the system due to prior experience with ChatGPT: “I don’t fully trust GPT when I use it. I only use it as a reference when completing tasks. I see the same with the drawing assessment analysis system — it’s something that could be a useful reference for my work, but I don’t trust it 100%. It would help me make judgments, but not be something I rely on entirely.” (AJ4). These patterns reveal that trust in AI is not unidimensional: seniors build it through transparency and verifiability, while juniors’ extensive AI experience has paradoxically cultivated a more calibrated skepticism.

5.3.5 Concerns Regarding Adoption. When asked about concerns related to adopting the drawing assessment analysis system, three key themes emerged, varying by participant group. Junior art therapists who used *AlphaDAPR* expressed concerns about the authority of art therapists in the process. AJ2 mentioned, “This is something that belongs to the domain of experts. It involves diagnosing individuals in a safe space while considering their gestures, pen pressure, and other factors, none of which can be accounted for with AI. I’m concerned that, if widely adopted, AI could be misused in ways that undermine this expertise.” In contrast, most senior art therapists who used *AlphaDAPR* answered that junior therapists might rely too heavily on the system: “For more experienced therapists, this could serve as a useful assistant. However, I worry that for junior therapists, it might become so ingrained that they won’t consider alternative approaches.” (AS2).

On a different note, participants who used *BetaDAPR* raised privacy concerns. BJ1 reported, "I was curious about how privacy is handled. Is the data stored on the company's database or do I have control over it and can delete it afterward? Even if personal information is removed, there is still the concern of data leaks, which makes me uneasy." These findings highlight that concerns regarding adoption are vary systematically depending on both experience level and the type of AI support provided.

5.3.6 Intention to Use. Most participants expressed an intention to use the system, with some noting that it could be particularly effective in large group settings. However, the reasons for adoption differed between *AlphaDAPR* and *BetaDAPR* users.

AlphaDAPR users primarily valued the system's efficiency in automating time-consuming manual tasks. AJ1 stated, "I thought it would be incredibly useful in situations where quick results are needed within the constraints of a limited budget, especially in large group environments." AS4 noted, "Especially for research, teaching, or corporate lectures where I have 20 to 30 clients, this would be really helpful. Measuring sizes and calculating inches individually can be quite tedious, so if this tool could take over those tasks, it would be great."

BetaDAPR users, in contrast, emphasized the system's comprehensive support that transformed their professional workflow. Additionally, one participant mentioned the potential benefit in managing heavy workloads: "If this system becomes available, I plan to actively use it. Writing reports while working at a private institution is quite demanding, so I believe I would use it." (BJ1). BS2 highlighted how the system shifted her role from manual documentation to quality verification: "My physical time isn't particularly consumed because I'm just checking whether the assessment's objectivity is appropriate, so I think I would use it more frequently." BS3 added, "Looking at this, I thought it could actually be used in a hospital setting—the report draft, score table, everything was good," suggesting that comprehensive support met the standards required for clinical deployment.

These responses indicate that while both systems were perceived as useful for large-scale assessments, *BetaDAPR*'s qualitative support addressed a specific pain point in art therapy practice—the burden of report writing and documentation—thereby transforming the therapist's role from performing manual tasks to verifying and refining AI-generated outputs.

6 DISCUSSION

6.1 Impact of Qualitative Support and Professional Experience

In this paper, we conducted an experiment with two conditions: the presence or absence of qualitative support in the system and the professional experience level of the art therapist.

6.1.1 Effect of Qualitative Support. The qualitative support—PDI responses and client summary reports—addressed the two most labor-intensive and expertise-dependent aspects of large-scale drawing assessment: obtaining detailed client information without face-to-face inquiry, and generating comprehensive documentation for each participant. This workflow integration produced a counterintuitive result: *BetaDAPR* scored higher on perceived ease of use despite presenting more information than *AlphaDAPR*, challenging conventional assumptions that more features necessarily increase complexity [64]. Since both systems share the same user interface and experimental procedure, the difference can be attributed to the qualitative support itself. In art therapy practice, practitioners must juggle multiple analytical tasks simultaneously—observing drawings, recalling PDI responses, formulating interpretations, and drafting reports. Rather than adding to this cognitive burden, *BetaDAPR* consolidated these fragmented processes into a unified workflow, thereby reducing the articulation work [13, 79] required to coordinate parallel tasks. This suggests that for expert support systems in domains requiring both quantitative measurement and qualitative interpretation, perceived ease of use depends not on feature minimalism but on workflow completeness [98]: systems feel easier when they support the full scope of professional practice.

6.1.2 Effect of Professional Experience. Regarding art therapists' experience level, our findings diverge from prior research [46, 53], which documented significant job displacement anxiety among junior professionals when AI systems handled tasks once considered core to human expertise. Rather than viewing AI as encroaching on professional territory, junior participants in our study largely conceptualized it as expanding their capacity to manage articulation work [79] in large-scale assessments. This shift appears attributable to the recent proliferation of accessible AI tools such as ChatGPT and Gemini, which has normalized AI as a collaborative partner in everyday workflows. Consequently, future expert support systems may benefit from treating prior AI experience as a key design variable—for instance, by calibrating onboarding, explanation, and transparency features based on users' familiarity with AI tools.

6.1.3 Interaction Between Qualitative Support and Experience Level. The interaction between qualitative support and experience level produced a particularly notable pattern among senior art therapists using *BetaDAPR*, who reported heightened trust and a stronger sense of physical presence. Mediation analysis confirmed that qualitative support enhanced trust, which in turn led to this heightened sense of physical presence. This mediation effect can be understood through the lens of cognitive absorption theory [4, 11]: when users trust a system, they allocate fewer cognitive resources to verifying its outputs, freeing attentional capacity for the primary task [2].

In our context, senior therapists who trusted *BetaDAPR*'s comprehensive analysis could focus entirely on clinical interpretation rather than second-guessing the system's scoring or report generation. This undivided attention facilitated deeper engagement with the assessment process, manifesting as a heightened sense of physical presence—the feeling of being genuinely situated in a DAPR analysis environment rather than merely viewing information on a screen. This trust was grounded in the system's transparency, where features such as the DAPR score table and item-level breakdowns provided a verifiable basis for interpretations, reducing the need for independent verification. This finding suggests that for expert support systems, building trust is not merely about user satisfaction—it serves as a gateway to deeper task engagement. Future research could explore how different transparency mechanisms (e.g., confidence scores, reasoning traces, interactive explanations) differentially affect trust formation and subsequent task immersion across varying expertise levels.

In contrast, junior participants exhibited a paradoxical pattern: despite greater AI familiarity, they reported lower trust in the system. Rather than signaling rejection, this reflects a sophisticated awareness of current AI limitations cultivated through extensive prior experience with tools such as ChatGPT. Rather than treating AI outputs as authoritative, juniors tended to approach them as useful references requiring professional verification. As AI usage becomes increasingly prevalent, future research should consider how to design expert support systems that accommodate these divergent trust patterns shaped by users' prior AI experience.

6.2 Practical Adoption

The interview results revealed that the potential concerns regarding the actual application are diverged into three groups: *Juniors* with *AlphaDAPR*, *Seniors* with *AlphaDAPR*, and *BetaDAPR* users. Focusing on the *AlphaDAPR* users, their perspectives revealed a central paradox surrounding automation in expert domains. *Juniors* emphasized that the system should be restricted to experts as intended to prevent misuse by the general public, while *Seniors* worried that junior therapists themselves might over-rely on the AI, blindly trusting its scores without sufficient clinical experience. These opposing views are not contradictory; rather, they represent two facets of the same core anxiety: ensuring technology enhances, rather than erodes, professional expertise. The seniors' concern reflects a desire to preserve the integrity of their field, positioning AI as an assistive tool where the human expert remains the ultimate authority and guardian of professional standards [28, 90]. Meanwhile, the juniors' stance points to a more personal fear of skill devaluation. For emerging professionals, an AI that automates subjective analysis can feel less like a supportive tool and more like a direct threat to the very competencies they are trying to build [46, 53]. Their concern is not just about public misuse, but about the potential erosion of the professional value they hope to attain.

On the other hand, participants who used *BetaDAPR* expressed concerns about privacy issues. In art therapy, where client confidentiality is paramount, the primary concern was about where this data would be stored and whether there was any possibility of it being leaked externally. Participants particularly emphasized the need for a system that adheres to the stringent standards for handling patient information in hospitals. This feedback directly reflects the principle of an unobtrusive fit into existing procedures, a concept emphasized in CSCW [6, 98]. In a high-stakes therapeutic environment, data security is not merely a technical feature but a fundamental prerequisite of trust and acceptance as part of the workflow. Therefore, integrating tools like *BetaDAPR* requires the system to act as a trustworthy partner that respects existing collaborative and administrative structures [13]. This includes aligning with established data management protocols—for instance, by storing user information within the secure databases of the institutions (i.e., schools and companies) that request large-scale drawing assessments.

The contrast between these concern patterns carries important implications for practical adoption. When *AlphaDAPR* provided quantitative scores without interpretive context, therapists were left uncertain about how to integrate AI output into their clinical workflow—resulting in anxieties about professional boundaries. *BetaDAPR* users, however, raised no such concerns, focusing instead entirely on privacy. This shift indicates that when therapists could see not only what the AI concluded but also how it reached those conclusions in real workflow—through PDI responses, detailed score breakdowns, and draft reports—questions about professional boundaries gave way to practical implementation concerns. This pattern suggests that the nature of adoption barriers depends on the completeness of AI support. Partial automation, such as providing only numerical scores, creates a gap that therapists must fill with their own interpretation, generating uncertainty about role boundaries. Comprehensive support, by contrast, offers outputs that therapists can directly verify, refine, and take ownership of—transforming the AI from an opaque evaluator into a transparent collaborator. For system designers targeting expert domains, this finding suggests that resistance to AI adoption may be mitigated not by limiting AI capabilities, but by ensuring that AI assistance spans the full workflow while keeping its reasoning process visible and editable.

Despite the above concerns, nearly all participants expressed their willingness to use our system in their professional settings for future large-scale situations in the interview. Notably, all participants who used *BetaDAPR*, regardless of the experience level, indicated their intention for practical adoption. This suggests that they recognized the system's potential to alleviate the 'invisible labor' common in therapeutic settings [60], perceiving it as a valuable tool for streamlining the heavy workload of drawing assessment analysis in large-scale environments. This positive reception is rooted in our deliberate choice to conceptualize the system not as an autonomous decision-maker, but as a workflow-aware tool that supports the complex coordination work between therapists, clients, and institutions [45, 70, 87]. Building on this principle, future iterations should continue to focus on features that build trust through transparency, facilitate collaboration, and delineate clear roles for AI and human experts, ensuring the system empowers, rather than replaces, the professionals it is designed to serve.

6.3 Limitations & Future Work

This study has several limitations. First, the sample size was small. Since we divided participants into four groups based on two factors (*AlphaDAPR/BetaDAPR* and *Junior/Senior*), only 15 individuals were assigned to each group. To address this, we plan to conduct future experiments with a larger group of art therapists. Second, participants were recruited through snowball sampling from a specific therapist, which may not represent the broader population of art therapists. Future research should aim for a more diverse participant pool to improve generalizability, including considerations of cultural diversity. Third, to construct the drawing assessment analysis system, we collected data from 37 individuals using tablet PCs. While other information was easily gathered, pen pressure could not be represented. Future work will focus on incorporating pen pressure features into the drawing assessment system. Lastly, although our system was designed for large-scale scenarios, it was only tested in a lab setting. Future research should evaluate the system's effectiveness in real-world environments to better understand its practical applications.

6.4 Ethical Considerations

The design and evaluation of *BetaDAPR* are guided by a core ethical principle: it is intended as an expert support system, not a replacement for professional judgment. This principle informs our approach to the critical ethical challenges inherent in developing AI for art therapy, primarily revolving around data privacy and model reliability.

First, safeguarding data privacy is paramount, as the system handles sensitive sketch data that may reflect an individual's psychological state. Transmitting such data via external APIs introduces potential privacy risks. To address this, we obtained IRB-approved informed consent from all participants, ensuring they understood how their data would be used. Furthermore, all sketches and prompts were fully anonymized to prevent personal identification. For future deployment, we concur with our study participants' suggestions to process user data within institution-managed secure environments (e.g., schools, companies) or through on-premises inference, thereby minimizing the risk of external data leaks.

Second, ensuring module reliability required a multi-faceted approach, balancing cutting-edge capabilities with a clear-eyed view of their limitations. Our model selection was a deliberate division of labor based on an internal evaluation of leading MLLMs (e.g., GPT-4V [3], Claude 3.5 Sonnet [8], Gemini 1.5 Pro [72]). We chose OpenAI's GPT-4V for its nuanced qualitative interpretation in generating user-specific PDI and summary reports [57, 78], and Anthropic's Claude 3.5 for its high consistency in structured, quantitative DAPR score calculations [7]. To enhance the accuracy and explainability of these models' outputs, we employed reasoning techniques like the 'Multimodal Chain of Thought (COT)' [101].

Despite these measures, we must address the intrinsic limitations of current MLLMs. The primary concerns are 'hallucination,' the generation of plausible but factually incorrect content [10]), and the potential for cultural and contextual bias, where a model may misinterpret the rich personal or cultural symbolism in a sketch [16, 19]. Such inaccuracies could lead to flawed DAPR scores or misleading reports. A further challenge relates to reproducibility; our reliance on commercial, closed-source APIs means that vendor-driven updates can cause 'version drift,' where the same prompt yields different results over time [22, 42]. The black-box nature of these models makes perfect replication difficult, underscoring the need for future research with on-premise or open-weight alternatives.

These considerations—from data privacy to model fallibility—reinforce our core principle that *BetaDAPR* must function as a tool to augment, not automate, clinical judgment. To mitigate the aforementioned risks, the system is designed to promote transparency and critical oversight. It links all analytical outputs to tangible evidence, such as sketch recognition results and the DAPR score table, enabling therapists to trace and verify the basis of the AI's conclusions. The user interface and introductory instructions explicitly prompt clinicians to critically evaluate the AI's suggestions, ensuring the final clinical decision always rests with the human expert. We believe that upholding these ethical standards is crucial for the responsible development of AI-powered support systems in art therapy.

7 CONCLUSION

The goal of this study was to develop an AI-based expert support system capable of providing both qualitative and quantitative assistance, and evaluate the effect of qualitative support and experts' experience level on the system. To the best of our knowledge, this is the first attempt to propose an expert support system that supports both quantitative and qualitative aspects for art therapists and demonstrates its effectiveness from various perspectives. Our proposed system, *BetaDAPR*, elicited significantly more positive responses in key aspects (e.g., intention to use, usefulness, and satisfaction) compared to existing systems that only provided quantitative support. Notably, the high levels of trust and physical presence exhibited by *Seniors* suggest that the enhanced quantitative support and the novel qualitative support fostered a sense of trust, which, in turn, eliminated distractions and increased physical presence. In terms of experience level, the perception of AI, rather than the counselor's experience, was found to influence system usage. *Juniors*, who had more frequent exposure to the latest AI, tended to perceive the AI in our system as an assistant therapist and expressed a higher intention to use it. However, their trust in the system was relatively lower, indicating that their regular use of AI had ingrained the notion that AI cannot be completely trusted. We believe that these findings will be valuable in developing systems to support experts in art therapy.

ACKNOWLEDGMENTS

This research was supported by the MSIT (Ministry of Science, ICT), Korea, under the Global Scholars Invitation Program (RS-2024-00459638) and the Global Research Support Program in the Digital Field program (RS-2024-00425354) supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation). This work was also supported by NAVER AI Lab of NAVER Corporation.

REFERENCES

[1] Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y Lim, and Mohan Kankanhalli. 2018. Trends and trajectories for explainable, accountable and intelligible systems: An hci research agenda. In *Proceedings of the 2018 CHI conference on human factors in computing systems*. 1–18.

[2] Nirmal Acharya, Anne-Marie Sassenberg, and Jeffrey Soar. 2022. Effects of cognitive absorption on continuous use intention of AI-driven recommender systems in e-commerce. *foresight* 25, 2 (2022), 194–208.

[3] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).

[4] Ritu Agarwal and Elena Karahanna. 2000. Time flies when you're having fun: Cognitive absorption and beliefs about information technology usage. *MIS quarterly* (2000), 665–694.

[5] Lewis R Aiken. 1999. *Personality assessment methods and practice, 3rd rev.* Hogrefe & Huber Publishers.

[6] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, et al. 2019. Guidelines for human-AI interaction. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–13.

[7] Khizar Anjum, Muhammad Arbab Arshad, Kadhim Hayawi, Efstathios Polyzos, Asadullah Tariq, Mohamed Adel Serhani, Laiba Batool, Brady Lund, Nishith Reddy Mannuru, Ravi Varma Kumar Bevara, et al. 2025. Domain Specific Benchmarks for Evaluating Multimodal Large Language Models. *arXiv preprint arXiv:2506.12958* (2025).

[8] AI Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku. *Claude-3 Model Card* 1 (2024).

[9] Rita B Ardito and Daniela Rabellino. 2011. Therapeutic alliance and outcome of psychotherapy: historical excursus, measurements, and prospects for research. *Frontiers in psychology* 2 (2011), 270.

[10] Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. 2024. Hallucination of multimodal large language models: A survey. *arXiv preprint arXiv:2404.18930* (2024).

[11] Janarthanan Balakrishnan and Yogesh K Dwivedi. 2021. Role of cognitive absorption in building user trust and experience. *Psychology & Marketing* 38, 4 (2021), 643–668.

[12] Matthias Baldauf, Peter Fröehlich, and Rainer Endl. 2020. Trust me, I'ma doctor–user perceptions of AI-driven apps for mobile health diagnosis. In *Proceedings of the 19th International Conference on Mobile and Ubiquitous Multimedia*. 167–178.

[13] Jakob E Bardram and Steven Houben. 2018. Collaborative affordances of medical records. *Computer Supported Cooperative Work (CSCW)* 27, 1 (2018), 1–36.

[14] Woodrow Barfield and Suzanne Weghorst. 1993. The sense of presence within virtual environments: A conceptual framework. *Advances in Human Factors Ergonomics* 19 (1993), 699–699.

[15] Emma Beede, Elizabeth Baylor, Fred Hersch, Anna Iurchenko, Lauren Wilcox, Paisan Ruamviboonsuk, and Laura M Vardoulakis. 2020. A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–12.

[16] Abeba Birhane, Sepehr Dehdashtian, Vinay Prabhu, and Vishnu Boddeti. 2024. The dark side of dataset scaling: Evaluating racial classification in multimodal models. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*. 1229–1244.

[17] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101.

[18] Christoph Buck, Eileen Doctor, Jasmin Hennrich, Jan Jöhnk, and Torsten Eymann. 2022. General practitioners' attitudes toward artificial intelligence–enabled systems: interview study. *Journal of medical Internet research* 24, 1 (2022), e28916.

[19] Olena Burda-Lassen, Aman Chadha, Shashank Goswami, and Vinija Jain. 2025. How Culturally Aware Are Vision-Language Models?. In *2025 IEEE 6th International Conference on Image Processing, Applications and Systems (IPAS)*. IEEE, 1–6.

[20] Adrian Bussone, Simone Stumpf, and Dympna O'Sullivan. 2015. The role of explanations on trust and reliance in clinical decision support systems. In *2015 international conference on healthcare informatics*. IEEE, 160–169.

[21] Suellen M Carney. 1992. *Draw a person in the rain: A comparison of levels of stress and depression among adolescents*. Pace University.

[22] Lingjiao Chen, Matei Zaharia, and James Zou. 2024. How is ChatGPT's behavior changing over time? *Harvard Data Science Review* 6, 2 (2024).

[23] Shuqing Chen, Daniel Stromer, Harb Alnasser Alabdalrahim, Stefan Schwab, Markus Weih, and Andreas Maier. 2020. Automatic dementia screening and scoring by applying deep learning on clock-drawing tests. *Scientific Reports* 10, 1 (2020), 1–11.

[24] Zhuoyi Cheng, Pei Chen, Yiwen Ren, Wenzheng Song, and Lingyun Sun. 2025. Transitioning Focus: Viewing Human-AI Collaboration as Mixed-focus Collaboration. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–8.

[25] Seung-Je Cho, Geon-Woo Cho, and Young-wook Kim. 2021. Development of Deep Learning-Based House-Tree-Person Test Analysis Model. In *Proceedings of the Korea Information Processing Society Conference*. Korea Information Processing Society, 558–561.

[26] Andy Crabtree, Jacki O'Neill, Peter Tolmie, Stefania Castellani, Tommaso Colombino, and Antonietta Grasso. 2006. The practical indispensability of articulation work to immediate and remote help-giving. In *Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work*. 219–228.

[27] Fred D Davis. 1989. Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS quarterly* (1989), 319–340.

[28] Jojanneke Drogt, Megan Milota, Shoko Vos, Annelien Bredenoord, and Karin Jongsma. 2022. Integrating artificial intelligence in pathology: a qualitative interview study of users' experiences and expectations. *Modern pathology* 35, 11 (2022), 1540–1550.

[29] Ning Fang, Jon Pluyter, Saskia Bakker, Igor Jacobs, Misha Luyer, Joost Nederend, Jeroen Raijmakers, Lin-Lin Chen, and Mathias Funk. 2024. From Experience to Experience: Key Insights for Improved Interaction with AI in Radiology. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–7.

[30] Guo Freeman, Elizabeth D Mynatt, Cliff Lampe, Heloisa Candello, Kori Inkpen, and Nitesh Goyal. 2025. The Impact of Generative AI on the CSCW Landscape: Insights from HCI Education, Industry Dynamics, and Funding Perspectives. In *Companion Publication of the 2025 Conference on Computer-Supported Cooperative Work and Social Computing*. 62–65.

[31] Adam Graves, Leslie Jones, and Frances F Kaplan. 2013. Draw-a-Person-in-the-Rain: Does geographic location matter? *Art Therapy* 30, 3 (2013), 107–113.

[32] Simeng Gu, Yige Liu, Fei Liang, Rou Feng, Yawen Li, Guorui Liu, Mengdan Gao, Wei Liu, Fushun Wang, and Jason H Huang. 2020. Screening depressive disorders with tree-drawing test. *Frontiers in psychology* 11 (2020), 1446.

[33] Lijie Guo, Elizabeth M Daly, Oznur Alkan, Massimiliano Mattetti, Owen Cornec, and Bart Knijnenburg. 2022. Building Trust in Interactive Machine Learning via User Contributed Interpretable Rules. In *27th International Conference on Intelligent User Interfaces*. 537–548.

[34] David Gussak. 2001. *The work of the art therapist: An interactionist perspective.* Ph. D. Dissertation. Citeseer.

[35] Emanuel F Hammer. 1958. *The clinical application of projective drawings.* Charles C Thomas.

[36] Leonard Handler and Joseph Reyher. 1965. Figure drawing anxiety indexes: A review of the literature. *Journal of Projective Techniques and Personality Assessment* 29, 3 (1965), 305–313.

[37] Ellen G Horovitz. 2020. *The art therapists' primer: A clinical guide to writing assessments, diagnosis, and treatment.* Charles C Thomas Publisher.

[38] Jeonghwan Hwang, Taeheon Lee, Honggu Lee, and Seonjeong Byun. 2022. A clinical decision support system for sleep staging tasks with explanations from artificial intelligence: user-centered design and evaluation study. *Journal of medical Internet research* 24, 1 (2022), e28659.

[39] Chiara Ionio and Eleonora Mascheroni. 2021. Psychological well-being and graphic representations of self in child victims of violence. *The Arts in Psychotherapy* 72 (2021), 101740.

[40] Maia Jacobs, Jeffrey He, Melanie F. Pradier, Barbara Lam, Andrew C Ahn, Thomas H McCoy, Roy H Perlis, Finale Doshi-Velez, and Krzysztof Z Gajos. 2021. Designing AI for trust and collaboration in time-constrained medical decisions: a sociotechnical lens. In *Proceedings of the 2021 chi conference on human factors in computing systems*. 1–14.

[41] Glenn Jocher, Ayush Chaurasia, and Jing Qiu. 2023. *Ultralytics YOLO.* https://github.com/ultralytics/ultralytics

[42] Tim Johnson and Nick Obradovich. 2024. New methods for deprecating artificial intelligence systems will preserve history and facilitate research. *nature communications* 15, 1 (2024), 10254.

[43] Jiwon Kang, Jiwon Kim, Migyeong Yang, Chaehee Park, Taeeun Kim, Hayeon Song, and Jinyoung Han. 2024. SceneDAPR: A Scene-Level Free-Hand Drawing Dataset for Web-based Psychological Drawing Assessment. In *Proceedings of the ACM on Web Conference 2024*. 4630–4641.

[44] Daiki Kato and Miyako Morita. 2009. Form, content, and gender differences in Lego® block creations by Japanese adolescents. *Art Therapy* 26, 4 (2009), 181–186.

[45] Kensaku Kawamoto, Caitlin A Houlihan, E Andrew Balas, and David F Lobach. 2005. Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success. *Bmj* 330, 7494 (2005), 765.

[46] Jiwon Kim, Jiwon Kang, Taeeun Kim, Hayeon Song, and Jinyoung Han. 2023. AlphaDAPR: An AI-based Explainable Expert Support System for Art Therapy. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*. 19–31.

[47] Seong-in Kim, Kee-Eung Kim, and Seunghwan Song. 2024. Exploring artificial intelligence approach to art therapy assessment: A case study on the classification and the estimation of psychological state based on a drawing. *New Ideas in Psychology* 73 (2024), 101074.

[48] Terry K Koo and Mae Y Li. 2016. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of chiropractic medicine* 15, 2 (2016), 155–163.

[49] Kate Kravits, Randi McAllister-Black, Marcia Grant, and Christina Kirk. 2010. Self-care strategies for nurses: A psycho-educational intervention for stress reduction and the prevention of burnout. *Applied Nursing Research* 23, 3 (2010), 130–138.

[50] Christine P Krom. 2000. *Hospice nurses and the palliative care environment: Indicators of stress and coping in the Draw-a-Person-in-the-Rain test.* Ph. D. Dissertation. Albertus Magnus College.

[51] Heidi S Lack. 1997. *The person-in-the-rain projective drawing as a measure of children's coping capacity: a concurrent validity study using rorschach, psychiatric, and life history variables.* California School of Professional Psychology-Berkeley/Alameda. 156–174 pages.

[52] Jane Lessiter, Jonathan Freeman, Edmund Keogh, and Jules Davidoff. 2001. A cross-media presence questionnaire: The ITC-Sense of Presence Inventory. *Presence: Teleoperators & Virtual Environments* 10, 3 (2001), 282–297.

[53] Jie Li, Hancheng Cao, Laura Lin, Youyang Hou, Ruihao Zhu, and Abdallah El Ali. 2024. User experience design professionals' perceptions of generative artificial intelligence. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–18.

[54] Mengqi Liao and S Shyam Sundar. 2021. How should AI systems talk to users when collecting their personal information? Effects of role framing and self-referencing on human-AI interaction. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–14.

[55] Eva Fishell Lichtenberg. 2014. Draw-A-Person-in-the Rain Test. In *Drawings in assessment and psychotherapy: Research and application*, Leonard Handler and A. D. Thomas (Eds.). Routledge/Taylor & Francis Group., 164–183.

[56] Di Liu, Jingwen Bai, Zhuoyi Zhang, Yilin Zhang, Zhenhao Zhang, Jian Zhao, and Pengcheng An. 2025. TherAIssist: Assisting Art Therapy Homework and Client-Practitioner Collaboration through Human-AI Interaction. *arXiv preprint arXiv:2502.12443* (2025).

[57] Zhengliang Liu, Hanqi Jiang, Tianyang Zhong, Zihao Wu, Chong Ma, Yiwei Li, Xiaowei Yu, Yutong Zhang, Yi Pan, Peng Shu, et al. 2023. Holistic evaluation of gpt-4v for biomedical imaging. *arXiv preprint arXiv:2312.05256* (2023).

[58] Matthew Lombard, Jennifer Snyder-Duch, and Cheryl Campanella Bracken. 2002. Content analysis in mass communication: Assessment and reporting of intercoder reliability. *Human communication research* 28, 4 (2002), 587–604.

[59] Pin Luarn and Hsin-Hui Lin. 2005. Toward an understanding of the behavioral intention to use mobile banking. *Computers in human behavior* 21, 6 (2005), 873–891.

[60] Elle Christine Lüchau, Finn Olesen, Helen Atherton, Jens Søndergaard, and Elisabeth Assing Hvidt. 2024. The invisible work of video consultation use in Danish general practice: An ethnographic study. *Digital Health* 10 (2024), 20552076241291333.

[61] Karen Machover. 1949. Personality projection in the drawing of the human figure (A method of personality investigation). (1949).

[62] Maria Madsen and Shirley Gregor. 2000. Measuring human-computer trust. In *11th australasian conference on information systems*, Vol. 53. Citeseer, 6–8.

[63] Tanya Nazaretsky, Mutlu Cukurova, and Giora Alexandron. 2022. An instrument for measuring teachers' trust in AI-based educational technology. In *LAK22: 12th international learning analytics and knowledge conference*. 56–66.

[64] Don Norman. 2013. *The design of everyday things: Revised and expanded edition.* Basic books.

[65] Gerald D Oster and Patricia Gould Crone. 2004. *Using drawings in assessment and therapy: A guide for mental health professionals.* Routledge.

[66] Cecilia Panigutti, Andrea Beretta, Fosca Giannotti, and Dino Pedreschi. 2022. Understanding the impact of explanations on advice-taking: a user study for AI-based clinical Decision Support Systems. In *CHI Conference on Human Factors in Computing Systems*. 1–9.

[67] Sung Youl Park. 2009. An analysis of the technology acceptance model in understanding university students' behavioral intention to use e-learning. *Journal of Educational Technology & Society* 12, 3 (2009), 150–162.

[68] Robin Pennington, H Dixon Wilcox, and Varun Grover. 2003. The role of system trust in business-to-consumer transactions. *Journal of management information systems* 20, 3 (2003), 197–226.

[69] Ashish Viswanath Prakash and Saini Das. 2021. Medical practitioner's adoption of intelligent clinical diagnostic decision support systems: A mixed-methods study. *Information & management* 58, 7 (2021), 103524.

[70] Rob Procter, Mark Rouncefield, Peter Tolmie, and Clare Verrill. 2024. Everyday diagnostic work in the histopathology lab: CSCW perspectives on the utilization of data-driven clinical decision support systems. *Computer Supported Cooperative Work (CSCW)* 33, 4 (2024), 1261–1288.

[71] Ochilbek Rakhmanov, Nwojo Nnanna Agwu, and Steve Adeshina. 2020. Experimentation on hand drawn sketches by children to classify Draw-a-Person test images in psychology. In *The Thirty-Third International Flairs Conference*.

[72] Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530* (2024).

[73] CR Reynolds. 1985. Revised children's manifest anxiety scale. *Western Psychological Services* (1985).

[74] Cecil R Reynolds and Bert O Richmond. 1978. What I think and feel: A revised measure of children's manifest anxiety. *Journal of abnormal child psychology* 6 (1978), 271–280.

[75] Fujiko Robledo Yamamoto, Amy Voida, and Stephen Voida. 2021. From therapy to teletherapy: Relocating mental health services online. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–30.

[76] Jeremy D Safran and J Christopher Muran. 1994. Toward a working alliance between research and practice. (1994).

[77] Ahmet Atakan Salar, Hajar Faiyad, Elena Battini Sönmez, and Sibel Hafton. 2023. Artificial Intelligence Contribution to Art-Therapy using Drawings of the House-Person-Tree Test. In *2023 3rd International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME)*. IEEE, 1–6.

[78] Severi Santavirta, Yuhang Wu, Lauri Suominen, and Lauri Nummenmaa. 2025. GPT-4V shows human-like social perceptual capabilities at phenomenological and neural levels. *Imaging Neuroscience* 3 (2025), IMAG–a.

[79] Kjeld Schmidt and Liam Bannon. 1992. Taking CSCW seriously: Supporting articulation work. *Computer supported cooperative work (CSCW)* 1, 1 (1992), 7–40.

[80] Lilach Shaul, Lars De Vroege, and AHS van den Broek. 2024. Dealing with workload and stress complaints: The impact of resilience. *Psychology* 15, 9 (2024).

[81] Rania Shibl, Meredith Lawley, and Justin Debuse. 2013. Factors influencing decision support system acceptance. *Decision Support Systems* 54, 2 (2013), 953–961.

[82] Sarah C Slayton, Jeanne D'Archer, and Frances Kaplan. 2010. Outcome studies on the efficacy of art therapy: A review of findings. *Art therapy* 27, 3 (2010), 108–118.

[83] Michael E Sobel. 1982. Asymptotic confidence intervals for indirect effects in structural equation models. *Sociological methodology* (1982).

[84] Charles Donald Spielberger. 1970. Manual for the State-Trait Anxiety Inventory (self-evaluation questionnaire). *(No Title)* (1970).

[85] Charles D Spielberger, Fernando Gonzalez-Reigosa, Angel Martinez-Urrutia, Luiz FS Natalicio, and Diana S Natalicio. 1971. The state-trait anxiety inventory. *Revista Interamericana de Psicologia/Interamerican journal of psychology* 5, 3 & 4 (1971).

[86] Jingjing Sun, Jingyi Yang, Guyue Zhou, Yucheng Jin, and Jiangtao Gong. 2024. Understanding Human-AI Collaboration in Music Therapy Through Co-Design with Therapists. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–21.

[87] Reed T Sutton, David Pincock, Daniel C Baumgart, Daniel C Sadowski, Richard N Fedorak, and Karen I Kroeker. 2020. An overview of clinical decision support systems: benefits, risks, and strategies for success. *NPJ digital medicine* 3, 1 (2020), 17.

[88] Anja Thieme, Maryann Hanratty, Maria Lyons, Jorge Palacios, Rita Faia Marques, Cecily Morrison, and Gavin Doherty. 2023. Designing human-centered AI for mental health: Developing clinically relevant applications for online CBT treatment. *ACM Transactions on Computer-Human Interaction* 30, 2 (2023), 1–50.

[89] JS Verinis, EF Lichtenberg, and L Henrich. 1974. The Draw-a-Person in the rain technique: Its relationship to diagnostic category and other personality indicators. *Journal of Clinical Psychology* (1974).

[90] Smrithi Vijayakumar, V Vien Lee, Qiao Ying Leong, Soo Jung Hong, Agata Blasiak, and Dean Ho. 2023. Physicians' Perspectives on AI in Clinical Decision Support Systems: Interview Study of the CURATE. AI Personalized Dose Optimization Platform. *JMIR Human Factors* 10 (2023), e48476.

[91] Mikayla Voller. 2018. *The Current Use of Art Therapy Assessments: A Thematic Analysis*. Drexel University.

[92] Katharina Weitz, Dominik Schiller, Ruben Schlagowski, Tobias Huber, and Elisabeth André. 2021. "Let me explain!": exploring the potential of virtual agents in explainable AI interaction design. *Journal on Multimodal User Interfaces* 15, 2 (2021), 87–98.

[93] Fan Wenjuan, Jingnan Liu, Zhu Shuwan, and Panos M Pardalos. 2020. Investigating the impacting factors for the healthcare professionals to adopt artificial intelligence-based medical diagnosis support system (AIMDSS). *Annals of Operations Research* 294, 1-2 (2020), 567–592.

[94] Sigit Widiyanto and Jhordy Wong Abuhasan. 2020. Implementation the convolutional neural network method for classification the draw-A-person test. In *2020 Fifth International Conference on Informatics and Computing (ICIC)*. IEEE, 1–6.

[95] Lisa R Willis, Stephen P Joy, and Donna H Kaiser. 2010. Draw-a-Person-in-the-Rain as an assessment of stress and coping resources. *The Arts in Psychotherapy* 37, 3 (2010), 233–239.

[96] Barbara H Wixom and Peter A Todd. 2005. A theoretical integration of user satisfaction and technology acceptance. *Information systems research* 16, 1 (2005), 85–102.

[97] Ankita Yadav, Shalini Garg, ANKIT Shrivastava, Anil Gupta, Shikha Dogra, and Sakshi Joshi. 2020. Child drawing: A projective tool for dental anxiety assessment. *International Healthcare Research Journal* 4, 1 (2020), 19–25.

[98] Qian Yang, Aaron Steinfeld, and John Zimmerman. 2019. Unremarkable AI: Fitting intelligent decision support into critical, clinical decision-making processes. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–11.

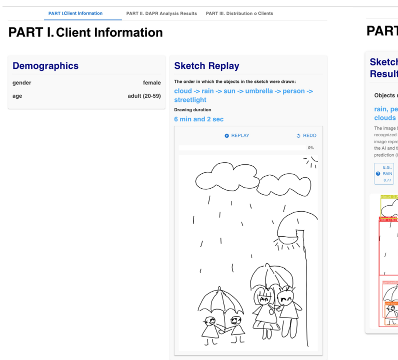
[99] Rui Zhang, Nathan J McNeese, Guo Freeman, and Geoff Musick. 2021. " An ideal human" expectations of AI teammates in human-AI teaming. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW3 (2021), 1–25.

[100] Yunfeng Zhang, Q Vera Liao, and Rachel KE Bellamy. 2020. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 295–305.

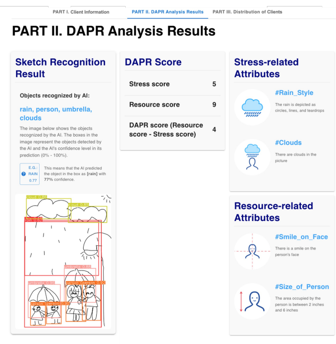
[101] Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. 2023. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923* (2023).

A RESEARCH METHOD

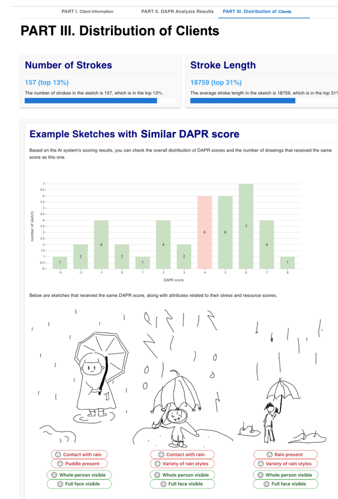
A.1 Comparisons between *AlphaDAPR* & *BetaDAPR*



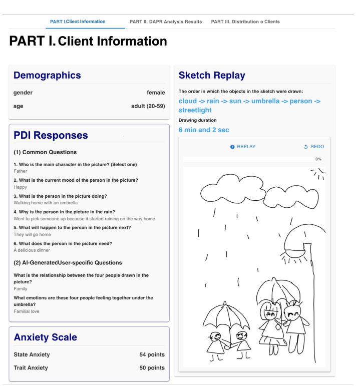
(a) Part I. Client Information of *AlphaDAPR*



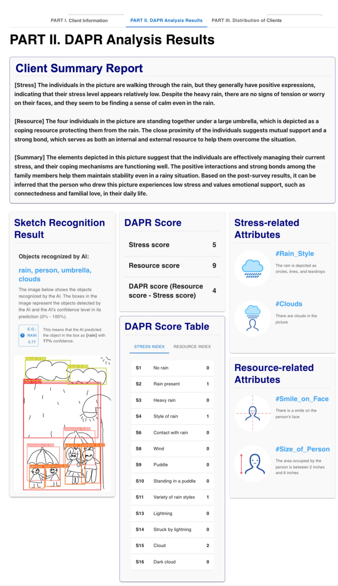
(b) Part II. DAPR Analysis Results of *AlphaDAPR*



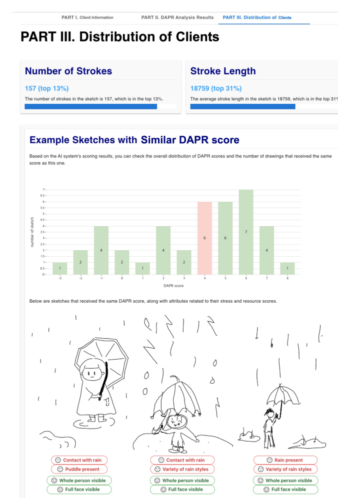
(c) Part III. Distribution of Clients of *AlphaDAPR*



(d) Part I. Client Information of *BetaDAPR*



(e) Part II. DAPR Analysis Results of *BetaDAPR*

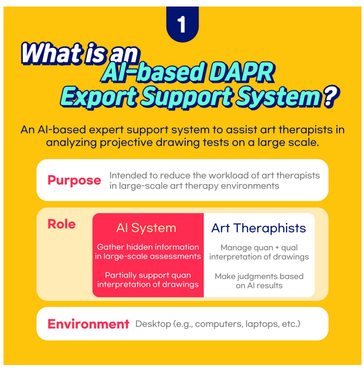


(f) Part III. Distribution of Clients of *BetaDAPR*

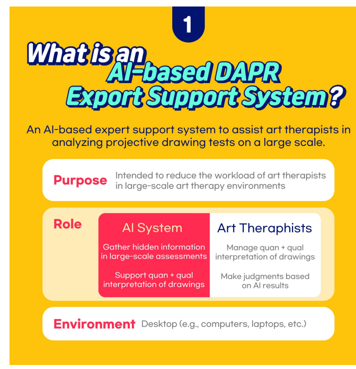
Fig. 7. Comparison of the user interface of *AlphaDAPR* and *BetaDAPR*.

Figure 7 shows a comparison of the user interface for *AlphaDAPR* and *BetaDAPR*. Figures 7(a), (b), and (c) are the three parts of *AlphaDAPR*'s user interface, whereas Figures 7(d), (e), and (f) are those of *BetaDAPR*'s user interface. The distinct components between the two interfaces are highlighted with a bold box border. As illustrated in Figures 7(a) and (d), in the client information part, *AlphaDAPR* provides a brief client demographic information and a sketch replay function, whereas *BetaDAPR* further displays PDI responses and anxiety scale results. In the DAPR analysis results part, while both interfaces offer basic DAPR score analysis, *BetaDAPR* additionally provides a client summary report and detailed DAPR item-level scores (see Figures 7(b) and (e)). Lastly, in distribution of clients part, both *AlphaDAPR* (Figure 7(c)) and *BetaDAPR* (Figure 7(f)) share the same user interface, which includes the number of strokes, stroke length, and example sketches with similar DAPR scores.

A.2 System Notice



(a) Explanation Page of *AlphaDAPR*



(b) Explanation Page of *BetaDAPR*

(c) Design Component Page of *AlphaDAPR*

(d) Design Component Page of *BetaDAPR*

(e) Final Page of *AlphaDAPR*

(f) Final Page of *BetaDAPR*

Fig. 8. Comparison of the system notices for *AlphaDAPR* and *BetaDAPR*.

Figure 8 compares the system notices for *AlphaDAPR* and *BetaDAPR*. While the overall format remains consistent, there are subtle differences in the specific details. Figures 8(a) and (b) display the explanation pages for *AlphaDAPR* and *BetaDAPR*, respectively. Figure 8(a) indicates that the AI system supports the quantitative interpretation of drawings, while Figure 8(b) explains that the AI system supports both quantitative and qualitative interpretations of drawings. Regarding the design component page, Figure 8(c) shows that design components of *AlphaDAPR* focuses on the quantitative aspects of DAPR analysis. In contrast, Figure 8(d) for *BetaDAPR* includes design components such as PDI responses and the client summary report, which require qualitative analysis performed by AI-based methods. The final page for both *AlphaDAPR* (Figure 8(e)) and *BetaDAPR* (Figure 8(f)) is identical, emphasizing that the final decisions are made by art therapists.

A.3 Measures

Table 8 presents the complete list of survey items which are used in our study.

Construct	Items	Source
Understandability	I understand how the system will assist me with decisions I have to make. It is easy to follow what the system does. I recognize what I should do to get the advice I need from the system the next time I use it.	[62]
Explainability	The textual information provided by this system helped me understand the decision-making process of the system. The visualisation provided by this system helped me understand the decision-making process of the system.	[92]
Perceived Ease of Use	It is easy for me to remember how to perform tasks using the system. Overall, I find the system easy to use.	[27]
Perceived Usefulness	I think using this system would enable faster diagnosis. I think using this system would improve my job performance. I think this system would enhance my diagnostic ability. I think this system would help reduce my work-related pressure. IOverall, I believe that this system would be useful for my work.	[27]
Satisfaction	I am satisfied with the information obtained from this system. My interaction with this system is satisfactory. Overall, I am satisfied with this system.	[96]
Trust	I trust the information provided by this system. I believe the results provided by this system are reliable. This system is trustworthy.	[68]
Technology Acceptance	I usually keep an eye on emerging technology products. I always try out new technology products earlier compared to others.	[93]
AI Familiarity	AI will be a good human assistant. AI will bring positive changes to art therapy in the future.	[46]
Co-Presence	I felt like there was an agent assisting with the actual DAPR drawing assessment while using this system. I felt like an agent assisting with the DAPR drawing assessment was talking to me nearby while using this system.	[58]
Physical Presence	I felt like I was actually analyzing DAPR drawing assessments while using this system. I felt like I was in a place where DAPR drawing assessments are actually conducted while using this system.	[52]
Intention to Use	I would be willing to use this system if it becomes available. Assuming that I have assess to the system, I would consider using this system.	[59, 67]

Table 8. Survey items used in the study. All items were measured on a 5-point Likert scale (1 = Strongly Disagree, 5 = Strongly Agree).

B RESULTS

B.1 Two-way analysis of variance (ANOVA)

Table 9 presents the detailed results of the two-way analysis of variance (ANOVA) for each measurement questionnaire item. These results provide further detail on the box plot shown in Section 5.2.

		df	F	p-value			df	F	p-value
Understandability	AB	1.00	1.66	0.203	Trust	AB	1.00	2.93	0.093
	JS	1.00	0.10	0.748		JS	1.00	4.21	0.045*
	AB:JS	1.00	0.65	0.424		AB:JS	1.00	3.54	0.065
	Residual	56.00				Residual	56.00		
Explainability	AB	1.00	2.17	0.146	Technology acceptance	AB	1.00	0.05	0.816
	JS	1.00	2.17	0.146		JS	1.00	0.15	0.698
	AB:JS	1.00	0.06	0.807		AB:JS	1.00	0.15	0.698
	Residual	56.00				Residual	56.00		
Satisfaction	AB	1.00	5.12	0.028*	AI familiarity	AB	1.00	0.52	0.474
	JS	1.00	0.50	0.483		JS	1.00	0.96	0.330
	AB:JS	1.00	0.02	0.888		AB:JS	1.00	0.35	0.558
	Residual	56.00				Residual	56.00		
Preceived ease of use	AB	1.00	5.04	0.029*	Co-presence	AB	1.00	0.07	0.793
	JS	1.00	0.25	0.620		JS	1.00	0.01	0.930
	AB:JS	1.00	0.25	0.620		AB:JS	1.00	0.62	0.433
	Residual	56.00				Residual	56.00		
Perceived usefulness	AB	1.00	6.16	0.016*	Physical presence	AB	1.00	0.77	0.384
	JS	1.00	0.20	0.659		JS	1.00	0.12	0.727
	AB:JS	1.00	0.00	1.000		AB:JS	1.00	6.03	0.017*
	Residual	56.00				Residual	56.00		
Intention to use	AB	1.00	6.01	0.017*					
	JS	1.00	0.32	0.574					
	AB:JS	1.00	0.00	1.000					
	Residual	56.00							

Table 9. Results of ANOVA by measurement questionnaire items. 'AB' denotes the main effect of factors in *AlphaDAPR* and *BetaDAPR*, whereas 'JS' denotes the main effect of factors in *Juniors* and *Seniors*. 'AB:JS' represents the interaction effect between both factors.

Received May 13, 2025; revised January 13, 2026; accepted March 17, 2026