

A VR-based Method for Observing Caregiver Intervention Timing under Standardized Childcare Conflict Scenes

Shigeru Owada ^{1,*}, Kokone Sato ^{1,*}, Shunsuke Kikuchi ², Hidenori Aoki ^{1,*}, Momoko Nakatani ²,
Yuki Taoka ³, Tetsuhito Sakata ⁴, Midori Takahashi ⁵, and Mikako Ishibashi ⁶

¹ Sony Computer Science Laboratories, Inc., Research Lab., Tokyo, Japan

² School of Engineering, Institute of Science Tokyo, Tokyo, Japan

³ School of Environment and Society, Institute of Science Tokyo, Tokyo, Japan

⁴ Department of Child Studies, Faculty of Home Economics, Otsuma Women's University, Tokyo, Japan

⁵ Graduate Course in Clinical Psychology, Graduate School of Humanities, Gakushuin University, Tokyo, Japan

⁶ Department of Psychology and Humanities, College of Sociology, Edogawa University, Tokyo, Japan

Email: owd@hoikutech.com (S.O.); 22b099@shohoku.ac.jp (K.S.); kikuchi.s.ag@m.titech.ac.jp (S.K.); Hidenori.Aoki@sony.com (H.A.); nakatani.m.ad@m.titech.ac.jp (M.N.); taoka.y.aa@m.titech.ac.jp (Y.T.); tsakata@otsuma.ac.jp (T.S.); hisui.bluegreen@gmail.com (M.T.); mishibas@edogawa-u.ac.jp (M.I.)

*Corresponding author

Manuscript received January 16, 2026; revised February 26, 2026; accepted April 23, 2026; published October 16, 2026

Abstract—Childcare practice is highly context-dependent and difficult to replicate, making it challenging to observe the behavior of multiple caregivers under standardized conditions. This study proposes a Virtual Reality (VR)-based methodological framework in which a virtual nursery environment was developed where caregivers experience conflict situations involving young children, and their intervention behavior is systematically observed using a head-mounted display. Twenty-nine caregivers participated in an initial questionnaire-based profiling phase, followed by VR-based intervention tasks. The results indicate that the VR environment can capture substantial individual differences in the timing of caregiver interventions under controlled conditions. Exploratory analyses further suggest that caregiver profiles—such as beliefs about child-initiated problem solving and perceived workplace stress—were associated with variation in intervention timing; for example, lower perceived workplace stress was associated with later intervention timing (Nursery Teacher's Stress Scale (NTSS): Workplace Relationships and Lack of Time; $|r| > 0.45$, $p < 0.05$). Furthermore, not only intervention timing but also active visual exploration (head rotation) showed a moderate association with a core pedagogical belief (Social Conflict Inventory (SCI): Facilitation, valuing child self-resolution) ($r > 0.43$, $p = 0.02$). These results suggest that the VR system can capture an important, yet practically underexplored, dimension of deliberate non-immediate intervention skill valued in pedagogical standards. This VR-based framework provides a foundational method for objective behavioral observation, pointing toward future large-scale validation studies and the development of VR-supported professional development programs. However, the sample size in this study is limited, with biases such as in years of experience and the presence or absence of compensation; therefore, statistically significant results should be interpreted as purely exploratory associations. Furthermore, it should be noted that standardizing the experience in this way inherently creates a critical trade-off with the ecological validity of the scenes.

Keywords—childcare, virtual reality, caregiver, intervention timing, Early Childhood Education and Care (ECEC)

I. INTRODUCTION

Improving the quality of childcare is an important issue, regardless of cultural or historical context [1]. High quality Early Childhood Education and Care (ECEC) has been demonstrated internationally to have a modest long-term effect on children's cognitive and socioemotional

developmental trajectories, especially for children from more at-risk backgrounds (often referred to as a “compensation effect”, the long-term benefit for children from at-risk backgrounds) [2]. Heckman further emphasizes that investment in the early years is crucial, as early skill formation determines the efficiency of later learning (dynamic complementarity, which suggests that early skills influence later learning efficiency) [3]. While childcare quality encompasses multiple dimensions, in organized settings such as nurseries, particular attention is directed toward the quality of everyday childcare practices—namely, the “process quality” (how staff interact with children and implement curriculum) [1]. However, there is generally no clear consensus regarding which specific practices best constitute high process quality, and caregivers continue their practice while struggling and experiencing uncertainty. In the Japanese context where this study was conducted, quality is evaluated against pedagogical standards such as the National Curriculum Standard for Nurseries (Hoikujo Hoiku Shishin) [4]—a national guideline that defines pedagogical quality standards in Japan, but practitioners face practical challenges in applying these standards, particularly in the complex task of caring for one-year-olds [5], which leads to ongoing debates regarding quality assessment methodologies [6]. Such uncertainty arises from the caregiver's own beliefs, institutional policies, relationships with parents, and cultural context [7].

For example, conflicts and acts of aggression between children occur relatively frequently [8]. While they may appear negative and sometimes require immediate intervention, it is widely recognized that such conflicts can play an important role in child development [9]. International research further explores the conditions that lead to educator intervention and the importance of promoting competent conflict resolution skills [10]. The necessity of high-level professional judgment in this universal dilemma is underscored by studies showing that emotional or physical intervention by educators, such as when “words do not work”, can sometimes lead to new conflicts between the child and the caregiver [11]. Furthermore, as children progress through developmental stages, their ability to solve problems independently increases [8]. Accordingly, caregivers must assess each child's personality and developmental stage in

order to determine appropriate intervention timing and methods. This requires highly sophisticated judgment, and as a result, approaches vary considerably across caregivers.

Critically, the timing of intervention—the decision of when to step in—is an underexamined yet pedagogically crucial dimension of high process quality. The Japanese National Curriculum Standard for Nurseries (Hoikujo Hoiku Shishin) [4] emphasizes fostering children's self-resolution, making the decision of when to intervene a core element of expert practice.

Prior ECEC research has extensively examined the diversity of caregiver responses to peer conflicts, focusing primarily on: (1) Intervention Styles and Strategies [12–14], (2) The Influence of Personal and Contextual Factors (e.g., beliefs [8], experience [15], and workplace culture [7]), and (3) The inherent Moral and Professional Dilemmas caregivers face [7, 16]. Critically, these studies predominantly address the intervention strategy (how to intervene) rather than the intervention timing (when to intervene). The existence of such diversity should not necessarily be viewed negatively. What constitutes an appropriate intervention depends on the caregiving philosophy of the individual caregiver, the team, or the organization as a whole. Moreover, for children, learning that adults differ in their approaches to caregiving offers valuable insight into the inherently diverse nature of the world. Therefore, this study does not aim to evaluate the appropriateness of specific intervention behaviors.

This extensive body of work has not yet empirically examined a critical element: real-time intervention timing under controlled conditions.

Specifically, traditional research methodologies impose significant limitations that prevent the objective, comparative analysis of intervention timing under standardized conditions. Most previous studies relied on questionnaires, interviews, or field observations; however, each of these approaches has limitations. Questionnaires and interviews depend on subjective self-reports, whereas field observations make it difficult to directly compare practices due to limited reproducibility. Consequently, the real-time dynamics of intervention timing have remained methodologically inaccessible.

Despite extensive research on caregiver beliefs and practices, there is limited empirical work that systematically observes when caregivers intervene under standardized conditions.



Fig. 1. Childcare workers in the virtual nursery are required to intervene in conflict situations involving children. The system was used to explore the relationship between the intervention timing and the caregivers' prior profiling.

Therefore, this study proposes a novel, VR-based methodological framework uniquely designed to address this

research gap, allowing for the systematic observation of real-time intervention timing under controlled conditions. In this study, a virtual nursery was constructed within a VR environment, and conflict scenes between children were created (Fig. 1). Caregivers wore Head-Mounted Displays (HMDs) and were asked to intervene in the virtual space as naturally as possible. When a caregiver touched a child, this action was registered as an intervention, and the scene was terminated. The actions and elapsed time until intervention were recorded.

The methodological power of this approach lies in its ability to standardize the observational context:

- Scenes can be designed to suggest individual differences among caregivers.
- The same situation can be presented to all caregivers.
- The sensors built into the HMD allow for accurate recording of caregivers' behaviors.

This standardization makes possible the objective, quantitative evaluation of the pedagogically critical dimension of intervention timing. While these attributes provide high internal validity and reproducibility, they inherently impose critical limitations on the ecological validity, which must be addressed in the Discussion section.

Prior to the experiment, caregivers completed a profiling phase based on several scales. The following three validated scales were used to assess the corresponding caregiver factors (aptitude, stress, and beliefs).

- Childcare Worker Aptitude Scale (AS) [17]: measures general professional aptitude and qualities
- Nursery Teacher's Stress Scale (NTSS) [18]: measures daily workplace stress factors
- The Social Conflict Inventory (SCI) [19]: measures beliefs about intervention during peer conflicts

By linking these responses with behaviors observed in the virtual nursery, it becomes possible to identify factors that influence childcare practices. Although this study is exploratory in nature, identifying the factors that shape such practices through VR has the potential to contribute to improved care quality by providing objective data for reflection. This paper reports on the design of the VR system and the insights gained from an initial experiment involving 29 caregivers.

This study addresses the following exploratory research questions:

RQ1: Can a VR-based nursery environment capture individual differences in the timing of caregiver interventions under standardized conditions?

RQ2: What caregiver-related factors may be associated with differences in intervention timing when examined exploratorily?

Note on Scope: The analysis for RQ2 is purely exploratory, seeking to identify associations and indicative trends; it is not intended to provide explanations or predictions of behavior. Consequently, causal inference is outside the scope of this study.

II. LITERATURE REVIEW

A. Caregivers' Hesitation and Contributing Factors

The question of whether to prioritize children's safety (care) or their learning and development (education) has

remained a source of hesitation among caregivers [16]. However, the conflicts faced by caregivers have not been solely considered from the perspective of care and education. Rooijen investigated the factors influencing caregivers' attitudes toward children's risky play and developed a caregiver behavior model based on Bronfenbrenner's ecological model [7, 20]. According to this model, caregivers experience a dilemma (moral friction) between ensuring children's safety or allowing them autonomy (to protect the child versus to let them go), as well as between team unity versus diversity (should we act united or is professional diversity acceptable?). The need to resolve this moral friction, combined with the process of professional identity formation, is realized in practice as the complex, real-time decision of when to intervene (the temporal dimension of decision-making). The identified contributing factors include differences in personal beliefs and values, years of experience and training history, peer pressure within teams (i.e., regarding "what is normal"), external pressures (such as parents, supervisors, or regulations), and cultural background [7]. This cultural influence is critical, as cross-cultural studies, such as *Preschool in Three Cultures*, highlight how different national educational styles reflect distinct cultural values regarding how children should be raised [21]. Recent studies have further divided the process of caregivers' professional identity formation into sources and determinants. Sources include both internal (personal growth needs or childhood experiences) and external factors (other caregivers, social media, and books). The determinants include various aspects such as psychological burden from parents or teams, workplace atmosphere, and views on the profession [15]. Thus, caregivers navigate various dilemmas and find ways to reconcile them while responding to a continuous flow of events.

B. Caregiver Profiling

The "Childcare Worker Aptitude Scale (AS)" developed by Fujimura, was designed to measure aptitude for becoming a caregiver [17]. This scale aims to capture the qualities, attitudes, and abilities that aspiring caregivers should possess from multiple perspectives and consists of seven subscales including "interpersonal skills", "self-understanding", "understanding of children", and "emotional stability". Its reliability and validity have been verified through questionnaire surveys, and it is recognized as a meaningful indicator that can be used for education and support within caregiver training programs.

To measure caregiver competencies and stress internationally, tools such as the Child Care Worker Job Stress Inventory are widely used [22]. However, our study required a measure of daily workplace stress specific to Japanese nursery teachers.

The "Nursery Teacher's Stress Scale (NTSS)", developed by Akada, was designed to measure the stress factors that nursery teachers experience in their daily work [18]. It is widely known in Japan as an effective tool for supporting mental health and improving work environment in childcare settings. The scale consists of six subscales, including "difficulty in understanding children", "interpersonal relationships", "dealing with parents", and "lack of time". Its reliability and validity have been confirmed.

Caregivers' behavior is likely to be influenced by personal beliefs about what constitutes appropriate intervention. For example, some may believe that it is important for children to return home without injury, whereas others may believe that it is better to emphasize learning from failure despite occasional trouble. It is assumed that caregivers hold a variety of such beliefs and that these fundamental attitudes affect the timing and nature of their interventions. One method for investigating beliefs about interventions is the Social Conflict Inventory (SCI), which consists of 20 items [19]. The SCI is a self-report tool and includes three subscales. The first subscale, "General Orientation to Classroom Peer Conflict", includes items asking whether the caregiver would intervene immediately in various situations involving peer conflict. The second subscale, "Cessation", assesses whether stopping conflicts immediately is prioritized. The third subscale, "Facilitation", differs slightly in emphasis and asks whether adults should resolve problems or emphasize children's self-resolution. The original SCI study did not link responses with actual practice; therefore, the relationship between beliefs and behavior remains unclear.

The professional competencies required for high-quality childcare extend beyond domain-specific knowledge to encompass a range of interpersonal, observational, and decision-making skills. Internationally, frameworks such as the Classroom Assessment Scoring System (CLASS) [23] conceptualize caregiver quality in terms of emotional support, classroom organization, and instructional support. However, caring for infants and toddlers—particularly one-year-olds, who are the focus of this study—demands a distinct skill profile characterized by heightened sensitivity to non-verbal cues, tolerance of ambiguity during conflict situations, and the capacity for timely yet deliberate intervention [5]. This skill profile is shaped by a complex interplay of personal dispositions (e.g., self-efficacy, sociability), professional socialization within teams, and the broader cultural-pedagogical context in which caregivers operate [7, 21]. In Japan, this context is defined in part by the National Curriculum Standard for Nurseries (Hoikujo Hoiku Shishin) [4], which explicitly positions the facilitation of children's self-resolution as a core professional competency. The present study operationalizes one critical dimension of this skill profile—intervention timing, or professional delay—as a quantitatively observable behavioral construct.

C. The Combination of Childcare Settings and Virtual Reality (VR)

The combination of childcare settings and VR is not entirely new. Prior VR systems and research have primarily focused on: perspective-taking (e.g., experiencing the child's view) [24, 25], emotional training (e.g., influencing caregiver emotions) [26], or general simulation (e.g., virtual tours) [27, 28]. However, to the best of our knowledge, no system exists that is specifically designed to highlight differences between caregivers in intervention scenes through artificially constructed situations or that enables detailed recording and comparison of actions within the VR environment. The present study extends the VR-in-ECEC literature by shifting the focus from experience and training to objective, quantitative behavioral measurement and comparison.

III. MATERIALS AND METHODS

Our experiment consists of three phases.

- (1) Pre-experiment questionnaire
- (2) Intervention experiment in the virtual nursery
- (3) Post-experiment questionnaire

Given the exploratory nature of this study and limited sample size, all statistical analyses were conducted to identify potential trends rather than to test predefined hypotheses or establish causal relationships.

A. Pre-experiment Questionnaire

A pre-experiment questionnaire was administered online using Google Forms. The basic profile information included gender, age, childcare experience, final academic background, and current class assignments. Additionally, three scales and inventories referenced in the Related Work section were used: the Childcare Worker Aptitude Scale (AS) [17], the Nursery Teacher's Stress Scale (NTSS) [18], and the Social Conflict Inventory (SCI) [19] (Appendix A, B).

In this experiment, to avoid making the pre-experiment questionnaire excessively long, only the two questions with the highest factor loadings from each of the seven AS subscales were used rather than the full set. The original NTSS includes the "V. Stress over salary" subscale; however, because it was not relevant to the present experiment, it was excluded. To reduce the number of items, only items with factor loadings of 0.6 or higher were selected from the remaining subscales.

We acknowledge that this necessary reduction, undertaken to minimize participant burden during this exploratory phase, introduces a trade-off where the breadth of the original construct coverage is inherently compromised. This item selection strategy—prioritizing items with the highest factor loadings—was employed to ensure that the remaining items capture the core, central components of each subscale, which is adequate for their use as exploratory profiling variables in this preliminary study.

Table 1. Cronbach's α coefficients for reduced scales

Scale	Items	α (Large Sample, $N = 741$)	α (Current VR Study, $N = 29$)
AS: Altruism	2	0.845	0.835
AS: Empathy	2	0.875	0.836
AS: Logical Thinking	2	0.750	0.677
AS: Tactfulness	2	0.823	0.857
AS: Sociability	2	0.915	0.978
AS: Initiative	2	0.835	0.764
AS: Nurturance	2	0.883	0.904
AS Overall	14	0.907	0.899
NTSS: Interaction and Understanding	5	0.880	0.922
NTSS: Workplace Relationships	5	0.899	0.863
NTSS: Dealing with Parents	4	0.891	0.841
NTSS: Lack of Time	2	0.844	0.703
NTSS: Misalignment with Nursery Policies	2	0.842	0.830
NTSS Overall	18	0.914	0.900

To ensure the scales' reliability despite the item reduction, we performed internal consistency analyses (Cronbach's

alpha) for the reduced item sets. We utilized both the data from the current VR experiment ($N = 29$) and a separate, large-scale dataset ($N = 741$) where these identical reduced scales were administered. The results, presented in Table 1, indicate high internal consistency for all subscales and overall scales across both samples. Specifically, the overall Cronbach's α for AS was 0.899 ($N = 29$) and 0.907 ($N = 741$), and for NTSS, it was 0.900 ($N = 29$) and 0.914 ($N = 741$). These robust coefficients support the stability and appropriate internal consistency of the reduced scales for their use in this exploratory study.

The primary purpose of collecting the pre-experiment questionnaire data (AS, NTSS, and SCI scores) was to serve as a set of profiling variables (psychological factors and beliefs) for the exploratory analysis. These data were subsequently cross-referenced with the behavioral metrics obtained during the VR experiment (e.g., intervention timing and head rotation speed) and the post-experiment self-report data. This systematic linkage aimed to identify potential correlations and influences between caregivers' prior psychological profiles and their observed in-VR behavior, as indicated in the Results section.

B. Virtual Nursery

The intervention experiment in the virtual nursery was conducted at least one week after the participants completed the pre-experiment questionnaire. This interval was intended to prevent participants' memories of their questionnaire responses from influencing their intervention behavior.

After the VR-based experiment, a brief multiple-choice question was administered to determine the reason for the intervention, followed by an open-ended questionnaire to investigate any confusion related to the VR experience or other issues encountered during the experiment.

1) Practice stages

Before the three main experimental scenes, two practice stages were developed to help participants become accustomed to wearing the HMD and learn how to operate within the virtual space. The purpose was to reduce the likelihood of situations in which the participants might feel confused by the controls and fail to intervene before the scene ended.

2) Main scenes

To establish pedagogical representativeness, scene designs were grounded in informal interviews with four veteran caregivers and one education researcher. These experts identified scenes—specifically focusing on 1-year-olds—that represent core professional dilemmas where variations in intervention timing are most likely to occur in practice. The scenes were structured to gradually escalate conflict, culminating in a physical altercation, a key feature that provides a time-window for caregivers to exercise professional judgment, making them representative of pedagogically important real-world decision points. Three distinct scenes were created, each varying in the type of conflict and number of characters involved. An overview of this process is shown in Fig. 2.

Owing to budget constraints, we were unable to provide characters with facial expressions. The voices of the characters were created by recording babbling sounds in a nursery room.



Fig. 2. Important phases in the three scenes.

a) Scene 1: Tupperware walking

This scene involved only one child. First, the child walks to a toy shelf empty-handed, picks up a container filled with toys, dumps the contents onto the floor, drops the container near their feet, puts one foot into it, and begins walking. After a short time, the child slips while wearing the container and falls backward.

b) Scene 2: Blockplay to biting

Two children were included in each scene. One child is seated and playing with blocks. Another child arrives, sits next to them, takes the blocks away, and begins to play independently. Eventually, the first child realizes that the blocks have been taken and tries to retrieve them, leading to a tug-of-war. When they are unable to retrieve the blocks, the child falls backward. Finally, the child bites the arm of the one who took the blocks.

c) Scene 3: Child unable to play

Four children were included in each scene. Three of them sit and play, while the remaining child stands at a distance, observing the others. The standing child briefly approaches the closest one and then retreats. They then approach again and take a toy from one of the seated children. After returning to their original position, they finally run headfirst into one of the seated children, headbutting them to end the scene.

3) VR application development

A virtual nursery application was developed using Unity and deployed on the PICO4 Enterprise. The device records head position and orientation, hand positions and orientations, speech (audio data), and video screenshots of the user's view. The results were saved on the HMD as CSV files and MP4 video files with audio, which were then transferred to a PC via a direct connection.

Children's movements were performed by adult actors and captured using a motion capture system (OptiTrack by Acuity). The recorded data were then retargeted to animate the child models in Unity.

C. Post-experiment Questionnaire

After the virtual nursery experiment, participants completed a self-administered questionnaire to identify the factors that influenced their behavior.

Specific items were developed with reference to Triandis's behavioral model [29]. In this model, factors that determine behavior include "Beliefs about Outcomes", "Evaluation of Outcomes", "Norms", "Roles", "Self-concept", "Emotions", and "Frequency of Past Behavior". Based on these categories, corresponding questionnaire items were created

(Appendix C).

Participants were also asked to freely describe their impressions of the experiment conducted in the virtual nursery.

D. Participants and Ethical Considerations

This experiment was conducted with signed consent from caregiver participants based on a consent form approved by Sony's Bioethics Committee. Due to institutional constraints, compensation varied across participants. Specifically, 15 participants (from university-affiliated nurseries and a child welfare institution) received 2000 yen compensation, while 14 participants from a public nursery declined compensation and received none. This unequal compensation structure is acknowledged as a potential source of bias, which is addressed through an exploratory analysis in the Results section.

The experiment was conducted with 29 participants: two caregivers each (totaling eight) from four university-affiliated nurseries in Osaka Prefecture, seven caregivers from a child welfare institution (kodomo-en) in Tokyo, and fourteen caregivers from a public nursery in Chiba Prefecture.

At all stages of the experiment, no personal information was collected from participants; instead, they were managed using unique IDs. To ensure strict anonymization, a QR code was displayed after the pre-experiment online questionnaire, which was then used at the reception of the virtual nursery experiment to link participant data. The same procedure was applied to the postexperiment questionnaire and interviews.

IV. RESULTS

The experiment was conducted between February and June 2025 with 29 participants (Fig. 3). Because the facilities were located in Tokyo, Chiba, and Osaka, and each session took more than 20 min per participant, it was not feasible to conduct the experiment with many participants in a single day. Therefore, the experiment was conducted across six separate days.



Fig. 3. Childcare workers wearing HMD.

A. Basic Profile

The basic profiles of the participants are shown in Fig. 4. Participants were fairly evenly distributed across their 20 s and 40 s. In Japanese childcare centers, the vast majority of caregivers are women, and in this experiment as well, only 2 out of the 29 participants were men. The highest level of education was junior college. More than half of the participants had over ten years of experience in early childhood education. Compared with the national distribution in Japan, this sample included a slightly higher proportion of

experienced caregivers (i.e., overrepresentation of veteran staff). This inherent sampling bias must be strongly acknowledged upfront as a critical constraint, as it limits the generalizability of our statistical inferences to the broader, national population of caregivers, particularly those with less experience.

Specifically, this demographic skew means that our findings predominantly reflect the behaviors and psychological profiles of experienced professionals. For instance, the observed trends regarding later intervention timing in the experienced cluster (C2 in Scene 3) and the early intervention of the veteran subgroup (C3, suggestive of “defensive caregiving” discussed in Section V.C) are specifically shaped by this experienced demographic. This overrepresentation is likely due to the large number of participants from public nurseries, which tend to have fewer new staff members.

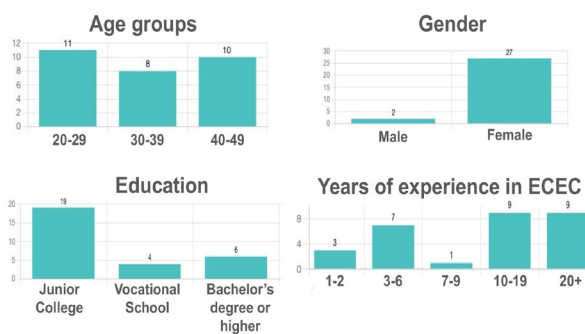


Fig. 4. Basic participant profiles.

B. Primary Finding 1: Individual Differences in Intervention Timing

Fig. 5 shows the time at which each caregiver intervened in each scene. The horizontal axis represents the elapsed time from the start of the scene, and the vertical axis represents the number of participants. The key phases of the scenes, as outlined in Fig. 2, are marked with balloons. As expected, the intervention timing was widely distributed across all scenes, indicating significant individual differences. Additionally, caregivers' actions formed clusters around certain phases of the scenes.

The total number of participants plotted in each scene did not add up to 29 because after each intervention, participants were asked whether the intervention reflected their true intent.

If they answered that it did not—for example, if they accidentally touched a child or if the scene ended while they were still unsure—those responses were treated as invalid and excluded from the data. The number of valid data points was eighteen in Scene 1, seventeen for Scene 2, and nineteen for Scene 3.



Fig. 5. Time until intervention.

Notably, every scene had some caregivers who intervened before the first event occurred. In Scenes 2 and 3, there was one such individual each, whereas in Scene 1, there were six individuals. Although it is unclear whether these interventions were truly intentional, these samples were not discarded as invalid. This may have influenced the subsequent analyses.

C. Primary Finding 2: Exploratory Association with Psychological Factors (Correlation)

Table 2 presents the correlation coefficients between the time until intervention in each scene and the subcategories of the pre-experiment questionnaire responses. Only the cells with larger absolute values of the coefficients are shown. Positive correlations are shown in purple and negative correlations are in blue. Cells with an absolute correlation coefficient greater than 0.4 are highlighted in red, while those with a p-value less than 0.05 are outlined in green. From this table, the following observations can be made:

Table 2. Intervention timing correlations

Intervention time scene	Intervention time Scene 1	Intervention time Scene 2	Intervention time Scene 3	AS: Sociability	AS: Initiative	AS: Nurturance	SCI: Facilitation	NTSS: Workplace Relationship	NTSS: Lack of Time
Scene 1	-	$r = 0.47$ $p = 0.10$	$r = 0.19$ $p = 0.56$	$r = 0.02$ $p = 0.93$	$r = 0.27$ $p = 0.27$	$r = 0.56$ $p = 0.02$	$r = 0.01$ $p = 0.98$	$r = -0.30$ $p = 0.23$	$r = -0.23$ $p = 0.36$
Scene 2	-	-	$r = 0.75$ $p = 0.00$	$r = 0.47$ $p = 0.06$	$r = 0.72$ $p = 0.00$	$r = 0.42$ $p = 0.09$	$r = 0.51$ $p = 0.04$	$r = -0.45$ $p = 0.07$	$r = -0.41$ $p = 0.10$
Scene 3	-	-	-	$r = 0.49$ $p = 0.03$	$r = 0.65$ $p = 0.00$	$r = 0.19$ $p = 0.43$	$r = 0.56$ $p = 0.01$	$r = -0.50$ $p = 0.03$	$r = -0.48$ $p = 0.04$

- The correlation analysis of intervention timing across scenes showed a strong correlation between Scenes 2 and 3, whereas correlations involving Scene 1 were low. This indicates that Scene 1 produces a different response pattern than the other two scenes. In other words, participants who intervened early (or late) in Scene 2 tended to do the same as in Scene 3; however, their timing in Scene 1 was not

strongly correlated.

- Among the pre-questionnaire subcategories, the only one that showed a moderate effect size correlation with intervention timing in Scene 1 was AS: Nurturance. All other correlations showed small descriptive trends (absolute values of 0.4 or lower). AS: Nurturance measured the degree to which a caregiver was willing to

invest effort for the sake of children. Interestingly, higher scores on this scale were associated with later interventions, which represents a noteworthy descriptive trend in this exploratory analysis.

- When focusing on Scenes 2 and 3, the pre-questionnaire subcategory with the highest positive correlation with intervention timing was AS: Initiative, followed by SCI: Facilitation. Since both correlations were positive, participants who reported having a strong sense of initiative or who valued children's self-resolution tended to intervene later.
- For Scenes 2 and 3, among correlation coefficients with absolute values greater than 0.4, a positive correlation was observed with AS: Sociability, whereas negative correlations were found with NTSS: Lack of Time and NTSS: Workplace Relationships. These results suggest that higher sociability and lower perceived workplace stress are moderately associated with later intervention.

D. Exploratory Supplementary Analysis: Multiple Regression

The correlation analysis identified the relationships observed when examining each variable individually. To investigate how these factors interact when considered simultaneously, an exploratory multiple regression analysis was conducted using the six subcategories that indicated the highest individual correlations ($|r| > 0.4$)—AS: Nurturance, AS: Initiative, SCI: Facilitation, AS: Sociability, NTSS: Lack of Time, and NTSS: Workplace Relationships—as predictor variables to estimate the time until intervention.

Ideally, all items from the pre-experiment questionnaire would be included as predictors. However, this study's primary objective is exploratory, aiming to identify indicative trends for future research, while simultaneously navigating a significant methodological constraint: the small sample size ($N = 29$). Including all variables would result in an under-constrained system, rendering the model unstable and making it difficult to extract meaningful, albeit

exploratory, insights. These variables were pre-selected because they represent established theoretical constructs (AS, NTSS, SCI) that capture attitudes and stress factors highly relevant to the problem of intervention timing, ensuring that the selected predictors are theoretically grounded. Each variable was normalized such that the minimum value was zero and the maximum value was one.

The results of the multiple regression analysis are presented in Table 3. The subscales were sorted in descending order based on the absolute values of their coefficients. In Scene 1, AS: Nurturance, which showed the highest individual correlation, also had the largest standardized coefficient in the multiple regression analysis. This strong and consistent descriptive trend across both exploratory methods is noteworthy, even though the result did not meet the threshold for statistical significance ($p = 0.178$).

In the multiple regression analysis for Scene 2, NTSS: Workplace Relationships emerged as the variable with the largest coefficient (strongest effect). This subscale also made the second-largest contribution in Scene 3, indicating that workplace relationships may influence caregiver intervention behavior. AS: Initiative showed the highest contribution in Scene 3 ($p = 0.037$), but was not statistically significant in Scene 2 ($p = 0.244$). In Scene 2, NTSS: Lack of Time and SCI: Facilitation showed strong coefficients and were statistically supported ($p < 0.05$). SCI: Facilitation also showed a strong and supported effect in Scene 3 ($p = 0.022$). We reiterate that these statistically significant findings are exploratory, serving to generate hypotheses, and do not imply causal relationships or definitive predictive power due to the small sample size.

In summary, the multiple regression analysis suggested that later intervention timing was associated with the following factors: having lower perceived workplace stress, beliefs emphasizing children's self-resolution over adult intervention, and a strong sense of initiative.

Table 3. Results of the multiple regression analysis $R^2 = 0.402$ (Scene 1), 0.786 (Scene 2), 0.751 (Scene 3)

Scene	Statistic	AS: Nurturance	AS: Sociability	AS: Initiative	NTSS: Workplace Relationships	SCI: Facilitation	NTSS: Lack of Time	Constant
1	Rc	16.590	-14.847	10.543	-8.403	-4.286	1.569	8.972
	p	0.178	0.321	0.608	0.619	0.737	0.899	0.576
2	Rc	12.915	-5.629	21.399	-49.756	28.770	29.022	-9.204
	p	0.310	0.677	0.244	0.025	0.014	0.048	0.473
3	Rc	-28.325	-17.926	67.763	-41.246	35.658	10.295	7.921
	p	0.127	0.446	0.037	0.054	0.022	0.522	0.711

Note: Rc: Regression coefficient

E. Exploratory Supplementary Analysis: Experience and Intervention Timing

To examine the relationship between early childhood education experience and intervention timing, a scatter plot was created (Fig. 6, top). The horizontal axis represents years of experience in early childhood education, and the vertical axis represents the timing of the intervention; the higher a point is plotted, the later the intervention occurred.

When comparing Scenes 1 to 3, Scene 1 showed an even distribution of points with no clear trend. However, Scenes 2 and 3 suggested clustering patterns. Therefore, k-means++ clustering was applied to divide the sample into three groups for further analysis (Fig. 6, bottom).

The most noticeable feature was the small cluster located in the lower right (C3). This cluster appeared in Scenes 2 and 3 and consisted of three participants with over twenty-five years of experience who intervened significantly earlier than the others.

C1 consisted of relatively inexperienced caregivers, whereas C2 included caregivers with more extensive experience. First, Levene's test suggested a significant difference in the variance of intervention timing between C1 and C2 in Scene 2 ($p = 0.015$). In practice, C2 (the more experienced group) showed smaller variance, indicating a tendency for their intervention timing to be more consistent.

In contrast, for Scene 3, Levene's test yielded a p -value of

0.999, indicating no significant difference in the variance between C1 and C2. A t-test was conducted to compare the means of intervention timing. In Scene 2, no significant difference was found between C1 and C2 ($p = 0.59$). However, in Scene 3, a significant difference was observed ($p < 0.001$), with C2 (the more experienced group) tending to intervene later on average than C1.

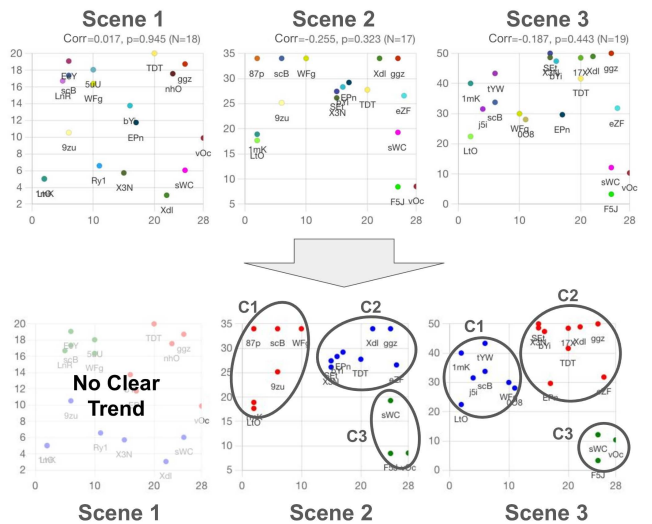


Fig. 6. Relationship between length of early childhood education experience and intervention timing (top), and corresponding trends (bottom).

Spearman's rank correlation coefficients were calculated for the samples included in C1 and C2 (excluding those included in C3). The coefficients were 0.319 ($p = 0.267$) in Scene 2 and 0.472 ($p = 0.065$) in Scene 3, none of which were statistically significant.

Our experiment did not reveal a consistent association between the level of experience and delay in intervention.

To investigate whether the three participants in Cluster C3 provided any common or non-standard responses to the pre-experiment questionnaire, Z-scores were calculated for each subscale response, and the absolute values of the average Z-scores across the three participants were ranked in descending order. The results are summarized in Table 4.

Although definitive conclusions could not be drawn owing to the small sample size ($N = 3$), it was found that all three participants in Cluster C3 had notably low scores in AS: Initiative and AS: Sociability.

Table 4. Z-scores of the pre-experiment questionnaire responses for the three participants in Cluster C3 (top 5 subscales only)

ID	AS: Initiative	AS: Sociability	SCI: Facilitation	NTSS: Lack of Time	SCI: Cessation
sWC	-0.591	-1.559	0.247	0.820	-0.114
vOc	-1.203	-0.474	0.247	-0.397	-0.383
F5J	-1.816	-1.559	-3.858	1.631	-1.461

F. Exploratory Supplementary Analysis: Motion Data (PCA and Behavioral Indices)

To examine inter-individual differences in participants' movement patterns during the VR intervention, Principal Component Analysis (PCA) was applied to motion-tracking data, including head and hand positions, orientations, and gaze direction. Time-series data were aggregated into phase-based feature representations for each participant and projected onto the first two principal components for visualization.

The first two principal components explained between 48.7% and 82.8% of the variance across phases. In the early phases (Phases 0–1), the explained variance was relatively low (approximately 50%–60%), indicating a high degree of variability in movement patterns. In contrast, later phases (Phases 3–4) showed higher explained variance (approximately 70%–80%), suggesting that participants' behaviors became more convergent as the scenes progressed. However, because participants who had already intervened were excluded from later phases, the number of observations decreased, and this reduction may have partially influenced the observed increase in explained variance.

Representative PCA plots are shown in Fig. 7. Visual inspection did not reveal clear cluster structures; instead, participant behaviors appeared to be distributed continuously, with a small number of outliers in several phases.

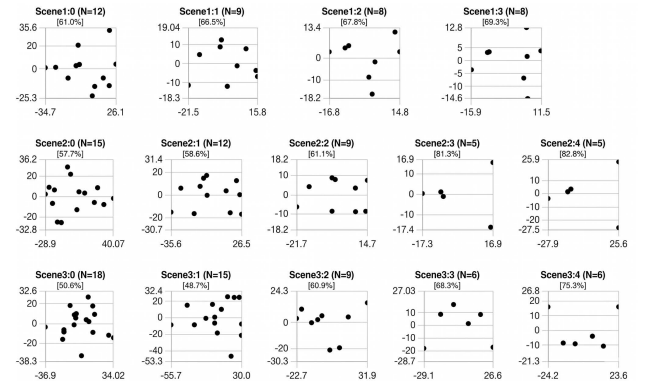


Fig. 7. Two-dimensional plot of participant behaviors.

Across scenes, Scene 3 showed the lowest explained variance in the early phases, indicating greater diversity in initial behavioral responses, whereas Scene 2 exhibited higher explained variance in later phases, suggesting relatively homogeneous behavior in the final stages.

These findings suggest that individual differences in movement behavior are more pronounced in the early stages of the scenes and become more structured as situations escalate. However, given the exploratory nature of this analysis and the limited sample size, these interpretations should be considered tentative.

In addition to the PCA analysis, four behavioral indices were derived from the motion data: mean translational speed, mean horizontal speed, mean vertical speed, and mean head rotation speed. These indices were calculated for each participant and used to examine associations with intervention timing and psychological variables.

Regarding intervention timing, moderate negative descriptive trends were observed between intervention timing in Scene 2 and movement speed indices ($r \approx -0.43$ to -0.44), suggesting that participants who moved more slowly tended to intervene later ($p \approx 0.08$ – 0.09), as illustrated in Table 5. No meaningful association was observed between head rotation speed and intervention timing ($|r| < 0.11$).

Correlations with pre-experiment questionnaire scores revealed limited associations overall; however, a moderate and statistically significant correlation was found between SCI: Facilitation and mean head rotation speed ($r = 0.43$, $p = 0.02$). This finding suggests that participants who place greater value on children's self-resolution tend to engage in more active visual exploration of the environment. Given the

exploratory design and small sample size, this result should be interpreted as hypothesis-generating rather than confirmatory.

Table 5. Intervention timing and movement speed

Intervention timing Scene	Mean speed	Mean horizontal speed	Mean vertical speed	Mean head rotation speed
Scene 1	$r = 0.32$ $p = 0.21$	$r = 0.33$ $p = 0.20$	$r = 0.24$ $p = 0.35$	$r = -0.07$ $p = 0.78$
Scene 2	$r = -0.44$ $p = 0.09$	$r = -0.43$ $p = 0.10$	$r = -0.44$ $p = 0.08$	$r = -0.10$ $p = 0.72$
Scene 3	$r = -0.36$ $p = 0.14$	$r = -0.40$ $p = 0.10$	$r = -0.33$ $p = 0.18$	$r = 0.03$ $p = 0.92$

G. Supplementary Finding: Self-Reported Reasons for Intervention

Fig. 8 shows the results of the post-experiment questionnaire on reasons for intervention based on Triandis's behavioral model. The overwhelming majority responded that they intervened "because this is how I always act when such a situation occurs", indicating that most participants acted primarily out of habit. This is likely because the scenes were intentionally designed to replicate common situations that occur frequently in childcare settings.

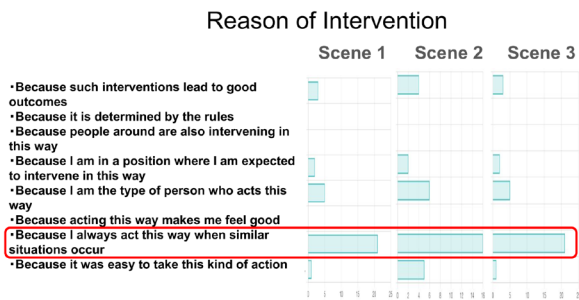


Fig. 8. Reasons for intervention.

H. Summary of Results

The exploratory VR experiment ($N = 29$) yielded three main findings regarding caregiver intervention behavior under standardized conditions:

Individual Differences (RQ1): The VR environment successfully captured substantial individual variability in intervention timing. Strong correlations between the multi-child scenes (Scenes 2 and 3) suggested consistent individual response patterns

Psychological Associations (RQ2): Factors associated with later intervention (delayed timing) included lower perceived workplace stress (NTSS: Workplace Relationships) and the pedagogical belief in child self-resolution (SCI: Facilitation). These results serve as hypothesis-generating findings.

Belief-Behavior Link: A statistically significant association ($r = 0.43$, $p = 0.02$) was found between the belief in child self-resolution (SCI: Facilitation) and active visual exploration (head rotation speed). This suggests the VR system can capture the active observation process critical for high-quality, non-immediate intervention. Additionally, self-reported reasons for intervention indicated that most participants acted out of habitual response patterns, consistent with the scenes' design to replicate common real-world childcare situations.

I. Impressions Regarding the Virtual Nursery Experiment

After completing the experiment, participants were asked to freely write their impressions of the virtual nursery experiment. Their responses were grouped into the following categories.

1) Prior knowledge about the children

- I feel that decisions about whether to intervene during childcare depend on how well the caregiver knows the child. Even if children behave in the same way, some can be monitored, whereas others—such as those with a history of biting—require immediate intervention. Therefore, it was difficult to determine how to intervene for children in virtual settings.
- It was difficult. There was little information about the children's personalities, which made it challenging to anticipate their behavior. It might feel more realistic if information such as age or tendencies were provided, similar to what we receive before a child enters the center. It was also challenging because the children were in an age group that I do not usually work with.

2) The realism of the CG

- Since the children's movements were not realistic, they were quite different from those in usual childcare settings, making it difficult to respond accordingly.
- The environment differed from that of actual childcare, and because the children were not real, I felt unsure how to respond.
- The children had almost no facial expressions; therefore, I could not tell what they were thinking. I usually read children's feelings from their facial cues, and since there were no expressive signals from them—such as "I want that"—it was hard to know when to speak up.

3) Confusion related to the VR environment and equipment

- Because it was my first experience with VR, it was difficult to respond immediately or speak to the children in sync with their movements.
- I missed the chance to speak because it was difficult to understand the situation. In the virtual environment, it was difficult to understand children's ages or characteristics. I wish I had asked whether the child wearing the same clothes across all scenes represented the same character.
- I understood that this was a limitation of viewing the environment on a screen, but the field of view felt narrower than in the real world, and I found it difficult to see the entire space at once.

4) Regarding freedom of movement and action

- I think having more options than are currently available would allow for a wider range of responses. For example, it would have been better if we could pick up objects that had been thrown.
- I felt it would be more realistic if there were more toys and other objects in the VR space.

J. Exploratory Analysis on Compensation Bias

To address the potential bias introduced by the unequal compensation structure, we conducted an exploratory analysis comparing the means of intervention timing and key psychological variables between the two groups (Table 6).

The results of the t -test (and Welch's test) showed no

statistically significant difference between the compensated and non-compensated groups for any of the tested variables ($p > 0.50$ for all comparisons). Although the compensation structure represents a methodological constraint, this exploratory analysis suggests that, within the current small sample, the absence of compensation did not introduce a measurable bias in the primary behavioral metric (intervention timing) or in the key associated psychological constructs (NTSS: Workplace Relationships, SCI: Facilitation, AS: Initiative). We acknowledge that the lack of observed difference may be due to the limited statistical power of the small sample.

Table 6. Comparison of key variables by compensation group

Variable	With Compensation (M±SD)	Without Compensation (M±SD)	t-test p	Welch p
Scene 2 Intervention timing	25.957 ± 9.264	24.972 ± 7.793	0.81509	0.81510
Scene 3 intervention timing	32.790 ± 16.485	35.168 ± 13.685	0.75117	0.75337
NTSS:Workplace Relationships	2.871 ± 0.910	2.733 ± 0.781	0.66550	0.66567
SCI: Facilitation	3.357 ± 0.807	3.517 ± 0.578	0.54818	0.54894
AS: Initiative	3.321 ± 1.085	3.400 ± 0.573	0.81111	0.81182

V. DISCUSSION

A. Methodological Contribution

This study primarily contributes a novel VR-based methodological framework that enables the observation of caregiver behavior under fully standardized conditions—an approach that is difficult to achieve in real childcare environments. By isolating behavioral timing from contextual variability, the proposed method complements traditional questionnaires and field observations.

B. Conceptual Implications: Intervention Timing as a Process

Exploratory analyses of the observed intervention timing suggested that several caregiver-related factors may be associated with how caregivers respond to conflict situations in the virtual nursery.

Multiple regression analysis was conducted to examine how these factors interacted when considered simultaneously. The results indicate potential associations between intervention timing and factors such as NTSS: Workplace Relationships, SCI: Facilitation, and AS: Initiative, suggesting these variables merit further investigation in future research.

NTSS: Workplace Relationships assesses whether caregivers feel stress related to five specific aspects: “relationships between other teachers”, “the caregiver’s relationship with other teachers”, “the attitudes and actions of surrounding teachers”, “difficulty in uniting emotionally with fellow staff”, and “disagreements with other teachers”. The consistent descriptive trend suggests that caregivers who experience greater stress from interpersonal relationships with colleagues tend to intervene earlier in children’s conflicts. This offers a noteworthy insight into how environmental factors shape decision-making, especially in Japanese childcare settings, where team-based caregiving is the norm. The inclusion of NTSS: Workplace Relationships in the multiple regression model is theoretically justified

despite its moderate individual correlation: its observed effect size is highly relevant because it represents a critical contextual factor—team-level interpersonal stress—that is likely to interact with individual beliefs (SCI: Facilitation, AS: Initiative) to shape behavior, thereby reflecting the complex ecological nature of caregiving practice.

SCI: Facilitation is a subscale that assesses whether a teacher believes it is more important for children to resolve conflicts themselves rather than having the teacher intervene and solve the issue for them. Expecting children to resolve conflict on their own naturally requires a period of observation to assess whether self-resolution is possible, leading to delayed intervention. The correlation between higher SCI: Facilitation scores and later intervention thus aligns with the core pedagogical approach of the Japanese National Curriculum Standard for Nurseries (Hoikujo Hoiku Shishin) [4], which emphasizes respecting and fostering children’s self-resolution. This emphasis on facilitation as a critical component of high process quality in the Japanese cultural context is supported by cross-cultural studies that illustrate the diverse approaches different nations take in raising children [21]. This reliance on careful observation before intervention is particularly vital in the complex task of caring for one-year-olds [5], who are the subject of this study, and is central to ongoing discussions regarding quality assessment methodologies in Japan [6]. Furthermore, the finding that AS:Initiative (proactive self-perception) was positively correlated with later intervention may reflect confidence and self-efficacy in one’s capacity to tolerate ambiguity and wait for the optimal moment for educational scaffolding rather than reacting impulsively. This variable, along with SCI:Facilitation, thus captures a critical psychological foundation for the ‘professional discretion’ necessary for high-quality, non-immediate intervention.

Critically, intervention timing should not be interpreted merely as a behavioral outcome (when the caregiver acted) but as a complex process-level pedagogical construct that encapsulates the antecedent judgment and observation skill. This argument is supported by the finding of a statistically significant correlation between SCI:Facilitation and mean head rotation speed ($r = 0.43$, $p = 0.02$). The alignment between the belief in child self-resolution (SCI:Facilitation) and increased active visual exploration (head rotation) suggests that expert discretion—the process of assessing whether self-resolution is possible—may be quantitatively captured using kinematic data. This finding suggests that the VR system captures not only the result of decision-making (timing) but also the process of information gathering (active observation). Caregivers who value facilitation tend to gather more visual information before acting, suggesting the potential for using VR motion data to assess pedagogical attitudes. This measured active visual exploration explicitly links the motion analysis findings to established theories of professional judgment and observation, such as Schön’s “*The Reflective Practitioner*” [30] and van Manen’s “*The Tact of Teaching*” [31]. The VR system thus provides a foundational tool for objectively quantifying the “thinking in action” or “pedagogical thoughtfulness” that these theories emphasize, offering a novel empirical bridge between psychological constructs and observable professional practice.

C. Practical Reflections and Future Directions

Cluster analysis, along with variance or rank correlation analyses, examining the relationship between the length of early childhood education experience and intervention timing did not suggest a stable or consistent association between experience and early intervention. However, in some cases, only some results approached significance; therefore, it cannot be concluded that there is no relationship. According to the intuition of practicing nursery teachers, increased experience is often believed to enhance observational skills, allowing caregivers to intervene only when necessary. Therefore, it is necessary for future research to examine the conditions under which this hypothesis holds true—or whether it fails to do so.

The study's exploratory findings indicate practical utility for identifying specific individual differences in intervention patterns. The analysis suggested a small group of highly experienced veterans who intervened very early (Cluster C3). These three participants shared a common trait: low scores on both AS: Initiative and AS: Sociability. AS: Initiative showed the highest contribution in Scene 3 in the multiple regression analysis. In contrast, AS: Sociability showed low contribution rates in Scenes 2 and 3 and did not yield statistically significant results. This pattern suggests that AS: Sociability may characterize a specific subgroup, namely the three participants who intervened particularly early. The AS: Sociability subscale consisted of two items: "Can interact with others casually" and "Can casually deal with anyone". A low score indicates difficulty or discomfort in interacting with others. Among experienced nursery teachers, those who believed that they had low behavioral proactivity and found it difficult to interact casually with others tended to intervene earlier. This pattern suggests a potential risk of "defensive caregiving"—prioritizing conflict avoidance over child development—where accumulated experience combines with low behavioral proactivity or interpersonal comfort. Teachers with interpersonal anxiety or avoidant tendencies may become increasingly prone to early intervention as they spend more time in the field. The VR system indicated the potential to identify specific behavioral patterns that are difficult to detect through conventional observation. However, even if this is true, no caregivers currently in the process of such intensification were identified in this study. This is another aspect that future research should explore in greater detail.

Furthermore, the VR system can be a tool to be validated in future work for reflective practice and professional training. Specifically, the system can allow caregivers to compare their quantified behaviors (timing, observation intensity, movement) against established quality metrics derived from the Japanese National Curriculum Standard for Nurseries (Hoikujo Hoiku Shishin) [4].

The exploratory findings illustrate how specific caregiver factors are associated with intervention timing, reflecting key aspects of professional performance. Specifically, the correlation between a belief in child self-resolution (SCI:Facilitation) and later, more actively observed intervention (measured by head rotation speed) suggests that the VR system may offer the potential to quantify a critical component of process quality [21]. This deliberate delay in intervention may reflect the sophisticated professional skill

required, particularly in the complex task of caring for one-year-olds [5], who are the focus of this study. This pattern directly aligns with the core pedagogical goals of the Japanese National Curriculum Standard for Nurseries (Hoikujo Hoiku Shishin) [4], which emphasizes fostering children's self-resolution and initiative. The capacity of the VR tool to capture this behavior may contribute to informing the ongoing debates regarding quality assessment methodologies [6] by providing objective, quantitative metrics for practices considered high quality in this specific cultural and pedagogical context.

Based on these findings and limitations, future work should be strategically focused on three key directions to promote and enhance knowledge about adult performance and its impact on children's well-being, development, and learning:

VR-Supported Professional Development: The VR system should be refined into a validated tool for reflective practice and professional training. Future efforts should focus on allowing caregivers to compare their quantified behaviors (timing, observation intensity, movement) with established quality metrics derived from the National Curriculum Standard [4]. This can help promote consensus among staff and directly impact practice quality.

Expansion of Behavioral Metrics: Future studies must incorporate additional motion data, such as gaze direction and detailed body positioning, into analyses. Integrating these data will enable a deeper understanding of the skills profile of professionals who care for younger children and clarify the causal pathways between observed behavior and psychological profiles (e.g., self-efficacy).

Large-Scale Validation: Further large-scale validation studies, ideally cross-cultural ones, are required to establish the stability and generalizability of the identified trends using more diverse and representative samples (addressing the current overrepresentation of experienced staff) and to confirm the existence of specific behavioral patterns, such as the potential "defensive caregiving" tendency observed in the early-intervening veteran subgroup.

D. Limitations

As an exploratory preliminary study, the current work is subject to several important limitations. These constraints primarily stem from the trade-off inherent in using standardized VR environments for behavioral measurement and must be acknowledged when interpreting the findings.

1) Methodological constraints

Sample Size and Generalizability: The study utilized a limited sample size ($N = 29$), which included an overrepresentation of experienced staff. This inherent sampling bias restricts the generalizability of our findings to the broader population of caregivers, particularly those with less experience.

Statistical Interpretation and Stability: Given the small and non-representative sample, all statistical findings should be regarded as purely hypothesis-generating. Consequently, the observed exploratory statistical significance ($p < 0.05$) does not imply stability, generalizability, or causality across different caregiver populations. Further large-scale validation studies are necessary to establish the stability of these trends.

Scene Reliability (Scene 1): Scene 1 showed low relevance

and reliability compared to the multi-child scenes (Scenes 2 and 3). This is likely attributable to its single-child design and the high number of participants who intervened before the critical event (Fig. 5). From a developmental perspective, one-year-olds exhibit highly diverse and unpredictable exploratory behaviors when playing alone, such as the “Tupperware walking” depicted here. Without the clear social cues or urgent risks inherent in peer conflicts (as seen in Scenes 2 and 3), caregivers’ intervention thresholds may have fluctuated more significantly based on their individual tolerance for “risky play” rather than standardized pedagogical timing. Therefore, future refinements should consider how the absence of social interaction impacts the stability of behavioral metrics in VR.

Limited Behavioral Metrics: The primary analysis focused solely on intervention timing. While other motion data (position, gaze direction, physical movements) were recorded, they were not fully incorporated into detailed behavioral analyses. Future research must utilize these additional data points to capture more nuanced, real-world behaviors and critically enhance the ecological validity of the findings.

Compensation Bias: The unequal compensation provided to participants, driven by institutional constraints, represents a key methodological constraint. Although an exploratory analysis indicated no statistically significant difference in primary outcomes between compensated and non-compensated groups, the potential for unmeasured systemic bias correlating with institutional type must be acknowledged.

2) Technical constraints (VR environment)

Representational Fidelity (CG Realism): The ecological validity was constrained by the VR environment’s technical limitations, particularly the lack of children’s facial expressions and limited CG realism. Participants noted the absence of crucial non-verbal cues (e.g., emotional states) that experienced caregivers rely on to assess children and inform their decision-making in real settings, potentially altering the intervention process compared to the field.

Interactional Constraints: The operational definition of ‘intervention’ was limited exclusively to a direct ‘touch’. This restricted the measurement of typical non-intervention supportive behaviors (e.g., verbal reassurance, fetching a toy, or moving closer without touching), thereby limiting the generalizability of the recorded timing data to the full diversity of real-world practice.

3) Contextual and cultural constraints

Standardization Trade-Off (Lack of Contextual Information): The methodological necessity of standardization led us to intentionally withhold essential prior information about the children’s personality and history. As noted by participants, this deliberate removal of contextual information significantly lowered the ecological validity by eliminating the foundation required for nuanced professional judgment.

Cultural Scope: The experiment was conducted exclusively in Japan. The observed caregiver behavior, such as the emphasis on fostering children’s self-resolution, may be influenced by the specific Japanese cultural and educational context (e.g., the Hoikujo Hoiku Shishin [4]). This limits the cross-cultural generalizability of the findings.

VI. CONCLUSIONS

The primary and fundamental contribution of this work is methodological innovation, not empirical generalization. We demonstrated the feasibility and validity of a VR-based virtual nursery as an exploratory methodological tool for observing caregiver intervention behaviors under fully standardized conditions. The system successfully captured significant individual differences in intervention timing, thereby establishing this system as a foundational tool for objective behavioral measurement in early childhood education. In addition, exploratory analyses indicated that caregiver profiles — particularly perceived workplace stress and beliefs about children’s self-resolution — were descriptively associated with variation in intervention timing, suggesting that the framework can also surface psychological factors linked to observed behavior.

The capacity for standardized observation uniquely enables critical future research. By overcoming the reproducibility limits of field observation, this methodology serves as a robust platform for:

- **Longitudinal Studies:** Quantitatively tracking the development of professional judgment and observation skills over a caregiver’s career.
- **Cross-Cultural Validation:** Systematically comparing intervention practices across different national contexts, such as the Japanese context and others, free from contextual noise.
- **Intervention-Based Research:** Providing objective pre- and post-intervention behavioral metrics for evaluating the efficacy of professional development and training programs, including VR-supported interventions.

This framework directly addresses the need for objective, comparable data, thereby clearly articulating the study’s primary contribution as a research enabler.

APPENDIX

English translation of questionnaires

A. AS: Childcare Worker Aptitude Scale [17] (Reduced Items)

- [Altruism] Always acts with others in mind
- [Altruism] Always tries to do things that make others happy
- [Empathy] Often feels happy or sad because of others’ experiences
- [Empathy] Frequently feels happy or sad by putting oneself in others’ shoes
- [Logical Thinking] Tries to understand things more deeply
- [Logical Thinking] Makes an effort to understand what has not yet been understood
- [Tactfulness] Understands people’s feelings
- [Tactfulness] Can sense emotions even from casual words
- [Sociability] Can interact with others casually
- [Sociability] Can deal casually with anyone
- [Initiative] Tends to act proactively
- [Initiative] Can act immediately in critical situations
- [Nurturance] Makes every effort for children without sparing labor
- [Nurturance] Makes an effort even when things are difficult for the sake of children

B. NTSS: Nursery Teacher's Stress Scale [18] (Reduced Items)

[Stress from Dealing with and Understanding Children] Cannot handle certain children well
 [Stress from Dealing with and Understanding Children] Do not know what kind of support children need
 [Stress from Dealing with and Understanding Children] Lack of knowledge
 [Stress from Dealing with and Understanding Children] Cannot understand children's feelings
 [Stress from Dealing with and Understanding Children] Cannot grasp the characteristics of children
 [Stress from Workplace Relationships] Relationships with other teachers
 [Stress from Workplace Relationships] My relationships with other teachers
 [Stress from Workplace Relationships] Attitudes and behaviors of surrounding teachers
 [Stress from Workplace Relationships] Teachers are not united when facing issues
 [Stress from Workplace Relationships] Disagreements with other teachers
 [Stress from Dealing with Parents] Differences in values between caregivers and parents.
 [Stress from Dealing with Parents] Differences in understanding of children between parents and caregivers
 [Stress from Dealing with Parents] My intentions as a caregiver are not communicated to parents
 [Stress from Dealing with Parents] Parents are uncooperative toward the nursery school
 [Stress from Lack of Time] Lack of private time
 [Stress from Lack of Time] Difficulty taking time off
 [Stress from Misalignment with Nursery Policies] Discrepancy between nursery policies and my own ideas
 [Stress from Misalignment with Nursery Policies] Some nursery rules are unacceptable to me

C. Post-experiment Questions Regarding Triandis's Behavior Model

Which of the following best explains whether you chose to intervene?

Because such interventions lead to good outcomes
 Because it is determined by the rules
 Because people around are also intervening in this way
 Because I am in a position where I am expected to intervene in this way
 Because I am the type of person who acts this way
 Because acting this way makes me feel good
 Because I always act this way when similar situations occur
 Because it was easy to take this kind of action

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

SO conducted the research; KS implemented the VR system; SK and HA designed and analyzed the field experiments; MN, YT, TS, MT, and MI were research supervisors who guided and oriented the entire research project; all authors have approved the final version.

FUNDING

This research was supported by the following grant: 2022 Academic Year Doshisha University Organization for Research and Development of Innovative Science and Technology, Collaborative Research Project at the Baby Science Research Center, "Prototype Development of a Childcare Coaching System Using VR." This research was also supported by JST RISTEX Grant Number JPMJRS25M2, Japan.

ACKNOWLEDGEMENT

We would like to express our sincere gratitude to the staff at Mifune-no-sato Nursery School, Okubo Wakakusa Kodomo-en, Seiai Nursery School, Seiai Kindergarten, Seiai Cross Nursery School at Sumiyoshi/Miyayama-cho, Kuraki Nagata Nursery School, and Oguno Nursery School for their participation in the experiment and for generously providing the experimental settings. We are deeply grateful to Prof. Dora Chen for kindly granting us permission to use the translated version of the SCI. We also extend our heartfelt thanks to Dr. Sachiko Nozawa of the University of Tokyo for her support in scene development. For their significant contributions to the creation of the 3D content, we would like to thank 3D Designer Shojiro Sakamoto and Actor Kuniharu Harada.

REFERENCES

- [1] OECD, *Starting Strong IV: Monitoring Quality in Early Childhood Education and Care*, 2015. doi: 10.1787/9789264233515-en
- [2] E. S. Peisner-Feinberg, M. R. Burchinal, R. M. Clifford *et al.*, "The relation of preschool child-care quality to children's cognitive and social developmental trajectories through second grade," *Child Development*, vol. 72, no. 5, pp. 1534–1553, 2001. doi: 10.1111/1467-8624.00364
- [3] J. J. Heckman, "Skill formation and the economics of investing in disadvantaged children," *Science*, vol. 312, no. 5782, pp. 1900–1902, 2006. doi: 10.1126/science.1128898
- [4] *National Curriculum Standard for Nurseries*, Ministry of Health, Labour and Welfare, Japan, 2017.
- [5] M. Koga, "What is the difficulty of childcare for one-year-Olds?" *Journal of Childcare Studies*, vol. 49, no. 3, pp. 248–259, 2011. doi: 10.20617/reccej.49.3_248 (in Japanese)
- [6] N. Horikoshi. (2024). Research on process quality in early childhood education: Drafting the quality assessment scale for early childhood education practices. *Bulletin of the National Institute for Educational Policy Research*. [Online]. 153. Available: https://www.nier.go.jp/kankou_kiyou/153/b03.pdf
- [7] M. Rooijen and G. Jacobs, "A professionalisation programme towards children's risk-taking in play in childcare contexts: Moral friction on developing attitudes and collegial expectations," *European Early Childhood Education Research Journal*, vol. 27, pp. 741–756, 2019. doi: 10.1080/1350293X.2019.1678715
- [8] D. W. Chen, "Preventing violence by promoting the development of competent conflict resolution skills: Exploring roles and responsibilities," *Early Childhood Education Journal*, vol. 30, no. 4, pp. 203–208, June 2003. doi: 10.1023/A:1023379306124
- [9] J. Piaget, *The Moral Judgment of the Child*, Routledge, 1932. doi: 10.4324/9781315009681
- [10] M. Majorano, P. Corsano, and G. Triffoni, "Educators' intervention, communication and peers' conflict in nurseries," *Child Care in Practice*, vol. 21, no. 2, pp. 98–113, 2015. doi: 10.1080/13575279.2014.1001812
- [11] R. I. Skoglund, "When 'Words Do Not Work': Intervening in children's conflicts in kindergarten," *International Research in Early Childhood Education*, vol. 9, no. 1, pp. 23–38, 2019.
- [12] L. Gloeckler and J. Cassell, "Teacher practices with toddlers during social problem solving opportunities," *Early Childhood Education Journal*, vol. 40, no. 4, pp. 251–257, 2012. doi: 10.1007/s10643-011-0495-4
- [13] S. T. Williams, A. M. Mastergeorge, and L. L. Ontai, "Caregiver involvement in infant peer interactions: Scaffolding in a social

- context,” *Early Child. Res. Q.*, vol. 25, no. 2, pp. 251–266, 2010. doi: 10.1016/j.ecresq.2009.11.004
- [14] M. Sims, T. Hutchins, and M. Taylor, “Young children in child care: The role adults play in managing their conflicts,” *Early Child Development and Care*, vol. 124, pp. 1–9, 1996. doi: 10.1080/0300443961240101
- [15] K. Heland-Kurzak, “‘Who I Am at Work’: Sources and determinants of childcare workers’ professional work,” *Elementary Education in Theory and Practice*, vol. 19, no. 2, 73, pp. 177–194, 2024. doi: 10.35765/eetp.2024.1973.13
- [16] C. L. Smart, “The dilemma of response among early childhood educators to rough-and tumble play in educational settings: moving beyond expectations and rediscovering play,” Ph.D. dissertation, University of New Hampshire, 2013.
- [17] K. Fujimura, “Development of a childcare worker aptitude scale for students aspiring to become nursery school or kindergarten teachers,” *Bulletin of Human Sciences, Osaka Shoin Women's University*, vol. 9, pp. 129–143, Jan. 2010. (in Japanese)
- [18] T. Akada, “Development of the nursery teacher’s stress scale and examination of its reliability and validity,” *Shinrigaku Kenkyu*, vol. 81, no. 2, pp. 158–166, June 2010. doi: 10.4992/jjpsy.81.158
- [19] D. W. Chen and K. E. Smith, “The Social Conflict Inventory (SCI): A measure of beliefs about classroom peer conflicts,” *Journal of Early Childhood Teacher Education*, vol. 23, no. 4, pp. 299–313, Jan. 2002. doi: 10.1080/1090102020230402
- [20] U. Bronfenbrenner, “Ecology of the family as a context for human development: Research perspectives,” *Developmental Psychology*, vol. 22, no. 6, pp. 723–742, 1986. doi: 10.1037/0012-1649.22.6.723
- [21] J. J. Tobin, D. Y. Wu, and D. H. Davidson, *Preschool in Three Cultures: Japan, China and the United States*, Yale University Press, 1989.
- [22] B. Curbow, K. Spratt, A. Ungaretti *et al.*, “Development of the child care worker job stress inventory,” *Early Childhood Research Quarterly*, vol. 15, no. 4, pp. 515–536, 2000. doi: 10.1016/S0885-2006(01)00068-0
- [23] K. M. La Paro, R. C. Pianta, and M. Stuhlman, “The classroom assessment scoring system: Findings from the prekindergarten year,” *The Elementary School Journal*, vol. 104, no. 5, pp. 409–426, 2004. doi: 10.1086/499760
- [24] Y. J. Katz, “The use of virtual reality three-dimensional simulation technology in nursery school teacher training for the understanding of children’s cognitive perceptions,” in *Learning in School, Home and Community: ICT for Early and Elementary Education*, G. Marshall and Y. Katz, Eds. Springer, 2003, pp. 41–51. doi: 10.1007/978-0-387-35668-6_5
- [25] D. Passig and T. Noyman, “Training kindergarten teachers with virtual reality,” in *Proc. Networking the Learner: Computers in Education, IFIP TC3 Seventh IFIP World Conference on Computers in Education, WCCE 2001*, 2001, pp. 733–744. doi: 10.1007/978-0-387-35596-2_74
- [26] I. Radu, C. Dede, M. R. Seyam *et al.*, “Using 360-video virtual reality to influence caregiver emotions and behaviors for childhood literacy,” *International Journal of Gaming and Computer-Mediated Simulations*, vol. 13, no. 1, pp. 12–33, 2021. doi: 10.4018/IJGCMS.20210101.oa2
- [27] Jollygood. (April 2026). *Childcare Education VR*. [Online]. Available: https://jollygood.co.jp/elementortest_komatsu/
- [28] Y. J. Katz, “Kindergarten teacher training through virtual reality three-dimensional simulation methodology,” *EMI Educ. Media Int.*, vol. 36, no. 2, pp. 151–156, 1999. doi: 10.1080/0952398990360211
- [29] H. C. Triandis, *Interpersonal Behavior*, Brooks/Cole Publishing Company, 1977.
- [30] D. Schön, *The Reflective Practitioner: How Professionals Think in Action*, Basic Books, 1983.
- [31] M. V. Manen, *The Tact of Teaching: The Meaning of Pedagogical Thoughtfulness*, State Univ of New York Press, 1991.

Copyright © 2026 by the authors. This is an open access article distributed under the Creative Commons Attribution License which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited ([CC BY 4.0](https://creativecommons.org/licenses/by/4.0/)).