MASTERARBEIT | MASTER'S THESIS

Titel | Title

Assessing Children's Empathic Responding in Virtual Reality
A Mixed-Methods Pre-Post Study Following Mental Health Literacy Training

verfasst von | submitted by

Iulia Pokorny BSc

angestrebter akademischer Grad | in partial fulfilment of the requirements for the degree of

Master of Science (MSc)

Wien | Vienna, 2026

Studienkennzahl lt. Studienblatt | Degree programme code as it appears on the student record sheet:	UA 066 840
Studienrichtung lt. Studienblatt | Degree programme as it appears on the student record sheet:	Masterstudium Psychologie
Betreut von | Supervisor:	Univ.-Prof. Giorgia Silani Privatdoz. PhD

Table of Contents

Introduction

Empathy can be described as the ability to share and understand other people's thoughts and feelings. It is often linked to positive behavior towards others and facilitates social interactions and interpersonal relationships (Allemand et al., 2015; Hoffman, 2000). For primary school children empathy is important in everyday life since many of their social experiences occur in school and in contact with peers. This often involves cooperation and conflict, in which empathic skills are needed (Barreda-Ángeles et al., 2021; Richaud et al., 2017). Empathy is a central component of moral as well as social competence development and is associated with better social skills, more supportive relationships, helpfulness towards others and less aggression (Allemand et al., 2015; Bayne et al., 2024; Sesso et al., 2021). Furthermore it is also associated with better problem-solving skills, school readiness and academic success, suggesting that empathy has both individual and interpersonal benefits (Imece & Arslan Cansever, 2019; McCreary & Marchant, 2017; Özdemir Beceren & Özdemir, 2019)

Empathy is also discussed in the context of mental health and prevention. Early adolescence is described as a vulnerable phase of life in which many mental disorders begin. At the same time schools provide a key setting for prevention targeting socio-emotional competencies such as empathy (Patel et al., 2007). It is therefore recommended that children develop positive concepts of mental health at an early stage, rather than intervening only when mental disorders have already emerged later in life (Bale et al., 2020). A central concept in this context is mental health literacy (MHL), which refers to knowledge and beliefs about mental health that facilitate the recognition, management, and prevention of mental health problems (Jorm, 2000). Bale et al. (2020) recommend that MHL programs for children should include content on understanding of their own and others' emotions, identifying psychological changes, and knowing about available support options.

Measuring empathy in children is methodologically challenging. Empathy is multidimensional and commonly assessed using questionnaires, whereby self-report and other-report may be influenced by social desirability, interpretation, context and children's development (Bayne et al., 2024; Neary, 2023; Richaud et al., 2017; Sesso et al., 2021). Virtual reality (VR) is a methodological tool that can offer experimental control by placing individuals in standardized, interactive and vivid social situations (Parsons et al., 2017; Parsons & Phillips, 2016). In addition, correspondence approaches that derive empathy from whether children report the same emotion as a target may combine various components of empathic responding, such as emotion identification and affect sharing (Coll et al., 2017).

Against this background the present study uses a VR-based stress paradigm in a classroom setting to operationalize empathy via children's emotion ratings across self- and other-related contexts. The VR scenario is structured as a social-evaluative task context, providing a standardized situation in which children do a stress-inducing task and observe a non-player character (NPC) facing the same challenge. This approach allows for interactive assessment of emotional responses while maintaining controlled task conditions. The study therefore examines whether an MHL program is associated with pre-post changes in VR-based indicators of empathy-related responding compared to a control group, and whether children's open-ended reports about their own emotions and those of an NPC support the plausibility of the VR paradigm and the interpretation of the quantitative indicators.

Theoretical Background and Hypotheses

Empathy

In research empathy is described as a multidimensional construct for which there is no universally accepted definition. This diversity of definitions has a direct impact on the processes that are measured and how findings are interpreted (Cuff et al., 2016; Singer, 2006). Depending on how the concept is used, empathy can be understood in very broad terms, including aspects such as emotional contagion, personal distress, sympathy or cognitive perspective-taking. At the same time there are more narrow definitions that define empathy as an affective state, that is elicited by and isomorphic to another's person state and includes awareness about the origin of the state (de Vignemont & Singer, 2006). Coll et al. (2017) refer to this definition as the standard definition of empathy.

A common approach is to distinguish between affective and cognitive empathy. Affective empathy refers to the ability to experience an emotion aroused by the emotion of others while cognitive empathy describes to understand what others are thinking or feeling. Although both components interact with each other, they are based on different processes and can change or be affected separately from one another (Cuff et al., 2016; Wu et al., 2024).

Like many other researchers, Coll et al. (2017) also distinguish between affective and cognitive components and emphasize the importance of measuring them separately, but they summarize both as processes of empathy that lead to an empathic response. More precisely, the terms affect sharing and emotion identification are used. They argue that under the standard definition of empathy it is necessary to correctly identify the state of another's person. This leads to the problem that even if there is an emotional response that is similar to a differently perceived emotion of the other person, this would not be defined as empathy.

Therefore, examining at these components separately and assessing how accurately someone identifies another's emotional state and to what extent that identification leads to an affective response allows us to understand individual differences in empathic responses (Coll et al., 2017).

In this framework emotion identification can rely on perceptual cues but also involves identifying a person's state from contextual information or through inferential reasoning. Importantly, individuals can vary in their ability to perceive, recognize and categorize the target's emotion. The similarity between the empathizer's evaluation of the target's state and its real emotional state is therefore called emotion identification accuracy (Coll et al., 2017), which closely relates to the notion of empathic accuracy introduced by Ickes et al. (1990). Affect sharing refers to the mechanism in which identifying another's emotional state leads to a similar affective state in oneself. Individual differences in affect sharing can be conceptualized as differences in the translation during this process, with greater accuracy reflected in a closer correspondence between the identified state and the empathizer's resulting state (Coll et al., 2017). The empathic response can therefore be the same for two individuals, even though they may not show the same amount of affect sharing. This explains why correspondence measures of matching should not be interpreted as empathy without considering both the accuracy of affect sharing and emotion identification accuracy (Coll et al., 2017).

Although Coll et al. (2017) already acknowledge that emotion identification can be based on perceptual cues as well as contextual information and interference, De Vignemont and Singer (2006) more explicitly specify which contextual factors modulate empathic responding. They focus on four main categories: the characteristics of the shared emotion, the relationship between empathizer and target, the characteristics of the empathizer and the situational context.

Regarding the target emotion, features such as valence, salience and intensity can influence how strong an empathic response is, and empathizing may be easier for primary emotions (de Vignemont & Singer, 2006). Relatedly, Ickes (2016) states that targets also differ in how "readable" their thoughts and feelings are, which can influence empathic accuracy.

Empathic responding can vary depending on the affective connection between two individuals and factors such as similarity, familiarity and the target's need for care or protection (Batson et al., 2005; Cialdini et al., 1997; Krebs, 1975; Singer et al., 2006).

Gender can be considered as part of the modulatory factors. Girls in late childhood show higher levels of affective empathy and more prosocial tendences than boys (Allemand et al., 2015; Neary, 2023; Wu et al., 2024). Furthermore, girls also seem to exhibit stronger emotional contagion and a higher sensitivity to other's emotions, suggesting that gender differences may not only concern deliberate social responding but also more automatic forms of affective resonance (Bensalah et al., 2016). Notably, research finds no clear or consistent gender difference in cognitive empathy, indicating that the female advantage may be more specific to affective components rather than to cognitive ones (Bensalah et al., 2016; Wu et al., 2024). Empathic accuracy research suggests that women are usually not more accurate than men. When gender differences emerge, they appear to reflect differences in motivation rather than differences in empathic ability (Ickes, 2016).

Another characteristic of the empathizer is their past experience. The more closely a target's state resembles something the empathizer has experienced before, the more likely and accurate affective resonance becomes (Preston, 2007). It is also argued that lacking a relevant experience can limit the empathic response (De Vignemont & Singer, 2006). This contextual view underlines once again that empathic responses can be selective and variable rather than constant across persons and situations, even when the same emotional cues are observed.

Situational context can hinder empathizing when emotions appear unjustified or ambiguous, or when an empathizer is confronted with multiple targets expressing different emotions. Some of these effects may operate via shifts in attention. Empathic accuracy research likewise suggest that paying more attention to the target is a predictor of accuracy (de Vignemont & Singer, 2006; Ickes, 2016).

The empathic response and emotion identification are not only shaped by the modulatory factors mentioned above but also by additional processes that are not always captured in experimental settings. This includes phenomena that broad definitions sometime subsume under the label empathy, which can complicate the interpretation of observed empathic responses (Coll et al., 2017).

Emotional contagion can be understood as the adoption of another person's emotion with limited self-other distinction, such that one's own state is not clearly distinguished from the other's (de Waal, 1996; Hatfield et al., 1993). In contrast, empathy is typically linked to recognizing the target as separate and attributing one's affective change to the target's state (Bird & Viding, 2014; de Vignemont & Singer, 2006). In the framework of Coll et al. (2017) emotional contagion can be described as a case in which emotion identification is accurate but unconscious and affect sharing is intact. Relatedly, egocentric bias can distort emotion identification when observers rely too heavily on their own current state as a reference point, resulting in projections that may not reflect the target's actual state (Bukowski et al., 2020).

Relatedly, affective reactions to another person's state can differ qualitatively and are not always isomorphic to the target's emotion. Personal distress occurs in situations in which witnessing another person's suffering leads to such intense involvement that the person feels overwhelmed, experiences discomfort or stress and tends to withdraw socially. In contrast, compassion described as empathic concern or sympathy, is understood as an other-oriented state in which one experiences feelings for a person that are focused on their well-being, without this being an isomorphic affective state (Singer, 2025). It should also be noted that empathy does not automatically translate into prosocial behavior, but first needs to be transformed into compassion, which supports the motivation to help. Even though empathy is often followed by behavior, many definitions do not specify an immediate behavioral outcome as a component of empathy itself (Cuff et al., 2016; de Vignemont & Singer, 2006; Eisenberg & Miller, 1987).

Finally the emotion identification component of empathy overlaps with constructs that are often labeled cognitive empathy, mentalizing or theory of mind (ToM), which refer to the socio-cognitive ability to understand the mental states of others, such as intention, beliefs or emotions (Bird & Viding, 2014; Kanske, 2018; Singer, 2006, 2025). However, perspective-taking can be understood as a common way of supporting this cognitive route, which is also made plausible by the fact that it is used specifically in experimental contexts to increase empathic responses (Zaki, 2014).

Children's Development of Empathy

From a very early stage in children's development, it is already possible to describe undifferentiated forms of empathy-related reactions. Newborns show distress reactions when they hear other newborns crying, which is interpreted as an indication of an automatic, reactive mechanism in the sense of emotional contagion (Richaud et al., 2017). In his four-stage model, Hoffman (2000) describes the phase between birth and the first year of life as *global empathy* and emphasizes that at this age, self-other distinction has not yet been formed. These emotional reactions trigger such personal distress that it seems as if the experience and emotions were the child's own. This developmental stage reflects the different ontogenetic trajectories of the underlying neural structures. While early automatic forms of affective resonance rely on limbic and para-limbic structures that develop early in ontogeny, more advanced cognitive forms of empathy depend on neocortical areas, particularly prefrontal and lateral temporal structures, which are among the last to fully mature (Singer, 2006).

The second stage, according to Hoffman, is a phase between approximately 1 and 3 years of age, which he refers to as *egocentric empathy*. During this period, children are able to distinguish themselves physically from others, but understanding of other's inner states is still limited and only begins to develop more clearly toward the end of this age range (Hoffman, 2000). In parallel, self-other distinction, perspective-taking and emotion regulation develop in the second year of life, allowing self-concern to shift toward concern for others (Knafo et al., 2008). Between the ages of 2-3 emotional understanding develops alongside ToM and executive functions. This enables children to separate and maintain the perspectives of different people (Decety, 2010, 2011a, 2011b). In this context, it is also reported that emotional reactions occur at this age that match the emotional states of others. These are accompanied by more organized prosocial behaviors such as comforting, helping and sharing (Decety & Jackson, 2004; Hoffman, 2000).

Hoffman (2000) defines the third stage as occurring between the ages of approximately 3 and 8 years old and refers to it as *empathy for another's feelings*. During this stage, children become increasingly aware that other people can have emotions that do not correspond to their own and they develop a more attentive perception of emotional cues. A critical challenge during this period is overcoming the emotional egocentricity bias (EEB), where children's own affective states tend to distort their perception of other's feelings. Neuroscientifically the ability to maintain self-other distinction and overcome this bias is specifically connected to the right supramarginal gyrus (rSMG). Because this region shows prolonged maturation, primary school children often show a higher EEB than adults, as they still lack the full functional capacity to decouple their own emotions from those of others (Silani et al., 2013; Steinbeis, 2016; Steinbeis et al., 2015). Furthermore, linguistic development makes it easier to name feelings and express empathic responses. Moreover, empathy at this age is no longer limited to people in one's direct environment but can also be directed toward people one does not interact with directly, such as victims of a natural disaster (Neary, 2023; Schonert-Reichl, 2013). In addition it is described for the preschool years that children already demonstrate perspective-taking in pretend tasks, which is often used to map progress in theory of mind development (Richaud et al., 2017; Wellman et al., 2001).

The fourth stage is associated with late childhood and early adolescence and is referred to as *empathy for another's condition*. In this stage of development, empathy is understood to increasingly apply to people or groups whose identity, life story or circumstances lie far outside one's own experience, which is linked to growing cognitive maturity (Neary, 2023; Schonert-Reichl, 2013). Such observations can be classified within a broader developmental assumption according to which empathy gradually progresses from more automatic mechanisms to more cognitively driven forms (Hoffman, 2000). In line with this developmental shift age-related maturation from infancy through middle to late childhood is also associated with improvements of empathic accuracy (Ickes, 2016). Overall, developmental psychological models suggest that empathy-related responses evolve from early, undifferentiated forms of affective resonance toward more differentiated and cognitively supported forms (Hoffman, 2000; Richaud et al., 2017). At primary school age, children are typically increasingly able to distinguish between their own experiences and those of others, to perceive and interpret emotional cues from other people more accurately and to describe emotional states verbally. This provides important requirements for examining empathy-related reactions not only as unspecific resonances but in a more differentiated way (Hoffman, 2000; Neary, 2023; Schonert-Reichl, 2013). Given that the functional connectivity between rSMG and prefrontal structures continues to mature throughout late childhood, the age between 8-12 represents a critical window of plasticity. During this phase children increasingly develop the executive resources that are needed to regulate automatic affective resonance and consciously engage in perspective-taking (Bensalah et al., 2016; Richaud et al., 2017; Steinbeis, 2016; Steinbeis et al., 2015).

At the same time, primary school context is characterized by frequent interactions with peers, in which children have to get along with others and regulate negative emotions, and empathic responses are repeatedly challenged (Webster-Stratton & Reid, 2004). Since these responses are context dependent, this setting is particularly suitable for investigating the variability and selectivity of empathic experience (de Vignemont & Singer, 2006). Against this background, late childhood appears to be a particularly relevant time window for assessing empathy in standardized but everyday social situations and examining whether preventive programs can support the transition to a regulated empathy response (Parsons et al., 2017; Parsons & Phillips, 2016; Patel et al., 2007; Singer, 2025).

Challenges in Measuring Empathy in Children

The fact that empathy is defined differently in literature and encompasses several sub-processes is not only a theoretical problem, but also directly determines what can be measured and how results are to be interpreted. For research with children, this situation is further complicated by the fact that measurement markers must be selected according to age, and the same reaction can mean something different depending on the stage of development (Bensalah et al., 2016; Maxwell & DesRoches, 2010; Ştefan & Avram, 2018). It is therefore not surprising that several approaches are being discussed for children. These include self-reports, other-report from parents and teachers, performance-based questionnaires and observations in structured or natural scenarios in which behaviors such as vocalization, facial expressions, or helping behavior are interpreted (Neary, 2023). However, this variety of methods also makes it likely that measured values may diverge.

Questionnaires and rating scales are nevertheless widely used because they are economical and easy to use in larger samples. However, measuring empathy in children can be challenging, as answering questionnaires requires attention and a certain level of language development, whereas the dimension to be measured is complex and the items must be formulated in a way that is both appropriate for children and valid (Lietz et al., 2011; Sartori & Pasini, 2007). Traditional assessment tools often rely on a written barrier. Bale et al. (2020) emphasize that standard written questionnaires disadvantage children with limited reading and writing abilities, which can lead to distorted scores which reflect literacy skills rather than the actual concept of interest. Therefore, evaluating knowledge through verbal or interactive methods is advocated as a more inclusive and accessible approach for school aged children (Bale et al., 2020). Furthermore, certain items can be highly metacognitive, for example when a child is asked to remember and describe their own emotional experiences at the same time (Lamm et al., 2007). Such requirements are closely linked to developmental stages and can mean that a low score reflects limited self-reporting ability rather than low empathy levels.

Difficulties can also arise at the level of item structure. Bayne et al. (2024) show in a meta-analysis that the use of dichotomous items is a significant predictor of lower internal consistency compared to Likert scaling. While many traditional scales for measuring children's empathy use such simple formats to reduce cognitive load, the resulting lack of accuracy can compromise the psychometric robustness and internal consistency of the measurement.

Standardized situations can help here because they create comparability and facilitate interpretation without having to claim that this isolates a pure empathy signal. It is suggested that empathy tasks use a correspondence measure consisting of two separate ratings, one for the assumed emotional state of the target person and one for one's own feelings. This allows one to derive the accuracy of emotion identification and affective sharing as different components (Coll et al., 2017). Traditional pictorial and story-based measures are criticized for their static and simplified nature which typically rely on hypothetical two-persons interactions that may fail to evoke sufficient affect (Eisenberg & Miller, 1987; Sesso et al., 2021). However, social interactions in real life are inherently dynamic and complex, involving multiple social actors and nuanced signals such as facial expressions, voice and situational context (Bayne et al., 2024; Nitschke & Bartz, 2023). VR offers a promising solution to these challenges by creating immersive environments for direct observation of behavior in ecologically valid scenarios, while bypassing the barriers of literacy and the static nature of traditional tests (Bale et al., 2020; Parsons & Phillips, 2016; Sesso et al., 2021).

A good approach is therefore to combine different methods and complement standardized indicators with open-ended verbal responses. These answers allow children to describe their experiences and perceptions in their own words and help to understand the reasoning behind an emotional judgement (Bale et al., 2020; Bayne et al., 2024; Neary, 2023; Richaud et al., 2017; Sesso et al., 2021). At the same time, self-reports should be interpreted cautiously, as they may be influenced by social desirability and expectation effects and can be distorted by post-hoc reconstruction after the experience (Bayne et al., 2024; Owais & Yaacoub, 2020; Richaud et al., 2017).

Empathy in Virtual Reality

Empathic experience and response are understood in the literature as context-sensitive and selective reactions that can vary depending on the situation and interpersonal context, which can give cause to a fundamental measurement problem (Bayne et al., 2024; de Vignemont & Singer, 2006). The immersive nature of VR environments evokes a strong sense of presence, which can create more ecological and dynamic social situations than traditional methods. It can be seen as an option that combines standardization with those realistic social situations. Due to its ability to simulate subjective experiences in a realistic manner, VR is often referred to in literature as a powerful “empathy machine” (Barreda-Ángeles et al., 2021; Parsons & Phillips, 2016). A key argument is that empathic systems are designed to function in interactions with other people, and that vivid directly experienced stimuli can elicit stronger empathic responses than merely imagined scenarios (Preston, 2007). For measuring empathy, it is particularly relevant that VR can enable controlled simulations with affectively embedding background narratives, thereby supporting emotionally relevant experiences and social interactions. In this sense, VR can be conceptualized as an active emotion elicitation method because participants directly take part in the situation rather than only observing an emotional stimulus in a passive format (e.g. watching a movie) (Somarathna et al., 2023). VR is described as a way of obtaining more ecologically plausible data in simulations of real-world environments, while at the same time keeping dynamic perceptual stimuli controllable. A methodological advantage of VR is the high degree of standardization and experimental control it offers, including the precise and consistent presentation of avatars, environments, sounds and other dynamic features across all participants (Morasse et al., 2021). This allows researchers to isolate and modify variables efficiently and to keep key aspects of the virtual character's behavior constant, which increases comparability and interpretability relative to less controlled real-life interactions. In addition, technical advances can make implementation more efficient and facilitate systematic recording of responses (Owais & Yaacoub, 2020; Parsons et al., 2017; Parsons & Phillips, 2016). It is also important that VR can create social presence, meaning the experience of a connection to another person who is perceived as a social interaction partner (Muravevskaia & Gardner-McCune, 2022). Especially for children, VR may provide a high engagement factor that helps maintain attention during assessments. Children in primary school age can experience virtual characters as social interaction partners, which can promote behavior and emotional responding that resembles everyday social situations (Bailey & Bailenson, 2017; Muravevskaia & Gardner-McCune, 2022).

Social-Evaluative Task Context

Given that empathic responses are understood to be context-sensitive and selective, it makes sense to not only choose the measurement context (VR) but also to specify the structure of the situation in which empathy-related reactions are to be assessed. A social evaluative paradigm is considered an established procedure for triggering stress and emotional arousal under standardized conditions (Kirschbaum et al., 1993; Seddon et al., 2020). For empathy assessment, the occurrence of a socially relevant, potentially stressful situation can be helpful, in which it is plausible to observe another person in the same situation and to be able to interpret the resulting emotional judgements (Brown et al., 2021; Nitschke & Bartz, 2023). In this context, task-performance under challenge can be considered as informative, as it may reflect how executive control resources remain available under stress, supporting self-other distinction and adequate social judgements in a demanding social evaluative situation (Decety & Meyer, 2008; Forbes et al., 2024; Steinbeis, 2016).

Mental Health Literacy for Children

The concept of mental health literacy was originally developed by Jorm et al. (1997) as „knowledge and beliefs about mental disorders which aid their recognition, management or prevention". While research has focused on adulthood a long time, the importance of early promotion of MHL in childhood is increasingly gaining attention (Bale et al., 2020). Since approximately 75% of all mental disorders manifest themselves before the age of 24, late childhood represents a critical period for preventive interventions, with schools providing an ideal setting for this (Patel et al., 2007). Bale adapted and described the concept specifically for children aged 8-12 including six key components: Recognizing changes in one's own mental health, help-seeking behavior, knowledge about available support, knowledge about mental health influences, coping strategies and resilience, attitudes toward help-seeking and -giving, stigma and positive mindset (Bale et al., 2020).

An essential component of MHL programs is the enhancement of socio-emotional skills. This includes the ability to correctly identify and regulate one's own and other's emotions, which are core processes that are closely linked to empathy. Socially oriented training approaches have been shown to be particularly effective in improving various dimensions of empathy (Bale et al., 2020; Wu et al., 2024). This implies that MHL may be relevant for children's emotion understanding and the way they interpret and verbalize emotions in self and others, while changes in immediate affective convergence may depend on situational and regulatory processes. Applied to the framework of Coll et al. (2017) MHL content would be expected to primarily support emotion identification by strengthening children's emotion vocabulary, their use of contextual information, and inferential understanding of why someone might feel a certain way, thus improving recognition and categorization of the target's state. In addition, because MHL programs addresses coping, emotion regulation and supportive attitudes, they may increase children's motivation and attention to engage with other's emotions (Bale et al., 2020; Ickes, 2016).

The Present Study

The previous discussion reveals the need of the present study. Empathy is particularly relevant from a developmental psychology perspective at primary school age because key prerequisites for empathic responding further differentiate during this period, especially the distinction between one's own emotions and those of others and the regulation of one's own emotional reactions (Hoffman, 2000; Neary, 2023; Schonert-Reichl, 2013). Furthermore, everyday school life confronts children with situations in which empathic skills are relevant (Bensalah et al., 2016). Empathy is also linked to beneficial developmental outcomes, which strengthens the rationale for fostering these skills early on (Bayne et al., 2024).

Nevertheless, measuring empathy in children remains challenging. Immersive VR therefore offers a promising context because it can deliver standardized, vivid and interactive social scenarios while retaining experimental control (Parsons & Phillips, 2016; Somarathna et al., 2023). Robust indicators of empathic responding allow school-based prevention programs to be evaluated over time and further improved based on evidence (Bale et al., 2020; Bayne et al., 2024; Patel et al., 2007).

Against this background the present study addresses two linked aims. First, as a methodological proof of concept, it evaluates whether a specific VR classroom paradigm can capture empathic responding in ways that are theoretically plausible and interpretable. Self and other emotion ratings are collected and translated into two correspondence measures: *Other-self match* serves as a VR-based correspondence indicator of empathic responding, whereas *Within-child match* is used as a correspondence-based indicator of affective resonance. Additionally, children's open ended explanations are used to examine convergence and to inform the interpretation of the quantitative indicators (Neary, 2023).

Secondly, we examine whether a MHL program that strengthens the recognition, classification and management of one's own and other's emotions is associated with changes in the correspondence measures over time (Bale et al., 2020). Based on the MHL content, effects are expected to be particularly relevant for children's emotion identification processes. This raises not only the question of whether the paradigm can represent empathy in this context, but also whether the indicators used are sensitive to reveal potential pre-post changes in connection with a school-based prevention. This leads to the following research questions:

1) Does mental health literacy training influence children's other-self match and within-child match toward an NPC over time compared with a control group?
2) To what extent do other-self match and within-child match converge with children's qualitative explanations of their own and an NPC's emotion, and what do discrepancies imply about the suitability and limits of this VR paradigm for measuring empathy?

Therefore, the following hypotheses have been formulated:

Hypothesis 1: Children in the intervention condition show a greater pre-post increase in VR-based correspondence indicators than children in the control condition.

Hypothesis 2: Children in the intervention condition show a stronger pre-post increase in coded emotion identification and affective resonance than children in the control group.

Hypothesis 3: VR-based correspondence indicators align with children's qualitative reports such that higher other-self match toward the NPC is associated with a higher likelihood of coded affective resonance and emotion identification, whereas higher within-child match is associated with a higher likelihood of coded affective resonance.

Gender and task performance are included as theoretically motivated covariates in the main models. Specifically, girls are expected to show higher levels of VR-based match than boys, and task performance is associated with VR-based match indicators.

Method

COMET4 Kids

The present study is part of the COMET for Kids project at the Faculty of Clinical and Health Psychology at the University of Vienna under the supervision of Prof. Dr. Giorgia Silani, Prof. Dr. Barbara Schober and Prof. Dr. Kerstin Zechner. The aim of the project is to develop, implement and evaluate a MHL program for primary school children. The project evaluation relies on multiple data sources, including questionnaire-based measures, physiological recordings, a VR-based experimental task and short open-ended responses to assess various mental health aspects of children. The present study draws on the VR-based assessment and focuses specifically on empathy related outcomes assessed in this setting. Data were collected at pre-intervention in September and October 2024 and post-intervention in June 2025.

Study Design

The present study followed a longitudinal mixed factorial design with repeated measures. Condition (intervention vs. control) was modeled as a between subject factor with two levels, whereas time was modeled as a within subject factor with two levels, pre- and post-intervention. Condition assignment was not randomized but implemented at the school level, resulting in a quasi-experimental pre-post design. Between the two assessments, children in the intervention condition participated in the COMET4Kids MHL program, which implemented a weekly curriculum throughout the final year of primary school with a focus on strengthening children's understanding of mental health needs, coping skills and help-seeking behaviors. Children in the control condition continued with their regular school activities during the same period. In addition, the study used a mixed methods approach by combining quantitative outcomes from the VR-based assessment with qualitative data from short open-ended responses that were subsequently coded for analysis.

Participants

The present study included 193 children aged 8-11 years (M_{age} = 9.27, SD = 0.52) at study entry (pre-intervention). 48.19% (n = 93) of the children were girls and 51.81% (n = 100) were boys. Schools in Vienna were contacted and invited to participate in the COMET4Kids project. Overall, twelve classes from four schools took part in the study. One integration class was included, comprising students who required additional individual support due to developmental disorders or (learning) disabilities. With consent of the school's headmasters, children in the participating classes were invited to participate. Participation required informed consent from children and their parents, therefore not all students in each class participated.

Virtual Reality Apparatus

The 3D VR environment was delivered via an Oculus Meta Quest 3 head-mounted display. Children remained seated and interacted with the virtual environment via head and hand tracking. Several features were held constant across participants to support standardization. For example, interaction and error sounds were identical across children. Charlie's avatar was matched to the participant's gender. Physiological arousal was recorded continuously using an arm-worn sensor capturing heart rate during the VR procedure.

Study Procedure

At pre- and post-intervention, data collection was conducted at the participating schools and followed a standardized protocol administered by trained students and the research team. Prior to the VR task, the procedure was explained in age-appropriate language and children were reminded that participation was voluntary, that they could ask questions at any time and that they could stop the task without negative consequences. Four VR head mounted displays were available, allowing up to four children to participate simultaneously while completing the assessment individually.

The VR scenario was set in a virtual classroom with two NPCs: a female teacher standing in front of the classroom and a peer character named Charlie seated next to the participating child (see Figure 1). Charlie's gender was matched to the participant's gender, with a female version presented to girls and a male version presented to boys. The procedure began with a habituation phase to familiarize children with the virtual classroom environment, the characters and the task format. The virtual teacher delivered standardized instructions, followed by a brief trial run to practice the response demands (see Figure 2). Across the VR procedure at predefined points children completed repeated emotion ratings presented on the virtual desk in front of them.

Figure 1. Positioning of Charlie in the Virtual Classroom

Figure 2. Virtual Teacher Explaining the Task Procedure

Following habituation and the trial run, children completed the first emotion rating. An anticipation phase then preceded the task, after which children provided a second emotion rating. Children subsequently completed the stress-inducing task. Immediately after the task, children completed the third emotion rating. The task phase was followed by a short recovery period within the VR procedure after which children provided the fourth emotion rating.

Next, the procedure transitioned to the observation part. Children were informed that Charlie would now perform the same task, which constituted an anticipation period for the upcoming observation. Children then watched Charlie perform the same task, with the scenario scripted such that Charlie made noticeable mistakes that were evident via error sounds. Immediately after the observation, children first rated Charlie's emotion, using the same response format as in the self emotion ratings. Children then completed the fifth emotion rating, capturing their affective response following Charlie's performance. This was followed by a short post-task break period, after which children provided the sixth and final emotion rating.

In close temporal proximity to the VR procedure, brief open-ended questions were administered verbally, asking children to describe their experience. Responses were written down by trained students and the research team.

Measures

Gender

Gender was assessed via a brief self-report item in which children selected one of two options ("I am a boy" or "I am a girl"). Responses were coded as male = 0 and female = 1. No additional gender categories were offered.

Stress-inducing Task

The VR stress task was designed to induce stress in an age-appropriate format based on the Trier Social Stress Test for Children (TSST-C) (Buske-Kirschbaum et al., 1997). In this setting, social evaluative pressure was created through the presence of a virtual teacher who delivered evaluative prompts and through the peer character Charlie, which introduced an explicit social comparison component. The task consisted of a speeded number-sequencing challenge. Children were required to press six buttons in ascending numerical order (1-6) and to complete the sequence as often as possible within a fixed time window. Although the response demands were kept simple, mild stress was induced through time pressure, salient performance feedback (error sounds) and repeated reminders by the teacher emphasizing speed.

Task Performance

Task performance was recorded automatically by the VR system. Two variables were used: the number of completed rounds and task accuracy. Task accuracy was summarized as a proportion score computed as trials won divided by trial rounds. Thus, a speed-related indicator and an accuracy-related indicator were captured.

Emotion Ratings

Children reported their emotional state repeatedly during the VR procedure using a forced choice display shown in the virtual environment (see Figure 3). Each rating required selecting one out of six emotions: happy, proud, sad, angry, anxious, neutral. Emotions were presented as emojis and accompanied by written labels. Intensity ratings were collected after each category choice but not analyzed in the present study.

For the analyses in this study, three self-report ratings were taken that were relevant to the research questions: the rating immediately after the child completed the stress task, the rating prior to Charlie's task and the rating immediately after observing Charlie's task. To derive valence-based indicators, emotion categories were recoded into three valence categories. Happy and proud were coded as positive, sad, angry and anxious were coded as negative, and neutral remained neutral.

Figure 3. Self-Emotion Rating in VR; "Wie fühlst du dich jetzt?" means "How are you feeling now?"; Emotion labels are in the same order as mentioned above.

Emotion Identification: Rating of Charlie's Emotion

After observing Charlie's performance children indicated how they thought Charlie felt. This rating used the same six emotion categories as the self-ratings. Charlie's emotion rating was recoded into the same three valence categories to allow direct comparisons with the child's own emotion rating.

Match Variables: Correspondence Indicators of Empathic Responding

Two binary match variables were computed based on exact matching in the recoded valence categories. Other-self match (VR-based correspondence indicator of empathic responding) was operationalized as a match between the child's self-rating after observing Charlie and the child's rating of Charlie's emotion (i.e., match = 1 if both ratings fall into the same valence category, no match = 0). Within-child match compared the child's self-rating after completing the stress task with the child's self-rating after observing Charlie (i.e., match = 1 if both self-ratings fall into the same valence category, no match = 0). Baseline match was defined as a match between the child's self-rated valence prior to Charlie's task and the child's estimated valence of Charlie (i.e., match = 1 if both ratings fall into the same valence category, no match = 0).

Qualitative Measures and Coding

Open-ended questions

Children answered three short open-ended questions administered verbally directly after the VR task: (1)"How did you feel during the task?" (2) "How did you feel while watching Charlie? Why?" (3) "How do you think Charlie was feeling? Why?". Responses were documented by trained students and the research team in writing. All three questions were administered at pre- and post-intervention.

Coding Approach and Category Framework

The open-ended responses were analyzed using a qualitative content analysis with a deductive-inductive category approach, following Mayring & Fenzl (2019). Initial categories were theory-informed, and the category system was iteratively refined by adding categories that emerged as recurrent but were not covered by the initial framework. This resulted in a final codebook including stress, affective resonance, emotion identification, prosocial behavior, competition, amusement and indifference. For each category a definition, anchor examples and coding rules were specified. Qualitative emotion identification was defined as the process of attributing a mental state to the target and justifying this attribution through inferential reasoning. In line with theoretical models of social cognition, attributions were also coded when they did not correspond to the situational cues (Bird & Viding, 2014; Coll et al., 2017; Kanske et al., 2015). Qualitative affective resonance was defined as the verbal expression of an affect that corresponds to the target's state. This category reflects that verbal reports do not allow for precise verification of isomorphism, narrative resonance can overlap with sympathy (de Vignemont & Singer, 2006; Neary, 2023). The full codebook is provided in Appendix C.

Coding was conducted by a single coder. The unit of analysis was the child at each assessment. All three responses were reviewed separately and then coded jointly, such that codes reflect whether a category was present in the child's open-ended material across the three questions. Because responses were often brief, yet multifaceted, multiple categories could be assigned to the same response. Coding was carried out in Microsoft Excel.

Quantifying Qualitative Data

To facilitate integration with the quantitative outcomes of the VR data, qualitative codes were converted into binary indicators per category, reflecting whether a given category was present in the responses of the child. In the present study, analyses primarily focused on qualitative affective resonance and emotion identification. Coding these two categories separately reflects the methodological recommendation to distinguish emotion identification from affect sharing, while acknowledging that open-ended responses can only approximate this separation rather than provide fully independent measures (Coll et al., 2017). Stress and competition were examined descriptively as contextual indicators of how children experienced and framed the task. The remaining categories were retained for transparency or future use. Binary indicators were used for descriptive summaries and convergence analyses with VR indicators. Missing qualitative data occurred when the qualitative interview component was not administered or not available, resulting in no recorded responses. These cases were excluded from the descriptive analyses. All percentages are based on the available qualitative interviews.

Statistical Analysis

Only complete cases were retained, meaning children who participated in both the pre- and post-assessments. The cases in which a glitch in the program occurred, that prevented Charlie from moving, were excluded. Emotion ratings were recoded into valence categories (positive, neutral, negative) and the two match variables described in the Measures section (other-self match, within-child match) were calculated. Baseline match was calculated as the correspondence between the child's self-rating before the observation and the child's rating of Charlie's emotion. Task accuracy covariate was calculated, which is composed of "trial won" divided by "trial rounds". Gender was treated as a binary covariate (i.e., boys coded 0, girls coded 1).

Analyses were conducted to address the research questions and hypotheses specified at the end of the theoretical background. Due to the hierarchical structure of the data on three levels (child, class, school) and the binary match outcomes of the VR-based correspondence indicators for empathic responding, a binomial generalized linear mixed model (GLMM) with a logit link was used to examine how condition affects other-self match and within-child match over time, while controlling for several potential confounding variables. All analyses were conducted in R using RStudio (Version 2025.09.2+418). GLMMs were estimated using the lme4 package (Bates et al., 2015), additional packages supported data preparation, diagnostics, robust inference and reporting. To determine an appropriate random-intercept structure a series of null models (intercept only fixed effect) that differed only in their random effects (child, child + class, child + class + school) were fitted. Models were compared using likelihood-ratio tests, Akaike Information Criterion (AIC), singular fits and Intraclass Correlation Coefficient (ICC) were used to evaluate whether additional clustering levels were supported. This model-fitting step was used to decide whether random intercepts beyond child level were needed.

Two GLMM were specified to address the first research question. The first model used the other-self match outcome and the second model used the within-child match outcome. For both outcomes, fixed effects included condition (intervention vs. control), time (pre- vs. post-intervention), and their interaction as the primary effect of interest. Gender and task performance were included as covariates. If the random-intercept variance was estimated as zero (singular fit) in the full models, a simplified model was used.

Model diagnostics were conducted for both models to evaluate model adequacy. Simulation-based residual diagnostics were performed using the DHARMa package to assess residual uniformity, dispersion and outliers (Hartig, 2026). Multicollinearity among fixed effects was evaluated using variance inflation factors (VIF) as implemented in the performance package (Lüdecke et al., 2021).

Regarding the second research question, coded affective resonance and coded emotion identification were examined separately. Frequencies were summarized using cross-tabulations (row percentages, *n*) by condition and time to describe how each code changed within the control and intervention groups over time. Gender differences in the prevalence of each code were examined using cross-tabulations across time points. To assess convergence between the VR-based indicators and children's qualitative reports, additional cross-tabulations were computed between the VR match outcomes and coded affective resonance. Exploratory analyses used additional descriptive cross-tabulations to further characterize response patterns and to examine a potential link between qualitative codes and an additional VR indicator.

Results

Descriptives

The final sample consisted of 174 children in total, with 85 children in the intervention group and 89 children in the control group. Gender distribution was well balanced in both groups, with 45.9% (*n* = 39) girls in the intervention group and 53.9% (*n* = 48) girls in the control group.

At pre-intervention 29.4% (*n* = 25) of children in the intervention group showed a baseline match, 37.6% (*n* = 32) showed other-self match and 57.6% (*n* = 49) showed within-child match. In the control group baseline match was 30.3% (*n* = 27), other-self match was 31.5% (*n* = 28) and within-child match was 67.4% (*n* = 60) at pre-intervention.

At post-intervention 22.4% (*n* = 19) of children in the intervention group showed a baseline match. Moreover, 25.9% (*n* = 22) showed other-self match and 56.5% showed a within-child match. In the control group baseline match was 21.3% (*n* = 19), other-self match was 31.5% (*n* = 28) and within-child match was 53.9% (*n* = 48). Table 1 summarizes the descriptives separately for the control and intervention groups at pre- and post-intervention on child-level.

Overall, the descriptives suggest different pre-post patterns between the groups. At pre-intervention the intervention group showed a more pronounced increase from baseline-match to other-self match, whereas at post-intervention this increase was smaller. Conversely, in the control group the difference from baseline match to other-self match was modest at pre-intervention but appeared larger at post-intervention. Within-child match showed a descriptive decline from pre- to post-intervention in the control group, whereas it remained comparatively stable in the intervention group.

Table 1

Sample Descriptives by Condition and Time.

Condition	Time	N	Girls	Baseline match	Other-self match	Within-child Match
Intervention	Pre	85	39 (45.9%)	25 (29.4%)	32 (37.6%)	49 (57.6%)
Control	Pre	89	48 (53.9%)	27 (30.3%)	28 (31.5%)	60 (67.4%)
Intervention	Post	85	39 (45.9%)	19 (22.4%)	22 (25.9%)	48 (56.5%)
Control	Post	89	48 (53.9%)	19 (21.3%)	28 (31.5%)	48 (53.9%)

Note. Entries are n (%). Percentages are calculated within each condition and time point. Descriptives are reported at the child level.

Figure 4 shows the distribution of children's self-valence at the 3rd emotion rating (i.e., immediately after the child's own task) by condition and time. In the intervention group positive valence was the most frequent response at both pre- and post-intervention and the distribution appeared broadly stable over time with only modest changes in the ratings. In the control group positive valence also dominated at both assessments. However, more neutral ratings were observed at post-intervention compared to pre-intervention.

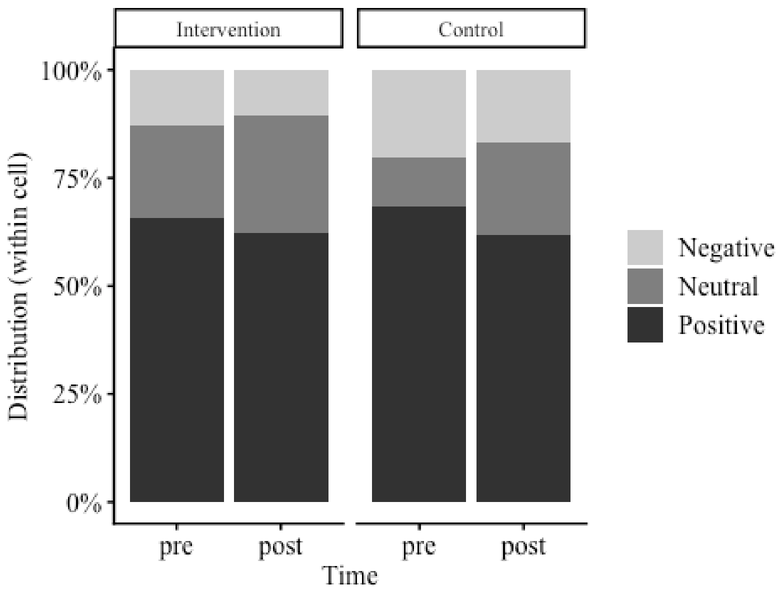

Figure 4. Self-Valence (3rd emotion rating) Distribution by Condition and Time

Figure 5 shows the distribution of the valence rating attributed to Charlie by condition and time. In the intervention group Charlie was predominantly perceived as negative at both pre-intervention and post-intervention, with more negative and fewer positive ratings at post-intervention compared with pre-intervention. A similar pattern was observed in the control group, where negative ratings also constituted the largest part at both assessments an appeared higher at post-intervention than at pre-intervention, while neutral ratings remained similar.

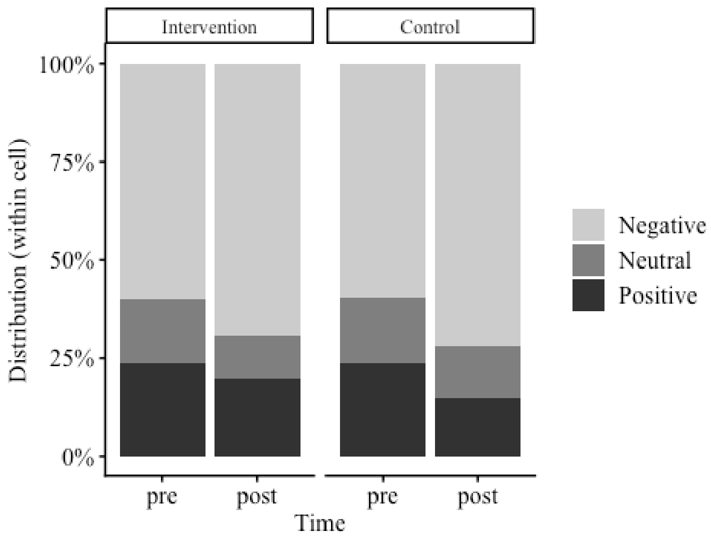

Figure 5. Charlie Valence by Condition and Time

Figure 6 shows the distribution of children's self-valence after Charlie's task by condition and time. In the intervention group self-valence was most frequently rated as positive at both assessments. The distribution appeared largely stable over time with only small shifts in the neutral and negative categories. In the control group positive ratings again represented the largest share at both pre- and post-intervention. However, fewer positive and more neutral ratings were observed at post-intervention compared to pre-intervention in the control group.

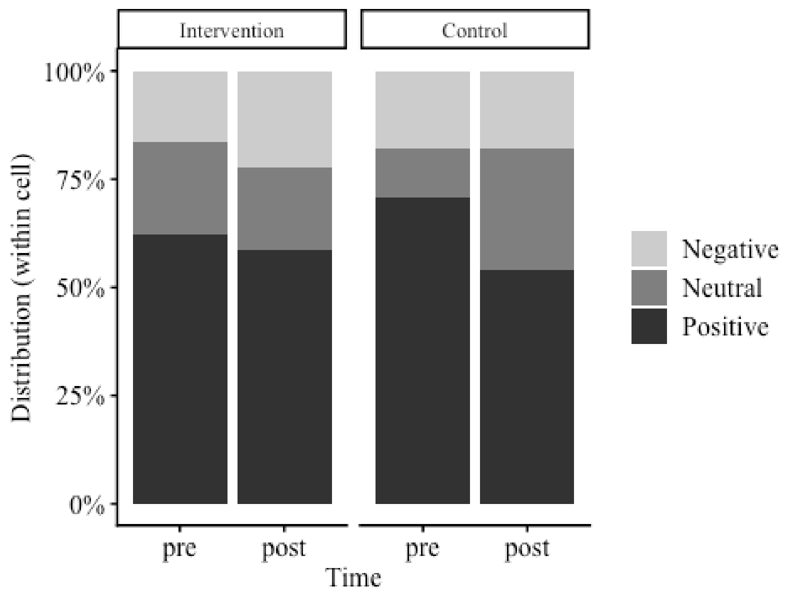

Figure 6. Self-Valence (5^{th} emotion rating) by Condition and Time

Model Fit

To evaluate whether the pre-post structure of the data and the nesting of children within classes and schools required multilevel modeling a set of null (intercept-only) logistic mixed models was fitted with successively expanded random-effects structures (Table 2). First a baseline model with a random intercept for child was calculated. Second a random intercept for class was added, followed by a model that additionally included a random intercept for school. Model comparison indicated that adding class did not improve model fit compared to the child-only model (ΔAIC = 0.45; χ^2 (1) = 1.55, p = .214). Adding school resulted in a boundary (singular) fit and provided no evidence of improved fit compared to the child + class model (ΔAIC = 2.45; χ^2 (1) = 0.00, p = 1.000). *ICC* estimates were small and nearly identical across the non-singular null models (*ICC* = .037 - .038), suggesting that clustering beyond child level contributed little additional variance in the outcome. Mixed models for other-self match and within-child match were specified with a child-level random intercept. In the models including all predictors, the child-level random-intercept variance collapsed to zero (singular fit), indicating no meaningful additional clustering at child, class or school level. The main analyses were therefore conducted using logistic regression models (GLM) with standard errors clustered at the child level. Model diagnostics were examined using simulation-based residual checks and standard collinearity diagnostics. No violation or notable irregularities were observed, indicating that the model assumptions were adequately met for inference.

Table 2

Null Model Comparisons for Random-effects Structure.

Model	Log-likelihood	k	AIC	ΔAIC	LRT χ^2	LRT df	p	ICC
1: child	-217.02	2	438.04	0.00				0.037
2: child + class	-216.25	3	438.49	0.45	1.55	1	.214	0.038
3: child + class + school	-216.25	4	440.49	2.45	0.00	1	1.000	

Note. Intercept-only logistic mixed models compared via likelihood-ratio tests (LRT) and AIC. Log-likelihood and k (number of estimated parameters) are reported for transparency. ICC is reported on the latent scale for a logit model (approximation).

Hypotheses Testing

Hypothesis 1

To test the first hypothesis, which posited that children in the intervention condition would show a stronger pre-post increase in VR-based match indicators than children in the control group, two logistic regression models were estimated with Wald confidence intervals based on cluster- robust standard errors clustered by child. The first model used other-self match as the outcome and included condition, time (pre- vs. post-intervention), their interaction (condition x time), while controlling for gender and task performance (Table 3). The second model examined within-child match using the same predictors and controlled for the same covariates (Table 4).

Contrary to Hypothesis 1, in the first model the condition x time interaction was not significant (*OR* = 0.60, 95% CI [0.24, 1.49], z = -1.10, p = .271), indicating that the change in other-self match over time did not differ reliably between groups. In line with this, neither the conditional effect of condition at pre-intervention (*OR* = 1.36, 95% CI [0.71, 2.60] , z = 0.93, p = .352) nor the conditional effect of time within the control group (*OR* = 1.07, 95% CI [0.56, 2.05], z = 0.20, p = .844) reached significance. Therefore, Hypothesis 1 was not supported for the primary outcome.

With regards to the covariates gender showed a significant association, with results for girls demonstrating higher odds of other-self match in comparison to boys (*OR* = 1.84, 95% CI [1.14, 2.97], z = 2.48, p = .013). However, task performance indicators were not significantly related to other-self match. Neither task accuracy (*OR* = 0.30, 95% CI [0.05, 1.76], z = -1.33, p = .183) nor trial rounds (*OR* = 0.96, 95% CI [0.83, 1.10], z = -0.59, p = .556) showed a significant association with the outcome.

Table 3

Logistic Regression Predicting Other-self Match (GLM; cluster-robust SEs).

Predictor	OR	95% CI	z	p
Condition	1.36	[0.71, 2.60]	0.93	.352
Time	1.07	[0.56, 2.05]	0.20	.844
Gender	1.84	[1.14, 2.97]	2.48	.013
Task Accuracy	0.30	[0.05, 1.76]	-1.33	.183
Trial Rounds	0.96	[0.83, 1.10]	-0.59	.556
Condition × Time	0.60	[0.24, 1.49]	-1.10	.271

Note. OR = odds ratio. 95% CIs are Wald confidence intervals based on cluster-robust standard errors (clustered by child). Reference categories: pre-intervention; control group; gender = male.

Results for within-child match converged with the first model. The condition x time interaction was not significant (*OR* = 1.65, 95% CI [0.69, 3.97], z = 1.12, p = .262), indicating no evidence for a pre-post change between conditions. Neither the conditional effect of condition at pre-intervention (*OR* = 0.65, 95% CI [0.35, 1.22], z = -1.34, p = .179), nor the conditional effect of time within the control group (*OR* = 0.56, 95% CI [0.29, 1.08], z = -1.73, p = .084) reached significance. Gender was not associated with the outcome (*OR* = 1.00, 95% CI [0.65, 1.55], z = 0.02, p = .988). The task performance indicators task accuracy (*OR* = 2.47, 95% CI [0.50, 12.21], z = 1.11, p = .268) and trial rounds (*OR* = 0.98, 95% CI [0.86, 1.11], z = -0.30, p = .766) did not show significant effects. Taken together, across both operationalizations, Hypothesis 1 was not supported.

Table 4

Logistic Regression Predicting Within-child Match (GLM; cluster-robust SEs).

Predictor	OR	95% CI	z	p
Condition	0.65	[0.35, 1.22]	-1.34	.179
Time	0.56	[0.29, 1.08]	-1.73	.084
Gender (female vs male)	1.00	[0.65, 1.55]	0.02	.988
Task Accuracy	2.47	[0.50, 12.21]	1.11	.268
Trial Rounds	0.98	[0.86, 1.11]	-0.30	.766
Condition × Time	1.65	[0.69, 3.97]	1.12	.262

Note. OR = odds ratio. 95% CIs are Wald confidence intervals based on cluster-robust standard errors (clustered by child). Reference categories: pre-intervention; control group; gender = male.

Hypothesis 2

To test Hypothesis 2, which predicted a stronger pre-post increase in coded emotion identification and affective resonance in the intervention group compared to the control group, descriptive cross-tabulations (row percentages, *n*) were calculated by condition and time and visualized in Figure 7 (affective resonance) and Figure 8 (emotion identification). The full cross-tabulation table is reported in the Appendix B (Table B1). The values are based on the non-missing qualitative responses. Missingness in the intervention group was 0.0% (n = 0) at pre-intervention and 11.8% (n = 10) at post-intervention and in the control group 1.1% (n = 1) at pre-intervention and 12.4% (n =11) at post-intervention, referring to cases in which no responses were received from the children.

In the intervention group, affective resonance increased from 24.7% (n = 21) at pre-intervention to 37.3% (n = 28) at post-intervention, whereas the control group showed an increase from 26.1% (n = 23) at pre-intervention to 32.1% (n = 25) at post-intervention. Descriptively, this pattern is consistent with Hypothesis 2 in that the pre-post increase appears larger in the intervention group than in the control group.

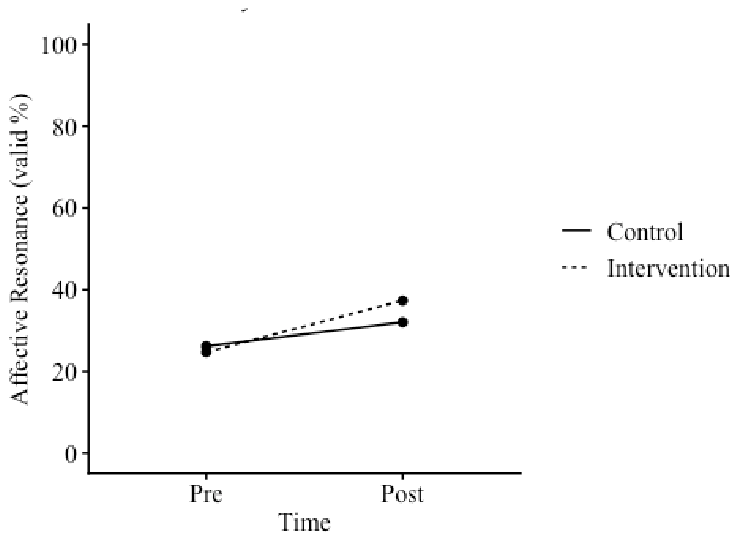

Figure 7. Affective Resonance by Condition and Time; Percentages based on non-missing responses.

For emotion identification, rates were very high already at pre-intervention across both groups, with 83.5% (*n* = 71) in the intervention group and 88.6% (*n* =78) in the control group. At post-intervention rates were higher across both groups. 92% (*n* = 69) of children in the intervention group and 93.6% (*n* = 73) in the control group showed emotion identification in their answers. The pre-post increase is slightly larger in the intervention group, which is consistent with Hypothesis 2.

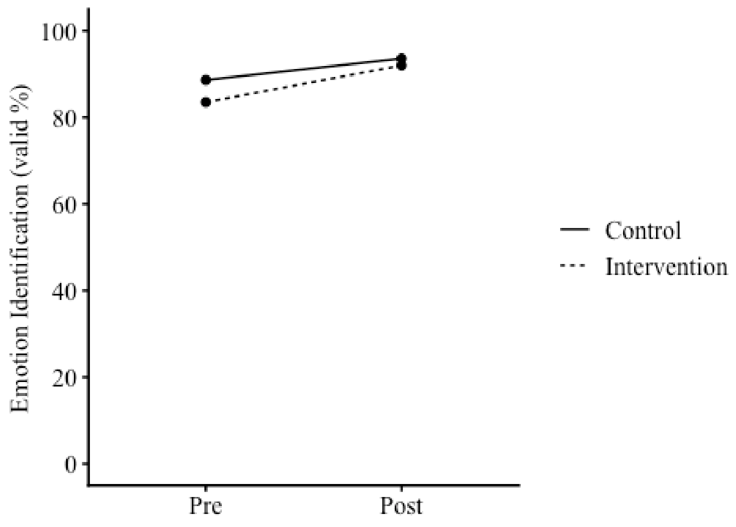

Figure 8. Emotion Identification by Condition and Time; Percentages based on non-missing responses.

Hypothesis 3

To test the third hypothesis, which suggested convergence between the VR-based indicators and children's qualitative reports, cross-tabulations were examined linking the VR match variables to coded affective resonance. Missing cases were excluded.

For other-self match, the descriptive pattern at pre-intervention did not support the expected association. At pre-intervention the proportion of affective resonance was slightly lower among children who showed other-self match (23.7%, n = 14) than among those who did not (26.3%, n = 30). At post-intervention, the proportions were very similar across groups, with 35.7% (n = 15) affective resonance in the match group and 34.2% (n = 38) in the non-match group, indicating only a negligible difference in the expected direction.

For coded emotion identification, the descriptive pattern was opposite to the expected association at pre-intervention. Specifically, emotion identification was less frequent among children who showed other-self match (78,0%, n = 46) than among those who did not (90.4%, n = 103). At post-intervention, emotion identification was very common in both groups and the difference largely attenuated, with 90.5% (n = 38) in the match group and 93.7% (n = 104) in the non-match group, indicating a minimal between-group difference.

A comparable pattern emerged for the within-child match variable. At pre-intervention, affective resonance was slightly less frequent among children who showed within-child match (24.1%, n = 26) compared to those who did not (27.7%, n = 18). At post-intervention, affective resonance was higher overall but remained lower in the within-child match group (32.6%, n = 28) than in the non-match group (37.3%, n = 25). Taken together, the descriptive cross-tabulations did not indicate that children who matched in the VR were more likely to express affective resonance in their qualitative answers. Therefore Hypothesis 3 was not supported. All descriptive cross-tabulations for Hypothesis 3 are provided in Appendix B (Tables B2-B4)

Exploratory Analyses

Gender

To mirror the role of gender as a theoretically motivated covariate in the main models and to descriptively describe potential gender differenced in children's qualitative reports, additional cross-tabulation were examined for affective resonance and emotion identification across gender and time. Rates of coded affective resonance were broadly similar for boys and girls at pre-intervention, with 26.4% (n = 23) of boys and 24.4% (n = 21) of girls expressing affective resonance. At post-intervention, affective resonance increased slightly for boys

(29.5%, *n* = 23) and more noticeably for girls (40.0%, *n* = 30), suggesting a descriptively stronger pre-post increase among girls. Emotion identification at pre-intervention was coded for 87.4% (*n* = 76) of boys and 84.9% (*n* = 73) of girls. At post-intervention, rates increased to 93.6% (*n* = 73) for boys and 92.0% (*n* = 69) for girls. The descriptive cross-tabulations are reported in Appendix B (Table B5-B6).

Stress

To descriptively characterize stress-related content in children's open-ended responses, the prevalence of the stress code was summarized by group and time. In the intervention group, stress was coded for 12.9% (*n* = 11) of children at pre-intervention and 12% (*n* = 9) at post-intervention. In the control group, stress was coded at comparable levels across assessments, with 13.6% (*n* = 12) at pre-intervention and 16.7% (*n* = 13) at post-intervention. Overall, stress-related content was coded for a relatively small proportion of children across conditions and time points.

Competition Framing

To characterize how children framed the task in terms of performance comparison with Charlie, competition-related statements were examined descriptively by group and time. Competition was coded less frequently at post- than at pre-intervention in both groups, decreasing from 16.5% (*n* = 14) to 9.3% (*n* = 7) in the intervention group and from 14.8% (*n* = 13) to 11.7% (*n* = 9) in the control group. The descriptive prevalence table summarizing stress and competition framing by condition and time is presented in Appendix B (Table B7).

Discussion

The present study examined whether a VR-based stress paradigm can capture children's empathic responding in a quantifiable manner and whether VR-based indicators are sensitive to pre-post change associated with a MHL program for children. Using a quasi-experimental pre-post design with an intervention and a control group, empathic responding toward an NPC was operationalized through correspondence indicators for empathic responding derived from self-related ratings and other-related ratings (Coll et al., 2017). To facilitate interpretation of these quantitative VR indicators, short open-ended responses were coded, with the main analyses focusing on affective resonance and emotion identification (Neary, 2023). This mixed-methods approach was motivated by the conceptualization of empathy as multidimensional and context-sensitive, and that correspondence measures can assess different sub-processes of empathic responding (Bayne et al., 2024; Coll et al., 2017; de Vignemont & Singer, 2006).

Building on this mixed-methods approach, the analyses yielded several findings regarding VR-based empathic responding. First, there was no evidence that the intervention group showed a stronger pre-post increase than the control group in the correspondence indicators for empathic responding across both operationalizations. With regard to the covariates, gender was significantly associated with other-self match, with girls showing higher odds of matching than boys, whereas task performance indicators were not significantly associated with other-self match. In contrast, neither gender nor task performance was associated with the alternative correspondence indicator. Second, qualitative coding suggested rates of affective resonance increased descriptively from pre- to post-intervention in both groups, with a marginally larger increase in the intervention group. Emotion identification was already highly prevalent at pre-intervention and increased slightly further at post-intervention across both conditions, descriptively showing a slightly larger increase in the intervention group. Third, the expected convergence between the VR-based correspondence indicators and qualitative reports was not supported, as children who matched in the VR did not show higher rates of affective resonance or emotion identification in their open-ended answers. Additional descriptive cross-tabulations indicated that affective resonance increased more noticeably for girls from pre- to post-intervention than for boys, whereas emotion identification was often coded in the responses of both genders.

Interpretation of Results

A key finding of this study is that the VR correspondence indicators provide a limited signal of empathic responding. According to the standard definition of empathy, empathic responding implies that the affective state of the empathizer is elicited by and isomorphic to the state of the target (de Vignemont & Singer, 2006). In correspondence approaches, this idea is often operationalized by comparing the affective state of the empathizer with the perceived or assumed affective state of the target (Coll et al., 2017). However, Coll et al. (2017) argue that such correspondence measures are ambiguous because they can conflate at least two different components. On the one hand emotion identification is measured and on the other hand the mapping from identification to self-state is measured, while additional modulatory and task-related processes can shape whether an empathic response occurs and how it is expressed (Coll et al., 2017; de Vignemont & Singer, 2006). The present results illustrate this ambiguity. The absence of an intervention effect over time in the correspondence indicators, the presence of a gender effect only for one correspondence indicator and the lack of convergence with qualitative reports together suggest that the VR indicators reflect a mixture of processes rather than a unitary empathy signal.

Boundary Conditions for Eliciting Empathic Responding in the Paradigm

One contextual observation that influences the interpretation of the hypotheses is that the intended stress-inducing framing was not strongly reflected in children's open-ended responses. Stress-related content was coded only in a small proportion of children at both assessments. This pattern is consistent with evidence that TSST variants using a virtual audience can elicit measurable stress responses but typically less strongly than comparable tasks conducted in front of a real audience (Kelly et al., 2007). In addition, competition-related framing was present in some children's responses, suggesting that the situation may have been appraised at least partly as a performance comparison or "game-like" challenge rather than as a stressful socio-evaluative threat. This suggests that for many children the VR task was not stressful or threatening. However, the assumed mechanism behind the paradigm is that a socially evaluative, potentially stressful situation creates an affectively meaningful context in which observing Charlie's struggles should elicit empathic responding. Without this aversive context empathic responding may be attenuated (Brown et al., 2021; Nitschke & Bartz, 2023). If children's own task experience is characterized by positive engagement, pride or fun rather than aversive stress, their affective state immediately after their own task and their state after observing Charlie may be driven by self-focused appraisal dynamics (de Vignemont & Singer, 2006).

Moreover, a competition frame can reduce the motivational tendency to attune to the target's affect, because attention and appraisal are pulled toward self-relevant outcomes (de Vignemont & Singer, 2006). In line with this, competitive rather than cooperative context can dampen empathic responding and may even promote counter-empathic reactions (Lamm & Majdandžić, 2015). Some children were perhaps more focused on comparing themselves and on observing whether they were better or worse at the task than Charlie than to empathize. Even if that was not the case, the short time between the own task and Charlie's observation might have made it difficult to focus entirely on Charlie and his task. Empathic accuracy research highlights attention to the target as a predictor of accurate and aligned responding, suggesting that limited target focus could weaken matching between children's judgement of Charlie and their own reported state (Ickes, 2016).

Another modulating factor for an empathic response could have been that Charlie's emotion was not conveyed with clear valence, salience and intensity, given his neutral and unreadable facial expression throughout the task. The only indicator for his inner state was therefore the error feedback sounds in the task. (de Vignemont & Singer, 2006; Ickes, 2016). Consequently, children's emotion attributions may have relied on inferential reasoning from performance cues, potentially anchored to their own error experience (Marheinecke et al., 2025)

Hypothesis 1

Contrary to Hypothesis 1, the intervention group did not show a stronger pre-post increase than the control group in either other-self match or within-child match. This finding requires separating conclusions about intervention effectiveness from conclusions about the construct validity of the correspondence operationalization. Within Coll et al.' framework an empathic response is the outcome of at least two separable processes: emotion identification and affect sharing. Although the present paradigm collected both components as two separate ratings, they were collapsed into a single binary match score and they were obtained sequentially. As a result, mismatch is difficult to interpret, because it can arise from differences in emotion identification, differences in affect sharing or contextual influences on either process (Coll et al., 2017). In addition, without an independent reference for Charlie's emotional state, emotion identification accuracy cannot be quantified in the strict sense that would be needed to assume empathy after the standard definition. Therefore, the null finding does not allow a specific conclusion about emotion identification or affect sharing, because the binary match outcome cannot adjudicate whether these components were unchanged or whether their effects would offset each other at the level of correspondence. Moreover, correspondence scores may be especially insensitive under the contextual condition described earlier, in which other-elicited responding may be less likely to emerge (de Vignemont & Singer, 2006; Ickes, 2016).

Finally, the outcome does not allow conclusions about whether valence alignment reflects affective sharing in the narrow sense, or personal distress. A child may report negative affect during observations because Charlie's performance is aversive or stressful to watch, rather than because the child shares Charlie's state triggered in an other-elicited way (de Waal, 1996; Hatfield et al., 1993; Singer, 2025).

Within-child matches may likewise have been affected by contextual constraints, such that the correspondence indicator could have been disturbed even if empathy-related processes were present. The underlying idea of within-child match was that observing Charlie would reinstate a similar affective state as during the child's own task, consistent with the argument that affective resonance becomes more likely when the target's state resembles the empathizer's own past experience and may be limited when such a relevant experience is lacking (de Vignemont & Singer, 2006; Preston, 2007). Interpreted as empathic responding, this requires that Charlie's situation is experienced as sufficiently similar to the child's own situation, especially regarding stress and perceived socio-evaluative threat. However, if the own task was often appraised as manageable, enjoyable or competition-like rather than threatening, Charlie's observation is unlikely to reinstate the child's affect after the own task in an empathy-specific way. In addition, within-child match does not incorporate an explicit emotion identification component, because it does not require an appraisal of Charlie's state. As a result, variation in within-child match may reflect self-regulatory processes, rather than target oriented empathic responding.

The covariate findings provide additional insights into the content of the VR correspondence indicators. For other-self match, girls showed significantly higher odds of matching than boys. This is consistent with literature reporting gender differences in socio-emotional responding and emotion-related reporting tendencies (Allemand et al., 2015; Bensalah et al., 2016; Wu et al., 2024). At the same time empathic accuracy research emphasizes that gender effects may reflect motivational and attentional factors rather than differences in ability. In the present context, girls may have been more likely to engage with Charlie's emotional situation or to rate their own affective state in a way that was more consistent with their attribution of Charlie's emotion (Zaki, 2014). This does not mean that other-self match is a pure measure of empathy, but it does show that a pattern was found here that matches existing literature. In contrast, for the second correspondence indicator, within-child match was not associated with gender. While this null finding should not be overinterpreted, the differing covariate pattern across the two indicators is compatible with the possibility that they capture overlapping but not identical processes (Coll et al., 2017).

Taken together, Hypothesis 1 results suggest that the VR correspondence indicators did not provide evidence for MHL-related change in valence alignment over time. The broader pattern supports the interpretation that these indicators capture a mixture of emotion attribution, self-focused appraisal and affective dynamics rather than a clearly other-elicited empathic response in the narrow sense (Coll et al., 2017; de Vignemont & Singer, 2006). More specifically, limited subjective stress, competition-related framing, ambiguous target cues, reliance on inferential reasoning from performance cues, as well as potential carry-over from the child's own task, may have reduced the likelihood that the paradigm elicited a consistent isomorphic response that could be captured by simple match variables. This interpretation follows the idea that empathic responding is shaped by situational modulators and attentional focus (Brown et al., 2021; Coll et al., 2017; de Vignemont & Singer, 2006; Ickes, 2016).

Hypothesis 2

Hypothesis 2 predicted a stronger pre-post increase in qualitative emotion identification and affective resonance in the intervention compared to the control group. The descriptive pattern was broadly consistent with this expectation for affective resonance, which increased in both groups from pre- to post-intervention, with a larger increase in the intervention group. Cognitive empathy was very common already at pre-intervention and increased slightly at post-intervention in both conditions. It also increased slightly more in the intervention group than in the control group. This is theoretically plausible given the content focus of MHL programs for children. Bale et al. (2020) emphasize components such as understanding one's own and other's emotions, recognizing psychological changes and learning coping and support strategies. These components can be applied to children's ability to name emotions, explain why someone has certain feelings and express affective and socially meaningful responses. Correspondingly, changes in these processes that can be particularly well captured in open-ended responses (Bale et al., 2020). However, these patterns should be interpreted with caution, as they are descriptive trends that were not examined using inferential statistical analyses.

Qualitative gender patterns suggested a stronger pre-post increase in affective resonance among girls, whereas emotion identification was similarly high for both genders. This aligns with prior research showing that gender differences are more pronounced in affective components (Bensalah et al., 2016; Ickes, 2016; Wu et al., 2024). A plausible explanation for why this difference is only noticeable at post-intervention could be age-related gains in verbal expressiveness, making it easier to talk about inner states (Bale et al., 2020). This observation should likewise be interpreted cautiously, as it also reflects a descriptive trend.

Hypothesis 3

Hypothesis 3 predicted convergence between the VR-based correspondence indicators and qualitative empathy indicators, such that children who matched in the VR would be more likely to express affective resonance and emotion identification. This was not supported. The descriptive pattern was small and sometimes opposite to the expected direction, with children who did not match in the VR reporting comparable higher rates of qualitative emotion identification and affective resonance. More generally, as discussed earlier, the VR correspondence indicators do not unambiguously reflect an empathic response in the narrow sense, because they can reflect a mixture of emotion attribution, self-focused appraisal dynamics and contextual modulation rather than clearly other-elicited affective resonance

(Coll et al., 2017; de Vignemont & Singer, 2006). In the present paradigm, the VR correspondence indicators are based on forced-choice self-report states at a narrow time window and are subsequently compressed into three valence categories and then into a binary match. This operationalization can compress mixed affective states into a single value (Bayne et al., 2024). A child may feel relieved after finishing the own task and simultaneously feel sorry for Charlie. Such a response could be coded as affective resonance in open-ended responses but would not necessarily produce a VR-match.

In addition, the two measurement channels emphasize different aspects of empathic responding. The VR match indicators require self-report that is embedded in a sequence containing self-related appraisal and other-related appraisal, whereas qualitative responses allow children to articulate emotional reasoning (Bale et al., 2020; Bayne et al., 2024; Neary, 2023). In this sense, the qualitative indicators may capture a more reflective narrative account of children's interpretation of the situation that may be shaped by response tendencies such as social desirability (Bayne et al., 2024; Richaud et al., 2017; Zaki, 2014). At the same time, it is important to note that qualitative affective resonance likely captures a broader set of empathy-related expressions than affect sharing alone, including compassion (Singer, 2025). Therefore, qualitative emotion identification and affective resonance can be present even when there is no VR match, because the one depends on valence alignment at one time point rather than on the richness of the child's explicit interpretation (Coll et al., 2017).

Strengths and Limitations

A central strength of the study is the use of a standardized, immersive and vivid VR paradigm that is well suited for primary school children (Parsons et al., 2017; Parsons & Phillips, 2016). The VR procedure required little language and literacy, reducing barriers typical of questionnaire-based measurements and supporting more inclusive assessments in heterogeneous school samples (Bale et al., 2020; Parsons & Phillips, 2016). In addition, the VR scenario appeared to function as a concrete basis for subsequent conversation about the own and the target's state. Qualitative responses indicated that Charlie was treated like a real person whose emotions could be inferred and discussed (Muravevskaia & Gardner-McCune, 2022). Children were able to verbalize emotions, reasons, and the perspective of the NPC in the open-ended questions, suggesting that the paradigm can elicit meaningful reflection about self and other (Bale et al., 2020; Bayne et al., 2024; Neary, 2023).

The study further benefits from its scope and design. With 174 children providing pre-post data, the sample size is comparatively strong for a school-based VR assessment. The longitudinal pre-post design with a control group strengthens interpretability of change patterns. Finally, the mixed methods proof-of-concept approach is a further strength. Combining VR indicators with open-ended responses improved interpretability by allowing convergence and discrepancies to be examined rather than assuming that correspondence measures capture empathy directly (Coll et al., 2017; Neary, 2023). The classroom context with peer observation and evaluative teacher prompts also adds ecological relevance for late-childhood social experiences (Webster-Stratton & Reid, 2004).

Several limitations constrain inference about empathic responding. First, the paradigm provides no clear reference for Charlie's emotional state and offers limited affective cues, so emotion identification accuracy cannot be assessed, and children's judgements remain plausibility-based inferences under ambiguous target information (Coll et al., 2017; Ickes, 2016). Second, stress appeared weakly elicited in children's reports and some responses reflected a competition frame, which likely reduced conditions under which other-elicited, isomorphic responding would be expected (de Vignemont & Singer, 2006). Third, repeated emotion ratings may also have increased boredom or heuristic responding.

The qualitative component has additional constraints. Responses were collected by multiple interviewers, missing qualitative data occurred especially at post-intervention and coding was conducted by a single rater without assessing interrater reliability. Moreover, qualitative affective resonance likely captures a broader range of empathy-related expressions, which may further reduce direct convergence with valence-based correspondence measures. Finally, construct validity is limited as binary valence alignment may conflate emotion identification, affective responding, regulation and context-driven appraisal, which is consistent with the observed lack of convergence with qualitative indicators (Bayne et al., 2024; Coll et al., 2017; de Vignemont & Singer, 2006).

Implications for Future Research

Future research should aim to more rigorously disentangle empathy-related mechanisms. Building on component-based approaches, future paradigms could assess emotion identification and affective sharing as separate processes through independent and ideally objective measures and examine how they are shaped by other appraisal or empathy related processes (Coll et al., 2017; de Vignemont & Singer, 2006). This would allow a more precise interpretation of which components change over time and which are sensitive to interventions such as MHL.

A key improvement concerns the specification of the target's affective state. To enable stronger inferences about emotion identification and to increase the likelihood of other elicited affective responding, future VR implementations should provide richer and less ambiguous cues in the target character. This could be achieved through clearer facial affect, body language and verbal emotional expressions. This would increase target readability and reduce reliance on purely performance-based contextual inference (Ickes, 2016).

Future work could adjust the procedure to better separate self-focused task appraisal from the observation phase. One option is to increase the salience of stressors in a developmentally appropriate way to ensure that the context reliably constitutes a socially evaluative challenge. At the same time, to prevent carry-over of the child's own affect, a brief neutral reset segment could be included between the own task and Charlie's task. This would help isolate other-elicited affective change during observation and improve interpretability of within-child versus other-self correspondence patterns.

Beyond improving the existing VR-based setting, future research should also consider alternative VR scenarios that are more directly empathy-eliciting than a performance stress task. School-relevant contexts such as peer exclusion, unfairness or bullying-like episodes may evoke clearer other-oriented reactions and are closely aligned with children's everyday social experiences.

Conclusion

The present study examined whether a VR-based classroom paradigm can capture children's empathic responding in an interpretable way and whether VR-based indicators are sensitive to pre-post change associated with an MHL program. The VR scenario elicited meaningful attributions and explanations of emotions in children's open-ended responses, but the correspondence indicators showed no differential pre-post change between groups and did not converge with qualitative emotion identification or affective resonance.

A central contribution of the study is therefore conceptual rather than confirmatory. The lack of convergence between VR match indicators and qualitative emotion identification and affective resonance supports the argument that correspondence measures are not equivalent to empathy, as their interpretation requires independent component measures and careful attention to context. In the present paradigm, several boundary conditions likely constrained a match-based empathy signal, including weak subjective stress, competition-related framing, limited target readability and potential attentional carry-over from the child's own task. As a result, even when children described other-related reactions in their own words, this was not reliably reflected in the match outcomes. At the same time, the qualitative trends suggest that children did engage with the NPC like a real person whose emotions could be inferred and discussed, and that MHL-related content may be more likely to manifest in explicit emotion attribution and verbalized interpretations than in the VR correspondence indicators.

Overall, the present findings highlight both the promise and the interpretive constraints of VR-based empathy assessment in late childhood. VR appears to be a feasible, engaging and ecologically meaningful context for studying children's emotion understanding and perspective-related reflections, especially when complemented by open-ended responses. Future work should refine VR scenarios and measurement strategies to separate empathy-related mechanisms more explicitly, strengthen target readability and design contexts that more directly elicit other-oriented responding, thereby advancing VR as a rigorous tool for evaluating socio-emotional interventions in children.

References

Allemand, M., Steiger, A. E., & Fend, H. A. (2015). Empathy Development in Adolescence Predicts Social Competencies in Adulthood: Adolescent Empathy and Adult Outcomes. *Journal of Personality*, *83*(2), 229–241. https://doi.org/10.1111/jopy.12098

Bailey, J. O., & Bailenson, J. N. (2017). Considering virtual reality in children's lives. *Journal of Children and Media*, *11*(1), 107–113. https://doi.org/10.1080/17482798.2016.1268779

Bale, J., Grové, C., & Costello, S. (2020). Building a mental health literacy model and verbal scale for children: Results of a Delphi study. *Children and Youth Services Review*, *109*, 104667. https://doi.org/10.1016/j.childyouth.2019.104667

Barreda-Ángeles, M., Serra-Blasco, M., Trepat, E., Pereda-Baños, A., Pàmias, M., Palao, D., Goldberg, X., & Cardoner, N. (2021). Development and experimental validation of a dataset of 360°-videos for facilitating school-based bullying prevention programs. *Computers & Education*, *161*, 104065. https://doi.org/10.1016/j.compedu.2020.104065

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1). https://doi.org/10.18637/jss.v067.i01

Batson, C. D., Lishner, D. A., Cook, J., & Sawyer, S. (2005). Similarity and Nurturance: Two Possible Sources of Empathy for Strangers. *Basic and Applied Social Psychology*, *27*(1), 15–25. https://doi.org/10.1207/s15324834basp2701_2

Bayne, H. B., Swank, J. M., & Gage, N. (2024). A critical review of empathy assessments for children ages 3–12. *Psychology in the Schools*, *61*(2), 739–767. https://doi.org/10.1002/pits.23082

Bensalah, L., Stefaniak, N., Carre, A., & Besche-Richard, C. (2016). The Basic Empathy Scale adapted to French middle childhood: Structure and development of empathy.

Behavior Research Methods, *48*(4), 1410–1420. https://doi.org/10.3758/s13428-015-0650-8

Bird, G., & Viding, E. (2014). The self to other model of empathy: Providing a new framework for understanding empathy impairments in psychopathy, autism, and alexithymia. *Neuroscience & Biobehavioral Reviews*, *47*, 520–532. https://doi.org/10.1016/j.neubiorev.2014.09.021

Brown, C. L., West, T. V., Sanchez, A. H., & Mendes, W. B. (2021). Emotional Empathy in the Social Regulation of Distress: A Dyadic Approach. *Personality and Social Psychology Bulletin*, *47*(6), 1004–1019. https://doi.org/10.1177/0146167220953987

Bukowski, H., Boch, M., Lamm, C., & Silani, G. (2020). *Is Self-Other distinction malleable? Egocentric and altercentric biases in empathy are modulated by priming attachment style and similarity mindsets*. PsyArXiv. https://doi.org/10.31234/osf.io/bpyvz

Buske-Kirschbaum, A., Jobst, S., Wustmans, A., Kirschbaum, C., Rauh, W., & Hellhammer, D. (1997). Attenuated Free Cortisol Response to Psychosocial Stress in Children with Atopic Dermatitis: *Psychosomatic Medicine*, *59*(4), 419–426. https://doi.org/10.1097/00006842-199707000-00012

Cialdini, R. B., Brown, S. L., Lewis, B. P., Luce, C., & Neuberg, S. L. (1997). Reinterpreting the empathy–altruism relationship: When one into one equals oneness. *Journal of Personality and Social Psychology*, *73*(3), 481–494. https://doi.org/10.1037/0022-3514.73.3.481

Coll, M.-P., Viding, E., Rütgen, M., Silani, G., Lamm, C., Catmur, C., & Bird, G. (2017). Are we really measuring empathy? Proposal for a new measurement framework. *Neuroscience & Biobehavioral Reviews*, *83*, 132–139. https://doi.org/10.1016/j.neubiorev.2017.10.009

Cuff, B. M. P., Brown, S. J., Taylor, L., & Howat, D. J. (2016). Empathy: A Review of the Concept. *Emotion Review*, *8*(2), 144–153. https://doi.org/10.1177/1754073914558466 de Vignemont, F., & Singer, T. (2006). The empathic brain: How, when and why? *Trends in Cognitive Sciences*, *10*(10), 435–441. https://doi.org/10.1016/j.tics.2006.08.008

de Waal, F. B. M. (1996). *Good Natured: The Origins of Right and Wrong in Humans and Other Animals*. Harvard University Press.

Decety, J. (2010). The Neurodevelopment of Empathy in Humans. *Developmental Neuroscience*, *32*(4), 257–267. https://doi.org/10.1159/000317771

Decety, J. (2011a). Dissecting the Neural Mechanisms Mediating Empathy. *Emotion Review*, *3*(1), 92–108. https://doi.org/10.1177/1754073910374662

Decety, J. (2011b). The neuroevolution of empathy. *Annals of the New York Academy of Sciences*, *1231*(1), 35–45. https://doi.org/10.1111/j.1749-6632.2011.06027.x

Decety, J., & Jackson, P. L. (2004). The Functional Architecture of Human Empathy. *Behavioral and Cognitive Neuroscience Reviews*, *3*(2), 71–100. https://doi.org/10.1177/1534582304267187

Decety, J., & Meyer, M. (2008). From emotion resonance to empathic understanding: A social developmental neuroscience account. *Development and Psychopathology*, *20*(4), 1053–1080. https://doi.org/10.1017/S0954579408000503

Eisenberg, N., & Miller, P. A. (1987). The relation of empathy to prosocial and related behaviors. *Psychological Bulletin*, *101*(1), 91–119. https://doi.org/10.1037/0033-2909.101.1.91

Forbes, P. A., Aydogan, G., Braunstein, J., Todorova, B., Wagner, I. C., Lockwood, P. L., Apps, M. A., Ruff, C. C., & Lamm, C. (2024). Acute stress reduces effortful prosocial behaviour. *eLife*, *12*, RP87271. https://doi.org/10.7554/eLife.87271

Hartig, F. (2026). *DHARMa: Residual Diagnostics for Hierarchical (Multi-Level / Mixed) Regression Models.* R Package Version 0.4.7. https://github.com/florianhartig/dharma.

Hatfield, E., Cacioppo, J. T., & Rapson, R. L. (1993). Emotional Contagion. *Current Directions in Psychological Science*, *2*(3), 96–100. https://doi.org/10.1111/1467-8721.ep10770953

Hoffman, M. L. (2000). *Empathy and Moral Development: Implications for Caring and Justice* (1st edn). Cambridge University Press. https://doi.org/10.1017/CBO9780511805851

Ickes, W. (2016). Empathic accuracy: Judging thoughts and feelings. In J. A. Hall, M. Schmid Mast, & T. V. West (Eds), *The Social Psychology of Perceiving Others Accurately* (1st edn, pp. 52–70). Cambridge University Press. https://doi.org/10.1017/CBO9781316181959.003

Ickes, W., Stinson, L., Bissonnette, V., & Garcia, S. (1990). Naturalistic social cognition: Empathic accuracy in mixed-sex dyads. *Journal of Personality and Social Psychology*, *59*(4), 730–742. https://doi.org/10.1037/0022-3514.59.4.730

Imece, S., & Arslan Cansever, B. (2019). Investigation of the Relationship Between the Empathic Tendency Skills and Problem Solving Skills of Turkish Primary School Students. *International Journal of Progressive Education*, *15*(5), 19–31. https://doi.org/10.29329/ijpe.2019.212.2

Jorm, A. F. (2000). Mental health literacy: Public knowledge and beliefs about mental disorders. *British Journal of Psychiatry*, *177*(5), 396–401. https://doi.org/10.1192/bjp.177.5.396

Jorm, A. F., Korten, A. E., Jacomb, P. A., Christensen, H., Rodgers, B., & Pollitt, P. (1997). "Mental health literacy": A survey of the public's ability to recognise mental disorders and their beliefs about the effectiveness of treatment. *Medical Journal of Australia*, *166*(4), 182–186. https://doi.org/10.5694/j.1326-5377.1997.tb140071.x

Kanske, P. (2018). The social mind: Disentangling affective and cognitive routes to understanding others. *Interdisciplinary Science Reviews*, *43*(2), 115–124. https://doi.org/10.1080/03080188.2018.1453243

Kanske, P., Böckler, A., Trautwein, F.-M., & Singer, T. (2015). Dissecting the social brain: Introducing the EmpaToM to reveal distinct neural networks and brain–behavior relations for empathy and Theory of Mind. *NeuroImage*, *122*, 6–19. https://doi.org/10.1016/j.neuroimage.2015.07.082

Kelly, O., Matheson, K., Martinez, A., Merali, Z., & Anisman, H. (2007). Psychosocial Stress Evoked by a Virtual Audience: Relation to Neuroendocrine Activity. *CyberPsychology & Behavior*, *10*(5), 655–662. https://doi.org/10.1089/cpb.2007.9973

Kirschbaum, C., Pirke, K.-M., & Hellhammer, D. H. (1993). The 'Trier Social Stress Test' – A Tool for Investigating Psychobiological Stress Responses in a Laboratory Setting. *Neuropsychobiology*, *28*(1–2), 76–81. https://doi.org/10.1159/000119004

Knafo, A., Zahn-Waxler, C., Van Hulle, C., Robinson, J. L., & Rhee, S. H. (2008). The developmental origins of a disposition toward empathy: Genetic and environmental contributions. *Emotion*, *8*(6), 737–752. https://doi.org/10.1037/a0014179

Krebs, D. (1975). Empathy and altruism. *Journal of Personality and Social Psychology*, *32*(6), 1134–1146. https://doi.org/10.1037/0022-3514.32.6.1134

Lamm, C., Batson, C. D., & Decety, J. (2007). The Neural Substrate of Human Empathy: Effects of Perspective-taking and Cognitive Appraisal. *Journal of Cognitive Neuroscience*, *19*(1), 42–58. https://doi.org/10.1162/jocn.2007.19.1.42

Lamm, C., & Majdandžić, J. (2015). The role of shared neural activations, mirror neurons, and morality in empathy – A critical comment. *Neuroscience Research*, *90*, 15–24. https://doi.org/10.1016/j.neures.2014.10.008

Lietz, C. A., Gerdes, K. E., Sun, F., Geiger, J. M., Wagaman, M. A., & Segal, E. A. (2011). The Empathy Assessment Index (EAI): A Confirmatory Factor Analysis of a

Multidimensional Model of Empathy. *Journal of the Society for Social Work and Research*, *2*(2), 104–124. https://doi.org/10.5243/jsswr.2011.6

Lüdecke, D., Ben-Shachar, M., Patil, I., Waggoner, P., & Makowski, D. (2021). performance: An R Package for Assessment, Comparison and Testing of Statistical Models. *Journal of Open Source Software*, *6*(60), 3139. https://doi.org/10.21105/joss.03139

Marheinecke, R., Blasberg, J., Heilmann, K., Imrie, H., Wesarg-Menzel, C., & Engert, V. (2025). Measuring empathic stress – A systematic review of methodology and practical considerations for future research. *Psychoneuroendocrinology*, *171*, 107216. https://doi.org/10.1016/j.psyneuen.2024.107216

Maxwell, B., & DesRoches, S. (2010). Empathy and social-emotional learning: Pitfalls and touchstones for school-based programs. *New Directions for Child and Adolescent Development*, *2010*(129), 33–53. https://doi.org/10.1002/cd.274

Mayring, P., & Fenzl, T. (2019). Qualitative Inhaltsanalyse. In N. Baur & J. Blasius (Eds), *Handbuch Methoden der empirischen Sozialforschung* (pp. 633–648). Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-21308-4_42

McCreary, J. J., & Marchant, G. J. (2017). Reading and Empathy. *Reading Psychology*, *38*(2), 182–202. https://doi.org/10.1080/02702711.2016.1245690

Morasse, F., Vera-Estay, E., & Beauchamp, M. H. (2021). Using virtual reality to optimize assessment of sociomoral skills. *Virtual Reality*, *25*(1), 123–132. https://doi.org/10.1007/s10055-020-00443-9

Muravevskaia, E., & Gardner-McCune, C. (2022). Social Presence in VR Empathy Game for Children: Empathic Interaction with the Virtual Characters. *2022 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, 742–743. https://doi.org/10.1109/VRW55335.2022.00222

Neary, P. (2023). Questionnaire Measures of Empathy in Children: A Scoping Review. *Assessment (Odessa, Fla.)*, *30*(3), 798–824. https://doi.org/10.1177/10731911211069677

Nitschke, J. P., & Bartz, J. A. (2023). The association between acute stress & empathy: A systematic literature review. *Neuroscience & Biobehavioral Reviews*, *144*, 105003. https://doi.org/10.1016/j.neubiorev.2022.105003

Owais, W. B., & Yaacoub, E. (2020). Quantifying Empathy in Virtual Reality: An Outline. *2020 IEEE International Conference on Informatics, IoT, and Enabling Technologies (ICIoT)*, 457–462. https://doi.org/10.1109/ICIoT48696.2020.9089565

Özdemir Beceren, B., & Özdemir, A. A. (2019). Role of Temperament Traits and Empathy Skills of Preschool Children in Predicting Emotional Adjustment. *International Journal of Progressive Education*, *15*(3), 91–107. https://doi.org/10.29329/ijpe.2019.193.7

Parsons, T. D., & Phillips, A. S. (2016). Virtual reality for psychological assessment in clinical practice. *Practice Innovations*, *1*(3), 197–217. https://doi.org/10.1037/pri0000028

Parsons, T. D., Riva, G., Parsons, S., Mantovani, F., Newbutt, N., Lin, L., Venturini, E., & Hall, T. (2017). Virtual Reality in Pediatric Psychology. *Pediatrics*, *140*(Supplement_2), S86–S91. https://doi.org/10.1542/peds.2016-1758I

Patel, V., Flisher, A. J., Hetrick, S., & McGorry, P. (2007). Mental health of young people: A global public-health challenge. *The Lancet*, *369*(9569), 1302–1313. https://doi.org/10.1016/S0140-6736(07)60368-7

Preston, S. D. (2007). A perception-action model for empathy. In T. F. D. Farrow & P. W. R. Woodruff (Eds), *Empathy in Mental Illness* (1st edn, pp. 428–447). Cambridge University Press. https://doi.org/10.1017/CBO9780511543753.024

Richaud, M. C., Lemos, V. N., Mesurado, B., & Oros, L. (2017). Construct Validity and Reliability of a New Spanish Empathy Questionnaire for Children and Early

Adolescents. *Frontiers in Psychology*, *8*, 979. https://doi.org/10.3389/fpsyg.2017.00979

Sartori, R., & Pasini, M. (2007). Quality and Quantity in Test Validity: How can we be Sure that Psychological Tests Measure what they have to? *Quality & Quantity*, *41*(3), 359–374. https://doi.org/10.1007/s11135-006-9006-x

Schonert-Reichl, K. (2013). Promoting Empathy in School-Aged Children. In K. Nader (Ed.), *School Rampage Shootings and Other Youth Disturbances* (1st edn, pp. 159–203). Routledge. https://doi.org/10.4324/9780203855454-9

Seddon, J. A., Rodriguez, V. J., Provencher, Y., Raftery-Helmer, J., Hersh, J., Labelle, P. R., & Thomassin, K. (2020). Meta-analysis of the effectiveness of the Trier Social Stress Test in eliciting physiological stress responses in children and adolescents. *Psychoneuroendocrinology*, *116*, 104582. https://doi.org/10.1016/j.psyneuen.2020.104582

Sesso, G., Brancati, G. E., Fantozzi, P., Inguaggiato, E., Milone, A., & Masi, G. (2021). Measures of empathy in children and adolescents: A systematic review of questionnaires. *World Journal of Psychiatry*, *11*(10), 876–896. https://doi.org/10.5498/wjp.v11.i10.876

Silani, G., Lamm, C., Ruff, C. C., & Singer, T. (2013). Right Supramarginal Gyrus Is Crucial to Overcome Emotional Egocentricity Bias in Social Judgments. *The Journal of Neuroscience*, *33*(39), 15466–15476. https://doi.org/10.1523/JNEUROSCI.1488-13.2013

Singer, T. (2006). The neuronal basis and ontogeny of empathy and mind reading: Review of literature and implications for future research. *Neuroscience & Biobehavioral Reviews*, *30*(6), 855–863. https://doi.org/10.1016/j.neubiorev.2006.06.011

Singer, T. (2025). A neuroscience perspective on the plasticity of the social and relational brain. *Annals of the New York Academy of Sciences*, *1547*(1), 52–74. https://doi.org/10.1111/nyas.15319

Singer, T., Seymour, B., O'Doherty, J. P., Stephan, K. E., Dolan, R. J., & Frith, C. D. (2006). Empathic neural responses are modulated by the perceived fairness of others. *Nature*, *439*(7075), 466–469. https://doi.org/10.1038/nature04271

Somarathna, R., Bednarz, T., & Mohammadi, G. (2023). Virtual Reality for Emotion Elicitation – A Review. *IEEE Transactions on Affective Computing*, *14*(4), 2626–2645. https://doi.org/10.1109/TAFFC.2022.3181053

Ştefan, C. A., & Avram, J. (2018). The multifaceted role of attachment during preschool: Moderator of its indirect effect on empathy through emotion regulation. *Early Child Development and Care*, *188*(1), 62–76. https://doi.org/10.1080/03004430.2016.1246447

Steinbeis, N. (2016). The role of self–other distinction in understanding others' mental and emotional states: Neurocognitive mechanisms in children and adults. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *371*(1686), 20150074. https://doi.org/10.1098/rstb.2015.0074

Steinbeis, N., Bernhardt, B. C., & Singer, T. (2015). Age-related differences in function and structure of rSMG and reduced functional connectivity with DLPFC explains heightened emotional egocentricity bias in childhood. *Social Cognitive and Affective Neuroscience*, *10*(2), 302–310. https://doi.org/10.1093/scan/nsu057

Webster-Stratton, C., & Reid, M. J. (2004). Strengthening Social and Emotional Competence in Young Children—The Foundation for Early School Readiness and Success: Incredible Years Classroom Social Skills and Problem-Solving Curriculum. *Infants & Young Children*, *17*(2), 96–113. https://doi.org/10.1097/00001163-200404000-00002

Wellman, H. M., Cross, D., & Watson, J. (2001). Meta-Analysis of Theory-of-Mind Development: The Truth about False Belief. *Child Development*, *72*(3), 655–684. https://doi.org/10.1111/1467-8624.00304

Wu, X., Yao, S.-C., Lu, X.-J., Zhou, Y.-Q., Kong, Y.-Z., & Hu, L. (2024). Categories of training to improve empathy: A systematic review and meta-analysis. *Psychological Bulletin*, *150*(10), 1237–1260. https://doi.org/10.1037/bul0000453

Zaki, J. (2014). Empathy: A motivated account. *Psychological Bulletin*, *140*(6), 1608–1647. https://doi.org/10.1037/a0037679

List of Figures

List of Tables

Appendix A

Abstract

Empathy is important in primary school, when socio-emotional abilities are still developing and peer interactions are frequent. Empathy is associated with beneficial developmental outcomes, suggesting benefits of early fostering. Yet measuring empathic responding in children remains challenging. Immersive virtual reality (VR) can present standardized social situations and may support the evaluation of school-based prevention programs. The present study pursued two linked aims: (1) as a methodological proof-of-concept, to examine whether a VR classroom paradigm yields interpretable correspondence indicators of empathic responding, and (2) to test whether participation in a mental health literacy program, targeting recognition, classification and management of emotions, is associated with pre-post changes in these indicators compared to a control group. In a quasi-experimental pre-post design, 174 primary school children (intervention n=85; control n=89) completed a VR task in which they first performed a challenging task and then observed a peer virtual character facing the same situation. Correspondence indicators were derived from children's ratings of the peer's emotion and their own. Open-ended responses were coded into predefined categories to support interpretation. Results showed no evidence for a differential pre-post change between groups on the correspondence indicators. Qualitative patterns suggested descriptively stronger increases in empathic responding over time in the intervention group but showed limited convergence with the VR outcomes. Findings indicate that VR is feasible for eliciting children's attributions and explanations while correspondence indicators have interpretative limits in the present paradigm. Future research should refine VR paradigms and measurement strategies to increase sensitivity to intervention-related change.

Keywords: empathy, virtual reality, children, mental health literacy, mixed methods

Zusammenfassung

Empathie ist im Volksschulalter besonders relevant, da sich sozial-emotionale Fähigkeiten in dieser Phase weiterentwickeln und Peer-Interaktionen im Schulalltag häufig sind. Empathie ist zudem mit günstigen Entwicklungsoutcomes assoziiert, was eine frühe Förderung dieser nahelegt. Gleichzeitig bleibt die Erfassung empathischen Reagierens bei Kindern methodisch herausfordernd. Immersive Virtual Reality (VR) kann standardisierte soziale Situationen abbilden und damit die Evaluation schulbasierter Präventionsprogramme unterstützen. Die vorliegende Studie verfolgte zwei zusammenhängende Ziele: (1) als methodisches Proof-of-Concept zu untersuchen, ob ein VR-Klassenzimmerparadigma interpretierbare, korrespondenzbasierte Indikatoren empathischen Reagierens liefert, und (2) zu untersuchen, ob die Teilnahme an einem Mental-Health-Literacy-Programm, mit dem Fokus Emotionen zu erkennen, einzuordnen und damit umzugehen, mit Prä-Post-Veränderungen dieser Indikatoren im Vergleich zu einer Kontrollgruppe verbunden ist. In einem quasi-experimentellen Prä-Post--Design absolvierten 174 Volksschulkinder (Interventionsgruppe n=85; Kontrollgruppe n=89) eine VR-Aufgabe, bei der sie zunächst selbst eine herausfordernde Aufgabe bearbeiteten und anschließend ein virtuelles Kind in derselben Situation beobachteten. Korrespondenzindikatoren wurden aus den Ratings zu den Emotionen des anderen Kindes und den eigenen Emotionsratings abgeleitet. Offene Antworten wurden zur Unterstützung der Interpretation in vorab definierte Kategorien kodiert. Es zeigte sich keine Evidenz für eine differentielle Prä-Post-Veränderung zwischen den Gruppen in den Korrespondenzindikatoren. Qualitative Muster deuteten deskriptiv auf eine stärkere Zunahme empathischen Reagierens in der Interventionsgruppe hin, konvergierten jedoch nur begrenzt mit den VR-Indikatoren. Insgesamt spricht dies für die Machbarkeit von VR zur Anregung von Emotionszuschreibungen und Begründungen, bei gleichzeitigen interpretativen Grenzen korrespondenzbasierter Indikatoren im vorliegenden Paradigma. Zukünftige Forschung sollte VR-Paradigmen und Messstrategien weiter verfeinern, um Sensitivität für interventionsbezogene Veränderungen zu erhöhen.

Schlüsselbegriffe: Empathie, Virtual Reality, Kinder, Mental Health Literacy, Mixed Methods

Appendix B: Additional Tables

Table B1

Prevalence of Affective Resonance and Emotion Identification by Condition and Time.

Condition	Time	N	Affective resonance, n (%)	Emotion identification, n (%)
Control	Pre	88	23 (26.1%)	78 (88.6%)
Control	Post	78	25 (32.1%)	73 (93.6%)
Intervention	Pre	85	21 (24.7%)	71 (83.5%)
Intervention	Post	75	28 (37.3%)	69 (92.0%)

Note. Cells show n (valid %), i.e., percentages are based on non-missing responses within each condition and time.

Table B2

Affective Resonance (AR) by Other–self Match (Pre- and Post-Intervention).

Other–self Match	AR = 0 Pre	AR = 1 Pre	AR = 0 Post	AR = 1 Post
0	84 (73.7%)	30 (26.3%)	73 (65.8%)	38 (34.2%)
1	45 (76.3%)	14 (23.7%)	27 (64.3%)	15 (35.7%)

Note. Cells show n (row %). Percentages are computed within each row separately for pre- and post-intervention. Missing cases were excluded.

Table B3

Emotion Identification (EI) by Other–self Match (Pre- and Post-Intervention).

Other–self Match	EI = 0 Pre	EI = 1 Pre	EI = 0 Post	EI = 1 Post
0	11 (9.6%)	103 (90.4%)	7 (6.3%)	104 (93.7%)
1	13 (22.0%)	46 (78.0%)	4 (9.5%)	38 (90.5%)

Note. Cells show n (row %). Percentages are computed within each row separately for pre- and post-intervention. Missing cases were excluded.

Table B4

Affective Resonance (AR) by Within-child Match (Pre- and Post-Intervention).

Within-child Match	AR = 0 Pre	AR = 1 Pre	AR = 0 Post	AR =1 Post
0	47 (72.3%)	18 (27.7%)	42 (62.7%)	25 (37.3%)
1	82 (75.9%)	26 (24.1%)	58 (67.4%)	28 (32.6%)

Note. Cells show n (row %). Percentages are computed within each row separately for pre- and post-intervention. Missing cases were excluded.

Table B5

Affective Resonance by Gender and Time.

Gender	Time	Affective Resonance = 0	Affective Resonance = 1
Boy	Pre	64 (73.6%)	23 (26.4%)
Boy	Post	55 (70.5%)	23 (29.5%)
Girl	Pre	65 (75.6%)	21 (24.4%)
Girl	Post	45 (60.0%)	30 (40.0%)

Note. Cells show n (row %). Cases with missing data were excluded

Table *B6*

Emotion Identification by Gender and Time.

Gender	Time	Emotion Identification = 0	Emotion Identification = 1
Boy	Pre	11 (12.6%)	76 (87.4%)
Boy	Post	5 (6.4%)	73 (93.6%)
Girl	Pre	13 (15.1%)	73 (84.9%)
Girl	Post	6 (8.0%)	69 (92.0%)

Note. Cells show n (row %). Cases with missing data were excluded.

Table B7

Competition and Stress by Condition and Time.

Condition	Time	Competition, n (%)	Stress, n (%)
Control	Pre	13 (14.8%)	12 (13.6%)
Control	Post	9 (11.7%)	13 (16.7%)
Intervention	Pre	14 (16.5%)	11 (12.9%)
Intervention	Post	7 (9.3%)	9 (12.0%)

Note. Cells show n (valid %), i.e., percentages are based on non-missing responses within each group and time.

Appendix C: Codebook Qualitative Data

Multiple categories per answer can be assigned simultaneously. Coding is based on content, not just individual words. Translations are provided with a focus on conveying meaning rather than preserving literal phrasing.

Questions:

(1) How did you feel during the task? Why? (2) How did you feel while watching Charlie? Why? (3) How do you think Charlie was feeling? Why?

Category	**Definition**	**Anchor Examples**	**Coding Rules**
Affective Resonance	The child verbalizes the own affective response during the observation of Charlie and it is congruent in valence with the emotion attributed to Charlie.	"sad because he didn't do so well" "sad for her, because she made so many mistakes" "a little anxious, but I don't know why"[1]	There must be a verbalization of the own affect and an attribution of Charlie's emotion. They have to be the same valence. Explicit causal attribution of the child's affect to Charlie is not required
Emotion Identification	The child attributes an emotion to Charlie and provides at least a minimal justification for this attribution. Justification may be situation-based, dispositional or speculative.	"nervous because the teacher stressed him out" "anxious, because of many mistakes" "proud because he finished the task"	Attributions are coded as emotion identification regardless of whether the justification corresponds to situational cues or are objectively accurate.
Prosocial Behavior	The child describes supportive actions or intentions toward Charlie	"Charlie made mistakes much more often. I always showed him what to do, but he wasn't that smart" "If she were real and not animated, I would be happy to tutor her, because for some it's just more difficult"	The child explicitly states that they wanted to or would help, support or defend Charlie.
Indifference	The child does not show any affective or	"neutral, I didn't care about him"	They report feeling neutral, normal, relaxed, good or bored while

	cognitive engagement with Charlie's situation across their answers.	"bored, because there was nothing to do"	watching Charlie and do not provide a meaningful description of Charlie's feelings or perspective. Indifference is only coded when no empathy-related category applies
Amusement	The Child reports finding Charlie's situation funny, laughs at his mistakes, is entertained, without any sympathy. Focus is on having fun at Charlie's expense.	"funny, because Charlie made a lot of mistakes. The mistakes were absurd" "funny because she looked and moved weirdly"	The child states that watching Charlie was funny/amusing and explicitly links this to Charlie's mistakes or behavior, without concern for Charlie. General enjoyment of the task without reference to Charlie or shared joy for Charlie is not coded as Amusement.
Competition	The child frames the situation primarily as performance comparison or game	"I felt like I was the best because I only made one mistake" "proud because I was better than Charlie"	The child refers to a performance comparison between self and Charlie (e.g. better, worse, more/fewer mistakes, etc.).
Stress	The child reports feeling tense, pressured, worries or otherwise negatively aroused in response to the test situation. The focus is on psychological strain due to task demands, evaluation or uncertainty.	"I was so nervous during the task" "test was stressful because you had to do it so fast" "as if it were unreal, unfamiliar, sometimes I mistyped, I was annoyed and nervous"	The child reports negative high arousal and links it to the task situation; "excited" is coded as stress only if it is framed as nervousness/worry/performance pressure; "exhausting" is coded as stress if it refers to being emotionally overloaded; "annoying" without any indication of tension, overload or pressure is not coded as stress

[1]*Note*: The examples only refer to responses about the own affect. The prerequisite that they correspond to Charlie's valence is given in these explicit examples.

Erklärung zur Verwendung von Künstlicher Intelligenz (KI)

Hiermit erkläre ich, dass ich die folgenden KI-Hilfsmittel zur Erstellung meiner Masterarbeit mit dem Titel „Assessing Children's Empathic Responding in Virtual Reality: A Mixed-Methods Pre-Post Study Following Mental Health Literacy Training" verwendet habe. Ich versichere, dass die Kennzeichnung der verwendeten KI-Hilfsmittel vollständig ist, inkl. Verwendungszweck, Produktname, als auch ggfs. Prompts und Outputs des KI-Hilfsmittels. Ich verantworte die Auswahl und die Übernahme sämtlicher Ergebnisse der von mir verwendeten KI-Hilfsmittel vollumfänglich selbst.

Verwendungszweck*	**Produktname des KI-Hilfsmittels**	**Ggfs. Prompt (Input) und Output des KI-Hilfsmittels**
R Code Erstellung	ChatGPT	

* Hilfsmittel zur Überprüfung von Grammatik und Rechtschreibung oder der automatisierten Erstellung des Literaturverzeichnisses sind von dieser Erklärung ausgenommen.