OXFORD

Interacting with Computers, 2025, 1–16
https://doi.org/10.1093/iwc/iwaf016
Article

Using the Smileyometer to measure UX with children

Janet C. Read* and Matt Horton

University of Central Lancashire Preston, UK
*Corresponding author: jcread@uclan.ac.uk

Since 2002, the Smileyometer has been much used for measuring UX with children, and limitations, extensions, and idiosyncrasies of it have been reported. We summarise this body of work drawing out the interesting observations and commentaries. Adapting the Smileyometer for small children, considering the effect on scores when rating a sequence of activities or products, and understanding how children might discriminate between products or services are three emerging themes that we examine in case studies. These studies show that adapting the Smileyometer for use with young children is possible, that an order effect can exist when rating items one after the other and this merits further investigation and that the tool does facilitate discrimination. We conclude with three guidelines to assist researchers in getting the best out of the tool by considering preparation, completion, and reporting when choosing the Smileyometer as a tool.

RESEARCH HIGHLIGHTS

- A review of the use of the Smileyometer over 20 years highlighting its use in HCI
- Tips to expand its use to small children with new ideas on how to present the tool
- Examples of new areas for study with regard to children measuring their UX

Keywords: *child–computer interaction; UX evaluation; methods; fun; Smileyometer.*

1 INTRODUCTION

First introduced as a Visual Analogue Scale (VAS) drawn by children (Read *et al.*, 2002), the Smileyometer is one of a suite of tools referred to as the Fun Toolkit (Read, 2008, Read *et al.*, 2002). The historical positioning of the Fun Toolkit and the Smileyometer was to apply the ISO9241 usability metrics of effectiveness, efficiency, and satisfaction to a within-subjects study and to expand and test a limited set of tools for measuring satisfaction with children. In the ISO definition of usability, satisfaction is positioned in relationship to the achievement of goals, which is to say that a user is expected to be satisfied with a product if it meets their goals. The substitution of fun for satisfaction when considering usability and UX with children can be rationalised in two ways, firstly that fun is both a design requirement as well as a goal state for children's technology and also that, with usability more generally looking towards UX, the inclusion of fun within the concept of usability has been promoted (Carroll, 2004).

Fun became a focus of attention in HCI with the rise of computer games and other pleasure-related IT products. Within a HCI context, Carroll (2004) writes, on page 38, that "Things are fun when they attract, capture, and hold our attention by provoking new or unusual emotions in contexts that typically arouse none, or arousing emotions not typically aroused in a given context." This definition rather positions fun as surprise, which may indeed be how an adult might perceive fun; however, for children fun is possibly more aligned to the way that Draper (1999) describes it—in terms of playing for pleasure, relating to activities done for their own sake with freedom of choice, and therefore much less goal-oriented.

The primary publication that referenced the Smileyometer described how fun was a potential substitute for satisfaction with children and how pictures in a VAS needed to be child-friendly (Read *et al.*, 2002). The same publication described the components of the Fun Toolkit and encouraged adopters to use multiple measures that could be amalgamated to measure fun. As well as the Smileyometer, the Fun Toolkit is made up of the Again Again table, which is a simple table that asks the child "would you like to do this again?" and the Fun Sorter. The Again Again table comes with three possible responses (Yes, Maybe, No) and usually a child ticks one response. The Fun Sorter is an ordinal tool that is only used when comparing experiences; children order their experiences aligned to constructs like "Most fun" or "Easiest to learn"—sometimes using pictures for younger children. In the historical positioning of the Fun Toolkit, two later papers Read (2008, 2012) validate the Fun Toolkit by demonstrating how the different tools could be used together while showing how the results from the different tools correlate; these papers conclude with some observations on how to best use the toolkit and the Smileyometer. One of these observations was that in most instances the Smileyometer should be shown to children both before and after engagement with the artefact or experience being evaluated. The rationale for this is that experienced and anticipated fun are both important to capture and that it is generally preferable for a child to have a better experience than he/she imagined they would have.

Received: June 25, 2024. **Revised:** February 02, 2025. **Accepted:** February 26, 2025

FIGURE 1. The Smileyometer.

In early use of the Smileyometer, research papers on children were relatively rare—the Interaction Design and Children (IDC) community had only just started to form and venues like CHI Play and specialist journals like IJCCI had not yet developed. Over the subsequent fifteen or so years, the Smileyometer has regularly appeared in academic papers when studying, designing, or evaluating interactive technology and experiences, with children. While most of these have evaluated technical products, the Smileyometer has also been used within the HCI community to evaluate experiences where there is minimal or no technology (Benton *et al.*, 2012, Lochtefeld *et al.*, 2022) and has been seen used outside the community in many papers including to evaluate board games for environmental education (Mostowfi *et al.*, 2016), rating a sports science related physical obstacle course for 3–6 year old children (Klingberg *et al.*, 2019), participation in events with autistic children (Simpson *et al.*, 2022), and experience of flow in a language curriculum in a controlled study with 6 and 7 year olds (Argyriadi & Sotiropoulou-Zormpala, 2017).

Alongside this, several papers have challenged the effectiveness of the Smileyometer and proposed different methods and tools to measure children's UX (Dietz *et al.*, 2020, Hall *et al.*, 2016, Yusoff *et al.*, 2011, Zaman & Abeele, 2010).

This present paper looks at the use of the Smileyometer within HCI and IDC. An analysis of use in ACM venues and in the IWC journal suggests a set of questions and observations from its application across many contexts and situations. Selected questions are then explored in two case studies and these lead to updated guidelines for using the Smileyometer in studies with children.

1.1 The Smileyometer

As outlined above, the Smileyometer developed over time as a VAS with labels and smiley faces that were drawn by children with the midpoint being not exactly neutral. The main reason for this was that a traditionally drawn "mid-smile" would in fact be a straight line mouth and this is generally interpreted as being slightly negative and as showing irritation or boredom; it was interpreted in that way by children hence a small "up smile" was used for the mid point. The implication of this is that it could be argued the child has three positive emotions to choose from; however, the position and labelling of the mid face generally situates it as half way (Read *et al.*, 2002).

In the first description of the Smileyometer (Read *et al.*, 2002), it was proposed for use before and after engagement with an activity, technology, or task; this can then measure anticipated or expected fun as well as experienced fun. This supports expectation confirmation theory (Bhattacherjee & Premkumar, 2004, Oliver, 1980) enabling the researcher to be able to see if the experience outperforms the expectation and is preferred to just a post use measure. The 2002 paper described three studies, one with 16 children aged 6–9, one with 45 children aged 7 and 8, and one with 55 children aged 8/9 and 9/10. These early studies gave three main insights:

1. With the younger ages there was a large tendency to score as *Brilliant* the activity or technology

2. Comparing before and after use, around 60% of children chose the top scores both before and after and almost 80% of children moved than one point from anticipated (before) fun to experienced (after) fun; suggesting that most children got roughly what they expected.

3. Applying a numeric score to the five smiley faces, an arithmetic average (applied with a caveat) showed that younger children had a higher mean average (more overall reported satisfaction) than older children.

In a later study (Read & MacFarlane, 2006), 47 children aged 7/9 and 26 aged 12/13 each graded a set of online games using the Smileyometer after they had played. This study showed that older children were less likely to grade everything with the same score (<10%) than the younger children (>40%), suggesting that the older a child is, the more able he or she is to discriminate between different experiences. In this, as in the earlier study, younger children had a higher numeric mean score (3.86), than older children (3.49). The paper also explored how the other Fun Toolkit tools worked together and concluded with a set of guidelines for surveys with children while also suggesting that, although the Smileyometer on its own was a good tool for older children, its use was enhanced by also using another tool to provide better validation. In Read's work (Read, 2012), age effects were explored in studies that introduced three different products/experiences to children and began with a prediction of how they might be scored based on expert (albeit adult) judgment. In this study, 24 younger children and 48 older children looked at products as shown in Table 1, and this study demonstrated that children were discriminating, and it was noted that the averages did reflect what was predicted by the adult evaluators. As with earlier studies, the findings once again showed that, in a matched task, younger children tended to score things higher on the scale, with ratings skewed toward the upper end of the Smileyometer.

Since the 2012 study, there has been little research examining how the Smileyometer specifically performs; however, it has been cited in over 500 academic papers, and it has been used in different variations across continents and contexts. Lehnert *et al.* (2020) examined how children's responses on Smileyometers vary according to the way an experimenter interacts with the scale, finding that women's voices seemed to limit extreme positive ratings. Zaman *et al.* (2013) carried out an evaluation of the Smileyometer and the This or That (Zaman & Vanden Abeele, 2007) method and reported that the This or That method outperformed the Smileyometer with preschoolers. A limitation of this particular study is that it did not use any Smileyometers before play (as is recommended) and the two games compared were very similar. The end result, that children scored both games in a similar way, was unexpected.

Important questions, given the small number of studies exploring the Smileyometer's efficacy follow: How is the community using the Smileyometer and what has been learned? What needs to be studied to better understand if and where the Smileyometer can be used with confidence?

1.2 The Smileyometer in the Literature

Given that the Smileyometer is a "brand" searching for it in the literature did not require complex search terms. A single word search for "*Smileyometer*" in the ACM DL, of full text articles, revealed 140 results. This was considered a good sample for exploring how the community was using the Smileyometer but given the publication venue for this paper we also searched in the same way within IWC papers where a further 8 papers were found. Thirteen of the ACM DL results were proceedings so these

TABLE 1. Results from Read (2012) showing some discrimination.

	Scores Children aged 6/7			Scores Children aged 13/14	
Interface	Anticipated	Reported	Interface	Anticipated	Reported
A1	high, high	4.8, 4.8	A2	high, high	4.3, 4.6
B1	high, med	3.7, 3.5	B2	high, med	3.9, 4.0
C1	med, med	4.4, 3.6	C2	med, med	3.3, 3.4

TABLE 2. Categories of papers found in the literature.

	Number	Examples
Smileyometer and other FT products	18	MacFarlane *et al.* (2005), Leite *et al.* (2017), Zhang-Kennedy & Chiasson (2016), Graßl & Fraser (2023), Dijk *et al.* (2012), Sim *et al.* (2016b), Garcia-Sanjuan *et al.* (2016), Hernandez-Lara *et al.* (2023) Zhang *et al.* (2021), Xu *et al.* (2023). Greifenstein *et al.* (2022), Jurdi *et al.* (2018), Sim & Read (2024), Leite & Lehman (2016), Lochtefeld *et al.* (2022), Dawidowsky *et al.* (2021), Cesário *et al.* (2017), Chu *et al.* (2015)
Smileyometer before and after	8	Kuhn *et al.* (2009), Fowler (2017), Ferraz *et al.* (2016), Ferraz *et al.* (2010), Sargeant & Mueller (2018), Sim & Cassidy (2013), Sim *et al.* (2013), Cosentino *et al.* (2023)
Smileyometer after	54	Note—only a subset are listed here: Cibrian *et al.* (2021), Hastie *et al.* (2016), Chen *et al.* (2019b), Deshmukh *et al.* (2016), Al-Dawsari & Hendley (2024)
Adapted Smileyometer	7	Lagerstam *et al.* (2012), Delden *et al.* (2020), Oberhuber *et al.* (2017a), Salian & Sim (2014), Fowler (2019), Hyde *et al.* (2014), Alghabban & Hendley (2023)

were discounted. The remaining 135 papers were downloaded for study. The first filtering was of papers that were by Read and were describing the studies already commented on (this removed three papers), then of non-English language papers (one removed that was in French) and then for any that were not referencing the Smileyometer (one referred to "*a smileyometer*" unrelated to the Smileyometer of interest.

The remaining 129 papers were each read and coded according to how (if at all) they had used the Smileyometer. Eighteen papers referenced a Smileyometer paper but did not study or use it (e.g., Ooi *et al.* (2016), Çorlu *et al.* (2017), Price & Pontual Falcão (2011), fourteen papers referred to the Smileyometer as a candidate for use, for example in doctoral consortium studies or future work, but did not use it (e.g., Fowler & Schreiber (2017), Schafer *et al.* (2013), Tzortzoglou (2023)) and ten introduced the Smileyometer in the literature but then went on to propose a new method for measuring UX with children (e.g., (Hall *et al.* (2016), Yusoff *et al.* (2011)). These "different method" papers are particularly informative as they typically aim to critique the Smileyometer before proposing something different; they will each be revisited in a later paragraph.

Eighty-seven papers used some version of the Smileyometer in some way. This use can be considered in a hierarchy (see Table 2) with the top tier being those who used the Smileyometer along with other elements from the Fun Toolkit (FT), followed by papers that used the Smileyometer before and after an experience. The bulk of the papers examined only used the Smileyometer as an after-experience measure or used a version of it as a response-catcher with other tools. In some cases, the Smileyometer was significantly altered (adapted), and those papers are separately counted here. It is important to point out that there are many small alterations (different pictures used, different labelling, etc.) seen across many papers; some of these are discussed in the following narrative.

The 87 papers that used the Smileyometer also varied according to the children they worked with (see Table 3). It is worth pointing out here that two studies (Keskinen *et al.*, 2012, and Hernandez-Lara *et al.*, 2023) used Smileyometers with disabled adults—these articles are not included in this count of children, but are included in the 87 total. To look at ages of children, five categories were chosen and papers were distributed between these on the basis of closest fit. Where a paper had a large age range (e.g., 4–10), it was counted in more than one category, and the numbers mentioned in the paper were shared between categories. This gives an approximation in terms of the number of studies done with each age group and the number of children, and averages per study, in each age range.

These data confirm what has previously been anecdotally known; namely, that most research in Child–Computer Interaction is done with children aged 6 to 11, that the Smileyometer is infrequently used with older children and that there are relatively few studies using the Smileyometer with younger children. There are different explanations for this; one is that children under 6 years of age tend to be more difficult to access, so fewer studies are done with them anyway; another is that the efficacy of tools like the Fun Toolkit, for young children, has been contested (Read *et al.*, 2023, Yusoff *et al.*, 2011). On the latter point, this presumption has been used in papers as a reason to not use the Smileyometer (Wang *et al.*, 2019); we return to this later in this paper.

1.3 The Smileyometer in use

The papers using the Smileyometer, from the sample surveyed, had different motivations and took different approaches to presenting, describing, and analysing the data gathered.

1.3.1 *What is being evaluated*

The Smileyometer has been used to evaluate a wide range of products and services and experiences. From the 87 papers examined, the majority described the evaluation of a digital product or device but several described children's enjoyment of a method or a non-digital experience; these included Benton *et al.* (2012) that used the original Smileyometer faces, as response choices for 11 questions for children to rate a design session. Smileyometer faces were also used as response choices for an adapted computer science

TABLE 3. Ages of children using Smileyometers.

Ages	Number of studies	Total number of children	Average number of children per study	Example paper
3, 4, 5	13	171	14.3	Sargeant & Mueller (2018)
6, 7, 8	34	1034	31	Yarosh & Kwikkers (2011)
9, 10, 11	28	824	26.3	Jung *et al.* (2019)
12, 13, 14	10	141	15.6	Foster *et al.* (2014)
15+	0			

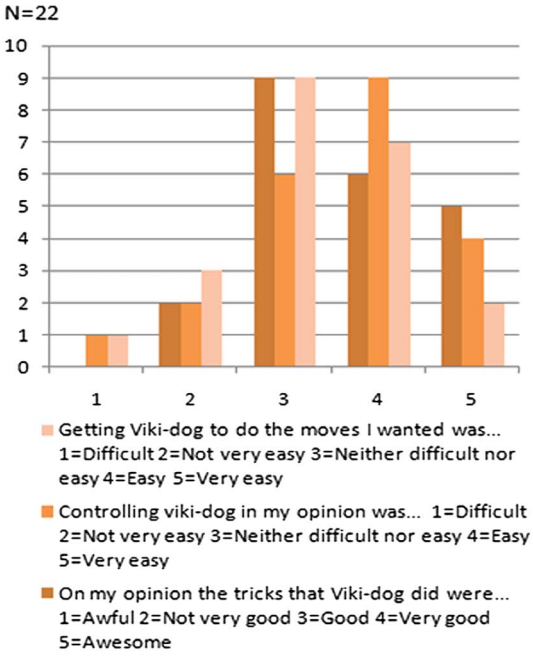

FIGURE 2. Smileyometer with different word labels after Lagerstam *et al.* (2012).

attitude survey to capture children's perceived changes in coding confidence (Fowler, 2017) and in a similar study, Greifenstein *et al.* (2022) used Smileyometers with other Fun Toolkit products to capture children's views on the help they were getting in programming workshops.

Products and digital experiences evaluated with Smileyometers have included museum technology (Cesário *et al.*, 2017, Jung *et al.*, 2019), iPad games (Bertou & Shahid, 2014), tangibles (Hijkoop *et al.*, 2020, Vonach *et al.*, 2016), storybooks (Sargeant & Mueller, 2018), robots (Ferraz *et al.*, 2016, Tsoi *et al.*, 2021), and educational products (Al-Dawsari & Hendley, 2024).

1.3.2 *How the Smileyometer is presented*

In presenting the Smileyometer to children, there were differences in procedure and in the pictorials. Some authors (e.g., Tsoi *et al.* (2021)) added colour to the smileys (Figure 3) or made the mid-smile neutral (Fowler, 2017), while others changed the word labels, with choices from 1–not at all to 5–very true Melniczuk & Vrapi (2023) and with a four-point scale from totally agree to totally disagree Oberhuber *et al.* (2017b)—see an example of an altered Smileyometer in Figure 2. The labels are important, and it is considered good practice to have labels as well as the faces (Borgers *et al.*, 2003). At least one paper had a translated Smileyometer, Godinez *et al.* (2017), and in one case the words were removed

FIGURE 3. An adapted Smileyometer from Tsoi *et al.* (2021).

entirely (Keskinen *et al.*, 2012) and the smiles were each presented on cards that could be selected.

Generally, the Smileyometer was presented with little fanfair. One exception was Leite *et al.* (2017) who worked with children aged 4–10 in three conditions (one control and two experimental) to explore interactions with a robot in a between-subjects study. In the initiation of the study the children practised the Smileyometer scales with three simple questions designed to elicit responses across the scale's range (*"How much do you like ice cream?", "How much do you like broccoli?"* and *"How do you feel when you stub your toe?"*). This seems like really good practice. In papers that describe how the Smileyometer is introduced, the general approach is to either ask a question verbally and ask the child to choose a response, or present a series of Smileyometer faces interspersed with other survey questions with written instructions. Examples of the former include from Leite *et al.* (2017) who asked *"How much did you like talking to Piper?"*, from Jurdi *et al.* (2018) who asked *"How much fun did you have with..."* and from Cosentino *et al.* (2023) with *"How easy was that...?"* Note that the questions asked often don't just ask about fun. In particular, where Smileyometers are being used as response catchers for many questions, there can be a lot of different things being captured on that single scale; Chu *et al.* (2015) captured children's self efficacy with questions like *"I am good at coming up with new ideas for stories"* and *"I have a lot of good ideas for stories"* being examples of statements that were then scored with a Smileyometer with an Agree/Disagree scale.

1.3.3 *How the Smileyometer data is presented and analysed*

Data from Smileyometers are presented in a variety of ways. In some instances, there is simply a narrative that the study used the Smileyometer and everyone was happy (e.g., Erfurt *et al.* (2019), Maqsood & Chiasson (2021)). Others just gave summary numeric data or percentages for the different scores (Ferraz *et al.*, 2017, Lochtefeld *et al.*, 2022). Most papers that say little about the Smileyometer scores tended to use the tool as an additional metric alongside what is often a performance study using logged or otherwise collected data. Where there is an enthusiasm for deeper analysis of self-reported engagement, in line with the suggestion by Read (2012) that arithmetic mean scores can show differences, means are often quoted (Bonner *et al.*, 2012, Duh *et al.*, 2010). In work by Xie *et al.* (2008), 132 children aged 8 and 9 compared physical tangible and graphical interfaces and a Smileyometer was used alongside the IMI (Intrinsic Motivation

TABLE 4. Using Smileyometers to differentiate.

	Examples
Between groups doing same thing	Children and adults both rated an experience with means for children (aged 5–12) being 4.6 for game enjoyment and 4.5 for design while adults rated 3.9 for experience and 4.8 for design. Vonach *et al.* (2016) Older and younger children's ratings were compared in a museum experience with older children scoring higher Dijk *et al.* (2012)
Before and After	In an exploration of five interaction modalities in Italy, using ANOVA after a Shapiro–Wilk test, significant differences were found for all questions when rated before and after Cosentino *et al.* (2023). When exploring different prototypes of games and rating them, means of 3.45 before and 3.83 after were shown to be significantly different in Sim & Read (2024)
Between products or experiences	In an evaluation in a museum with and without technology, using a Mann Whitney test, Cesário *et al.* (2017) showed a significant difference in favour of the technology. A repeated interaction study with three conditions and an ANOVA test showed significant differences on likeability Leite *et al.* (2017). In Dawidowsky *et al.* (2021), three conditions for reading fluency were compared with significant differences for fun, legibility and perceived speed all shown from Smileyometer data
No Differences	In several papers, while it was hoped that Smileyometers might show a difference, no difference was reported. Lochtefeld *et al.* (2022) compared two different tasks with resultant mean Smileyometer scores of 3.97 and 3.89, which did not show a significant difference; with most means above 4, but not explicitly given (derived from bar charts in the paper), Ahmad *et al.* (2016) used ANOVA tests on Smileyometer data from an evaluation of different robot experiences and found no significant differences.

Inventory) scale, coded from *"not at all true"* to *"very true"* to provide answers to the IMI questions. Mean scores and *SD*s were reported for children, with examples including means of 4.25, 4.26, and 4.32 for interest and enjoyment of the three interfaces to promote discussion. Non-parametric statistical tests are used in many papers and these are the statistical test of choice when comparing conditions and when looking at changes from before to after. As an example, when Smileyometers were used as response loggers with the E-learner Satisfaction questionnaire (Alghabban & Hendley, 2020, Wang, 2003), in a between-subjects study with nine-year-old children in two groups, means were quoted alongside a non-parametric test. In this study means of 4.90, *SD* = 0.18, and a median of 5, for the *"experimental"* group indicated a larger mean satisfaction score than from the control group (mean 4.68, *SD* = 0.38, median = 4.75), which was reported as a statistically significant difference between the overall satisfaction in the two conditions (Independent sample Mann–Whitney U test ($U = 277.5$, $p = .023$)).

1.3.4 *What the Smileyometer is delivering*

Most often, the Smileyometer is used to give a summative score on how much children enjoyed their activity. In these instances, some researchers asked the children to rate the activity before as well as after (Cosentino *et al.*, 2023, Garcia-Sanjuan *et al.*, 2016), whereas others had simply asked for an end score (Maqsood & Chiasson, 2021, Soleimani *et al.*, 2016).

Many papers used the Smileyometer as a response-catcher—this deviated slightly from the original idea of the Smileyometer as a fun gauge but does show its versatility. In these papers the scale was used to enable easy answering of questions by the children. Often, when used in this way, the Smileyometer faces were used with other tools; these included the IMI (McAuley *et al.*, 1989) scale (Cesário *et al.*, 2017, Dijk *et al.*, 2012, Xie *et al.*, 2008), the Godspeed (Bartneck *et al.*, 2009) survey Hastie *et al.* (2016), and the GEX (IJsselsteijn *et al.*, 2013) game experience questionnaire (Holz *et al.*, 2018).

Comparative studies typically presented either an experimental condition and a control condition or compared a couple of interaction modes or ideas where the Smileyometer was used to discriminate (Foster *et al.*, 2014, Graßl & Fraser, 2023, Jurdi *et al.*, 2018, Leite *et al.*, 2017, Park *et al.*, 2017, Sim *et al.*, 2016a). Given that one of the critiques of the Smileyometer is that younger children will skew towards high scores, and that therefore there may be little discrimination, it is important to see whether this is the case in published papers and to therefore understand the extent of this as a problem. Table 4 gives examples of both means and differentials across different study designs where comparisons were being explored.

1.4 Observations and Critiques of the Smileyometer

The method papers that are reported in this review each frame the development of their method as an "alternative" to other methods (including the Smileyometer) that are already out there. Table 5 shows that some of the new methods were clearly intended to measure new things while others were built on the basis of limitations of the Smileyometer. These included the limitations on age range (Dietz *et al.*, 2020, Yusoff *et al.*, 2011), the skew towards Brilliant (Dietz *et al.*, 2020, Fowler, 2013, Hall *et al.*, 2016), and the potential for poor, or scale limited, differentiation (Hall *et al.*, 2016, Sylla *et al.*, 2017).

Aside from in the "methods" papers, age related limitations of Smileyometers were addressed in part in several of the other papers that reported work with young children with some of these suggesting ideas on how to help children complete the Smileyometers. Sargeant & Mueller (2018) stated that 22 out of 26 children (aged 3, 4, 5) gave really good or brilliant as their scores when using the Smileyometer. The research team remarked, that while they know the Smileyometer is not recommended for the age group they are working with, they have used it before successfully and so continue to use it. Joly (2007), with five 4-year-old children, added stickers into the mix and reported that "*it was confirmed that when children had trouble to interact, their scores on the Smileyometer were low*", which suggests these children could differentiate between the images on the Smileyometer and also were mapping their choices to their experience; they report on how a small child rated their interaction as awful and concluded that the Smileyometer *"**should** be used with young children"*. In a later study from the same project, Leite *et al.* (2017) incorporated training and then successfully used the Smileyometer in homes. Given these positive reports, but also the *"fear"* of using something with young children when it was originally argued that young

TABLE 5. Alternative methods and their relationships to the Smileyometer.

Paper	Method	Tool measuring	Limitations of Smileyometer referenced
Hall *et al.* (2016)	5DG	Fun	The basis behind this tool is that there are two problems with the Smileyometer—one being that if comparing two versions and the first is rated highly then the second cannot go higher—even if it is better—the second critique is around the range of scores used by children. The new tool seemed to show the same skew as the Smileyometer when evaluated with children aged 9–11.
Yusoff *et al.* (2011)	Fun Semantic Differential Scale	Fun	The work of this author was with young children and the author took the statement by Read that the Smileyometer didn't work for small children so developed a scale that might work for small children.
Fowler (2013)	Gaze	Engagement	Had previously used Likert scales but noted that 80% of children put responses on anchor points so chose to explore gaze tracking to measure UX.
Dietz *et al.* (2020)	Giggle Gauge	Fun	Noted the tendency for extremes and skew with Smileyometers, the Giggle gauge aimed to solve this—also for younger children aimed to reduce the cognitive challenge of having to parse five images
Zaman & Abeele (2010)	Laddering	Preferences	This is offered as being complimentary to other tools. The authors suggest that there is not much learned from the Smileyometer, but that laddering can add more in-depth knowledge as well as elicit reasons for choices
Hijkoop *et al.* (2020)	Tangible SR	Fun	Without suggesting there is anything amiss with any existing tools this was more of a design challenge to make user reporting more fun.
Sylla *et al.* (2017)	Paper Ladder	Preferences	Refers to extreme biases and the difficulty children might have in differentiating using Smileyometers. Paper ladder removes the need to verbalise that is embedded in some self-report measures.
Zhang *et al.* (2019)	Emo Form	Emotions	Using retrospection in a single form, this is offered as an alternative to, for example, repeated Smileyometers. It is actually measuring a different thing than a Smileyometer and the authors make that clear.
Huisman *et al.* (2013)	LEMtool	Emotions	Like the Emo Form, this is measuring something other than fun. The Smileyometer is mentioned in the literature review but not discussed concerning the LEMtool

children could not use it, it seems pertinent to revisit that claim. This raises a research question, which is *Can young children use the Smileyometer? If they can, are their scores too polarized to be of any use?* We explore this question in Case Study 1.

As well as those who have suggested new solutions (Table 5), several of the users of Smileyometers have included insightful general observations on how the scale works, especially in regards to differentiation, in their write ups. Leite *et al.* (2017) noted that the before scores (after seeing the robot but not interacting with it) were so high that it would be hard for the children to record a positive change after the interaction; they additionally noted that there was more room for negative change! However, they point out that their aim, as is always the case with experiences for children in HCI, was to provide an enjoyable experience and so high scores would be expected. This point, on what a child should do if they score one thing high and there is no option to then go higher, was noted by Hall *et al.* (2016) the solution they proposed, which was to take away the negative faces, really doesn't change the problem. Additionally, as seen in Figure 4, Tsoi *et al.* (2021) showed that *"awful"* is, and should be, a valid choice in a study on robot interactions.

Alluding to this concern, as to when a score is the offered maximum and a better experience comes along, there needs to be work done. In our own group we have hypothesized some solutions but are still not too sure about what children currently do, nor as to the extent of the problem. There has not so far been any study to explore what the effect is on Smileyometer ratings, of meeting a really fun, or a less fun activity first in a sequence of interactions. Similarly, there has not been much studied on how

FIGURE 4. Example of Awful responses (Tsoi *et al.*, 2021).

a child who may rate their anticipated fun as 5 but then have a mega time, give a rating for experienced fun. *We hypothesize that scores for other items after meeting first the "most fun" product would tend to be lower than if those same items were encountered first or after the "least fun" product.* We explore this hypothesis in Case Study 2.

1.5 Summary

From this examination of a subset of literature on the use of the Smileyometer it is clear that it is used in many ways but that mainly it is used to gather a score after interaction and predominantly with children aged 6–10. The use of it to compare things in a systematic way is limited but when it is used in this way researchers seem happy to use it alone or with other tools. There are many things that can be explored, one of these is whether small children (under 6) can use the tool and if so, does it bring anything useful to an evaluation. A second area of interest is to explore the expandability and differentiability of

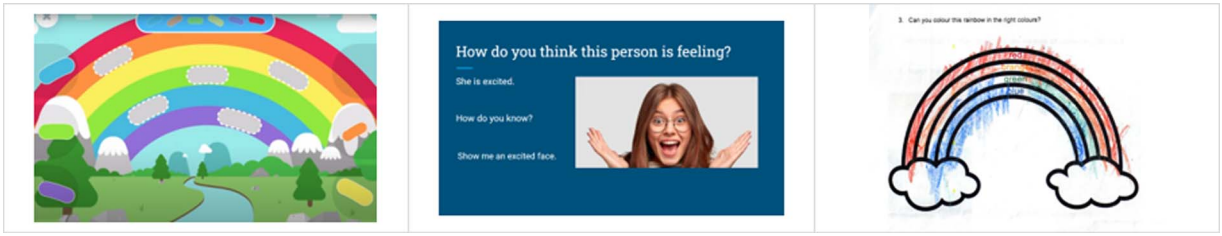

FIGURE 5. Example conditions (iPad, PowerPoint, worksheet) for Case Study 1.

the Smileyometer. Very few papers examine the effect on scores on the order in which children make their judgments; this is pertinent to before and after studies when the child, as per Hall *et al.* (2016) may start with *brilliant* and then where do they go if it gets better? This is an important limitation of any scale and worth further investigation. While most of the papers here were using the Smileyometer to either record responses from a set of questions or to get a measure of experienced fun, it is important to consider how data can be more useful to the UX community in regard to discrimination of products, services, or systems and the effect of the order in which competing things are met.

In the subsequent two sections of this paper we present new research, in the form of two case studies, that explores the use of the Smileyometer with young children and then, with older children, discrimination and order effects. These studies help show some of the methodological challenges of using Smileyometers as well as highlighting some possibilities for future exploration. We then conclude this paper with a summary of lessons learned for those seeking to use the Smileyometer in their work.

2 CASE STUDY 1—Small Children

From a small selection of published papers there is evidence that the Smileyometer has been used with pre-reading children; albeit it with some challenges and the need for some modifications (Chen *et al.*, 2019a, Leite & Lehman, 2016, Sargeant & Mueller, 2018). This case study explores Smileyometer scores given by small children aged 3 and 4, in two pre-school facilities, as part of a study to look at the UX of children with different learning activities. This study explores whether young children can use the Smileyometer and, if they can, what can be gleaned from their scores. Children attended twice a week for six weeks and each week they were introduced to a different topic (recognizing feelings, counting, animals, colours, hand-washing, body parts) with either a game on an iPad, a PowerPoint presentation delivered by an adult, or a set of worksheets facilitated by and adult. The six topics were all fairly similar and each had a game on the iPad, a PowerPoint introduction, and a worksheet that was similar across the topics—the content of these different presentations was made by the company that had made the iPad games as we sought to make a comparison of learning and engagement across these different learning media. More detail on this study can be found at Read *et al.* (2023). Here we focus only on the Smileyometer data from the children, looking at this for the first four weeks of the six weeks of the study.

2.1 Participants

The children were aged 3 and 4 in pre-school education in the UK. As described above they participated in a multi-week study comparing iPad games with two other learning activities (worksheets and a teacher led session using PowerPoint). As this was not full-time education, quite a few children did not attend all the sessions. No names were taken from the children who attended, in their own pre-schools, as part of their pre-school activities. The work was done with approval from the University Ethics committee. Parents had signed consent for the children to attend the sessions, which were attended by their nursery teachers and university staff. During the sessions children were free to not participate.

2.2 Procedure

When we first went to each pre-school, the nursery staff put the children into two small groups and noted their names for future weeks. The children then went to meet the researchers in one of two different areas of the nursery. One group used iPads and played a computer game, the others had the same topic delivered either by worksheets or by a talk with PowerPoint images (see Figure 5). In the second session of the same week, the children stayed in the same groups with the iPad or worksheets/PowerPoint and did follow on material on the same topic; thus in a single week they met a single topic twice. Staying in the same groups, in the following week, the children who had used the iPad in week one then did either the worksheets or PowerPoint and those who had not met the iPads met the second topic via a computer game. In week three, the iPads were again used by the same children as in week one—and so on. Note that as there were only two groups, the below data represents each group seeing the iPad for two weeks (four occasions) and seeing worksheets and PowerPoint for one week each (two sessions). Children were free to leave any group whenever they wanted to and a couple did walk away on a couple of occasions. Children were also told that while we would like to get their views about the things we were doing, they did not need to give these to us and they didn't need to fill anything in.

Before starting each activity, which was designed to last around 15 minutes, the children had their activity described; *"So we are going to play a game on these iPads to learn about X"* (the children could see the iPads), *"So we are going to learn about X with these worksheets"* (the children could see the worksheets), or *"So we are going to learn about X"* (with the first intro screen of the PowerPoint showing). They were then asked to fill a Smileyometer to indicate *"How much fun (how good) they thought this activity would be?"*. The first time we, the researchers, introduced this we showed the children the scales, read out the words and laid the Smileyometers on the floor in front of each child before giving them each a pencil and asking them to tick or mark one of the faces to indicate how much fun this activity would be. These children were not able to read, so this was read out to them, and the scales shown. After engaging with the activity, children completed a second Smileyometer—without being able to see their marks from the first—and this time were asked *"So, how fun do you think this activity was?"* in order to capture how much they had enjoyed it after the effect.

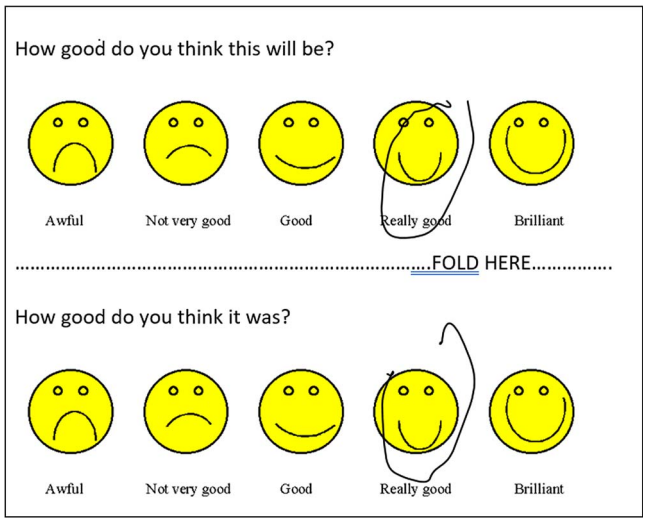

FIGURE 6. Smileyometer recording sheet for Case Study 1.

2.3 Apparatus

For the Smileyometer recording, children were given an A5 sheet of paper with two rows of Smileyometers (Figure 6) marked with a FOLD HERE dividing line for the researcher to use to *"hide"* the first answers (expected fun) from the second answers (experienced fun).

2.4 Results

The data collection with the Smileyometers was facilitated by several different researchers who did vary in how they followed the protocol and how they helped the children. Two researchers failed to fold the Smileyometers and so children working with those individuals were potentially able to see their earlier scores. An inspection of those answers, when compared to answers with another researcher doing the same activity with folded Smileyometers, didn't show any noticeable effect of this. Two of the researchers filled in most of the Smileys for the children having asked them where to point, a third filled in several, and the others did assist some children. Again, a visual inspection didn't show noticeable differences across these different completion patterns.

From a maximum potential "haul" of 168 pairs of Smileyometers (based on the numbers of children attending), 127 pairs were handed in (collected) and 117 had understandable scores on both scales. The non-understandable scores were where children had marked everything (1), had missed one or both of the two scales (6), or had identified more than one choice (3). Children were given agency to complete them or not and this accounts for the discrepancy between "haul" and handed in (see Figure 7 for examples).

When children filled in the Smileyometers themselves, they mainly used ticks or circles with roughly equal numbers choosing each method; some children ticked one of the pair and circled another.

It was hypothesized from earlier studies that (a) young children would struggle to complete Smileyometers and (b) they would score everything as *brilliant*. In terms of completion, it was perhaps surprising how many ended up being filled in but there is evidence, by looking at pen and style variation that a significant number were filled in by adults on behalf of the children and this would have helped completion—it also suggests that many of the children were "filling them in" by making a visual selection as opposed to directly marking the scales. When asked, researchers said they ticked or circled some (and in one case all) of the images directly as children pointed to their preferred score both to save time and limit distraction. It appears therefore that the scales can be used with young children, and can help to gather their opinions but that maybe adult help is needed.

Figure 8 shows the mean averaged before and after scores across the four weeks with scores hovering around four showing that the children were not rating everything a 5. The distribution of ratings for each numeric score, including before and after scores, over the four weeks are shown in Figure 9. This figure does demonstrate a skew but also shows what might be considered a "healthy" distribution of the lower scores in so far as all these scores have been chosen and all have been chosen in different weeks; this suggests some discrimination. In week one there is more discrimination than in subsequent weeks; this could be attributed to the effect of children becoming familiar with the research team and/or the activities and therefore feeling generally more positive in the following weeks, it could also be a sign that the children enjoyed the activities more in the second than in the first. It is not really possible to separate out the activity from the research team when considering a child's experience in this sort of study; further studies may be able to show if slight underenthusiasm is a regular outcome with children of this age in similar studies.

The numeric data from the children's ratings show two interesting findings. First, the children were discriminating and giving scores that averaged relatively low compared with other published work on small children using the Smileyometer. Looking at the range of ratings chosen in week 1 and week 4 it is clear that children did not always pick 5. Indeed, over the full set of pairs of before and after data, Table 6 shows the percentage of times that 5-5 (anticipated and experienced) was chosen and it can be seen that these frequencies are considerably lower than those reported in earlier studies. In a random exercise, the probability of a child picking 5-5 would be 1 in 25, so it appears that it was chosen between three and ten times more than it would be chosen by chance.

2.5 Discussion

It does appear that young children can use the Smileyometer to express an opinion and may also be able to mark that opinion themselves with a circle or a tick although some may need, or prefer, adult help. It also appears that young children do discriminate and do not score everything as *Brilliant*. In consideration of how they used the Smileyometer, the researchers who worked with these children were asked to reflect weekly on the use of the Smileyometer with such young children. The following comments were seen in the first two weeks:

(i) *"I helped them, one didn't want to do it."*
(ii) *"Seemed to work well, they seemed to understand the scale and put consideration into choosing an answer."*
(iii) *"With support they were able to select a smiley face."*
(iv) *"Easy to do, not convinced they understood."*
(v) *"Started doing the second sheet too soon, not really able to do it on their own."*
(vi) *"I needed to do it with them, they pointed to the face."*

Comment 4 points to a legitimate question as to whether such children associate the experience of the activity with the chosen smiley face; it is noted that comment 1 suggests contrariwise, that children did seem to know what they were doing; this clearly needs more study and we return to it in the discussion. Practical issues are highlighted in comments 1, 3, 5, and 6; however, in the comments later on in the study (see 7–11 below), there were signs

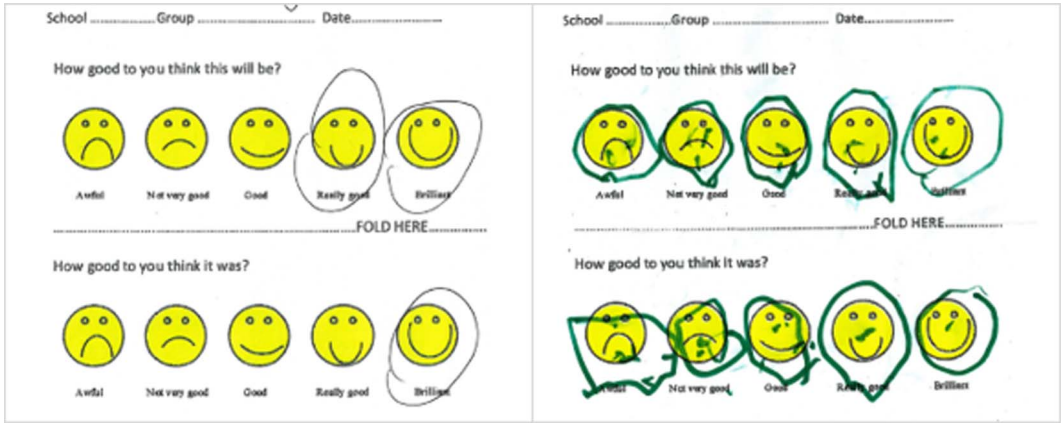

FIGURE 7. Example non-understandable Smileyometers from Case Study 1.

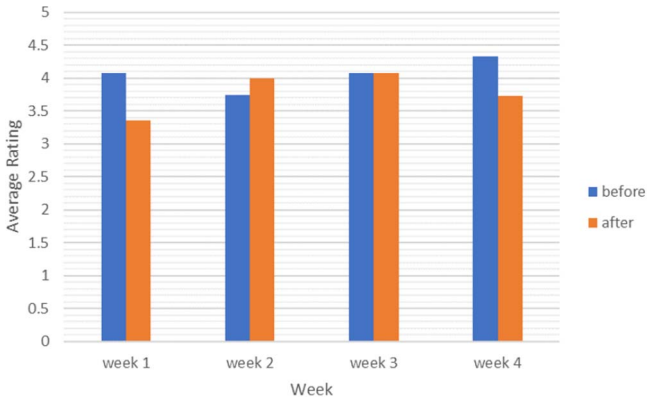

FIGURE 8. Children's ratings over the four weeks.

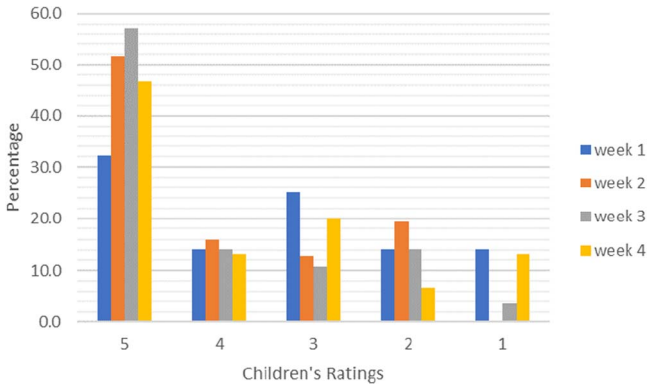

FIGURE 9. Percentage ratings of scores over the four weeks.

that the children were settling into the action, even if not entirely the meaning, of the task. The following comments were seen in the later weeks of the study:

(vii) *"Seem to be getting into the habit of filling it in now."*
(viii) *"They could point to a face, but, it wasn't clear whether they understood the purpose of the scale or were just picking a face because they understood the instruction 'point to a face'."*
(ix) *"One child chose 'not very good' but I observed them to be really enjoying the activity"*
(x) *"Children seem to be understanding the mapping better between experience and smiley face."*
(xi) *"I helped them with the first then they did their own second ticks."*
(xii) *"Completed it with ease."*

TABLE 6. Frequency of 5-5 scores over the weeks.

	Week 1	Week 2	Week 3	Week 4
Number of paired scores logged	28	31	28	30
Percentage being 5-5	14%	29%	36%	40%

These comments suggest that children were settling in to the activity with one suggesting that they could discern a link between experience and ratings (comment 8) but with another commentator still expressing some doubts on that score (comment 9)

In conclusion, this case study showed that young children can complete a Smileyometer, but probably do need some support at the start at least. It also shows that there is some differentiation although this is still small. For young children, without reading skills, the use of spoken language in surveys is very important and caution needs to be taken with anything that could be ambiguously interpreted especially with this age (Borgers & Hox, 2001) so the way the Smileyometers were presented to the children, in this study, is a potential confound in the results. It could be argued that fun could have been gathered from these children just from observations; however, that would take away from the children their agency to have their own opinions heard (Borgers *et al.*, 2000).

3 CASE STUDY 2—Beyond Brilliant

In this study the main aim was to look at Smileyometer data to explore order effects and discrimination to answer, *"How discriminating are children when evaluating several things?"* and *"How do previous encounters affect later encounters?"*. The aim was also to explore the hypothesis that, *"Scores for other items after meeting first the "most fun" product would tend to be lower than if those same items were encountered first or after the "least fun" product."* Specifically we also sought to explore the earlier published claim that 60% of children chose *Brilliant Brilliant* and that very little change happened between anticipated and experienced scores (Read *et al.*, 2002).

3.1 Participants

One hundred and thirty-five children aged 8 to 11 participated in a series of workshops that we organised in their schools. Children participated as class sets (nine in total) and selection of children for each event was organised by the school and consent collected by the school. Ethics had been gathered from the University and

FIGURE 10. The four games being compared in Case Study 2.

even though consent had been gathered from parents, assent was actively gathered from children too. Children completed the activity described here in week one of a multi-seek series of workshops about ocean health. The wider project is described in Read *et al.* (2024). Each session was attended by two researchers and the class teacher—the teacher was there to facilitate and did not actively engage in the session. In this study children played physical rather than digital games. The Smileyometer has been used in previous studies with non-digital experiences (Benton *et al.*, 2012, Sim & Cassidy, 2013) and has been used to compare non-digital and digital games (Oberhuber *et al.*, 2017b); the reason for using physical games in this session was that there was limited time, allowing 30 children to access digital games in a school was problematic (as most schools had lock down on most web games) and we wanted collaborative game play to take place as a precursor to later group work.

3.2 Procedure

On the day of the study, children were put into groups of two, three, or four by the teacher and were handed individual Smileyometer logging sheets as shown in Figure 11. This sheet was explained to them, and they were instructed to play a selection of games and log scores, on the three quadrants shown, for the first three games they played. The children were then given a game for their group and were asked to individually rate it before playing and then, after playing for 5 to 10 minutes, were asked to rate the experienced fun. The game they had was then collected in and another game was given to them. They again rated this before play and then after. This was repeated for the third game. The last game they met they simply played without scoring. The decision to score only three games was based on the children perhaps not having enough time to spend on four games as the time available for the activity was quite small. In this study, as the children were older than those in Case Study 1, adults did not actively observe the filling in of the Smileyometers, nor did they make any comments; however, the children all appeared to be able to fill them in with no need for assistance and with no difficulty. An important side activity related to this Smileyometer completion was that the data from the Smileyometers was being used to explain to children about different data types; thus, once the ratings were gathered children were asked what use this data might be to games developers and to the research team, they showed a good understanding of the potential use of such data and then they had data explained to them in an active effort to give them enough information as to whether they wanted to hand this data in or not. This protocol, around data, is explained further in Read *et al.* (2024).

3.3 Apparatus

Four games were provided for the children—multiple copies of the same games were available so there could have been two groups playing the same game at the same time. The games (shown in Figure 10 L to R) evaluated were Dobble, (Ocean) Bingo, Shark, and Top Trumps. Dobble is a matching game where two cards are placed down and the first child to spot the item that is on both wins Bingo is self explanatory but had an Ocean theme in this case, Shark was a plastic toy with teeth that were pushed down by a player with one randomly triggering the shark to "bite" the player and Top Trumps is a well known game where cards are pitched against each other against a characteristic of the card's image. We did not ask children if they had met these specific games before but when watching them play we noted that in most of the small groups at least one child had played a version of bingo before (but not our specific one) and that similarly in most groups one or more children had played a version of Top Trumps before (but not our version). In some groups there was a child who had experienced a version of Dobble before (but again not our version) but we were not aware of any child having previously played the Shark game. Thus, across the children there was generally some shared knowledge of how the games worked. Where there appeared to be no "expertise" coming into the game play we did observe the children rapidly figuring it out.

Scores were recorded individually by children on a single sheet of paper that had three before and after Smileyometers on it (see Figure 11).

3.4 Results

For analysis, each Smileyometer entry was coded from 1 to 5 and registered against a code to identify whether this was from the first, second or third game play experience. Due to the numbers varying in each group and there being so many pupils taking part in this activity, there was a small variation in the numbers meeting each of the games, but each game was met in the different orders in a reasonably balanced way as shown in Table 7.

Observations of the children playing the games suggested that one game (the Shark game) seemed more fun that the others. To establish if this was borne out by the results of the Smileyometers, an analysis of all the scores, by game, was completed (Table 8).

From these data it appears that the most fun game was Shark and the least fun was Dobble. This confirmed the authors' expectations in terms of the most fun game; we had not hypothesised on the least fun game.

To explore order effects, scores for these two games were compared by the order (out of the three games) when they were met , see Figure 12.

From these graphs it "appears" that Dobble (LHS) scored generally better on experienced fun, when it was the first thing that the pupils met, than when it was encountered later on. Counter wise, Shark scored much better on anticipated fun when it was not the first game encountered. A possible explanation for this is that, as the pupils were rotating the games around in the classroom,

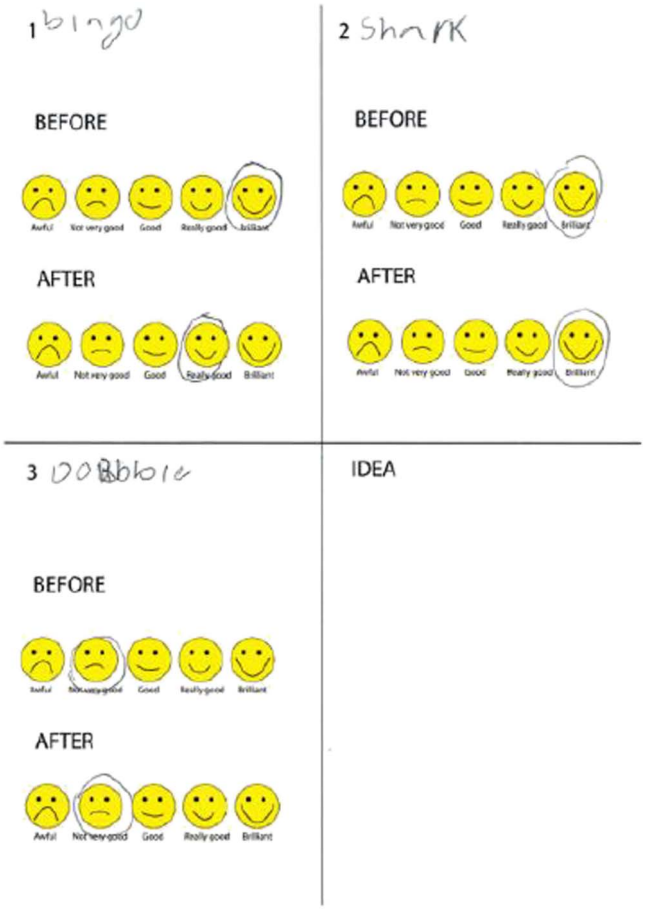

FIGURE 11. Smileyometer recording sheet for Case Study 2.

by the time a group came to rate the Shark game in second or third position, they might have already noticed others in the room having a lot of fun with it.

For Dobble the experienced fun was markedly rated as less by those meeting it third than those meeting it first. Top Trumps experienced a similar lack of support falling from an average score of 3.95 when rated first to an average of 3.33 when rated third. Table 9 shows how the ratings differed when being encountered first and third.

To explore the effect of meeting either Dobble or Shark first, rather than one of the more neutral choices, we chose to explore data from those who met Dobble first and Shark second with those that met Shark first and Dobble second. Across the study 16 pupils played Shark first and immediately after played Dobble, and another 16 played Dobble first and immediately after played Shark. Table 10 shows the averages.

These averages seem to suggest that expectations of the Shark being fun were potentially raised for those who had previously experienced the Dobble game but as noted above it could just be that they saw the Shark game in their periphery. There is less variation on the experienced fun. In wondering whether meeting the (apparently) best thing first might have an impact on later things, there is no evidence here to suggest that is what happens. The scores for experienced fun for Dobble, having met Shark, at 3.5 are not different from those for Dobble when chosen first.

We were also curious about the 5 scores. Firstly, we explored the entire data set to see how many pupils gave everything a 5. That would be *Brilliant Brilliant* across three games. There were only 4/135 that did this, which perhaps supports the notion that pupils of this age can, and do, differentiate. Of course it could be pure chance that these numbers occurred but the probability of this happening by chance is very low (1 in 15625) as opposed to the actual occurrence being 1 in 33.75. Following Hall *et al.* (2016)'s critique about the ceiling on scales, we were also interested to see if a child who gave Dobble (the perceived least fun game) a 5 as an experience was then more likely to give 5s to other games (thus potentially showing the impact of the ceiling on the scale).

Given that some children naturally give 5s, and 5s are quite common, the scores for Dobble starting as 5 are compared with those for TopTrumps (a more neutral game) starting as 5. As shown in Table 11, the average score given to other games from those who rated Dobble as 5, and those who rated Dobble lower, does not vary. However, given that with TopTrumps, those who gave a 5 for that subsequently went on to give higher marks than those who started their ratings lower,suggests that some children are generally more likely to give higher marks and so naturally, all things being equal, we might have expected those who started

TABLE 7. Numbers meeting each game and when played.

	When Played First	Second	Third	Overall
Shark	24	35	36	95
Dobble	36	40	29	105
Top Trumps	39	31	32	102
Bingo	33	29	33	95

TABLE 8. Mean and median averages from all children for all the games.

	Mean (anticipated)	Mean (experienced)	Median (anticipated)	Median (experienced)
Shark	4.03	4.3	4	4
Dobble	3.15	3.55	3	3
Top Trumps	3.29	3.54	3	4
Bingo	3.26	3.65	3	4

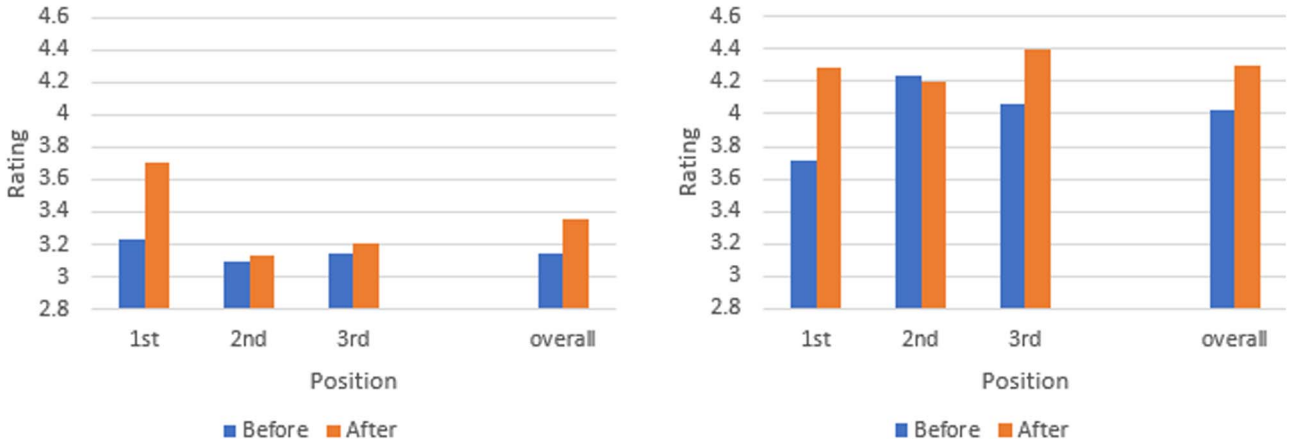

FIGURE 12. Before and after ratings according to the order in which the game was met for game (a - LHS) Dobble and (b -RHS) Shark.

TABLE 9. Comparing ratings according to when seen.

	First (anticipated)	(experienced)	Third (anticipated)	(experienced)	Change (anticipated)	(experienced)
Shark	3.71	4.29	4.06	4.4	9.4%	2.6%
Dobble	3.23	3.71	3.14	3.21	-2.8%	-13.5%
Top Trumps	3.44	3.95	3.56	3.33	3.5%	-15.6%
Bingo	3.42	3.58	3.09	3.73	-9.6%	4.2%

TABLE 10. Effect of preceding choice on immediate next choice.

	Shark (anticipated)	(experienced)	Dobble (anticipated)	(experienced)
Shark First	3.5	4	3.5	3.5
Dobble First	4.5	4.4375	3.25	3.75

TABLE 11. Does a high score for the weakest game predict more high scores after?

	Average score given to other games	Number
With 5 first on Dobble	3.77	13
With other scores on Dobble	3.76	21
With 5 first on Top Trumps	3.95	20
With other scores on Top Trumps	3.25	16

with a 5 on Dobble to have given higher marks in subsequent evaluations but this was not available to them in the scale. This certainly merits further studies.

3.5 Discussion

This case study show that children were discriminating and very few (<3%) chose *Brilliant* all the way. In terms of how previous encounters affect later encounters, the data in Tables 9 and 10 seem to suggest that there may be some effect but more work is needed to establish what this is. The hypothesis that, *"Scores for other items after meeting first the "most fun" product would tend to be lower than if those same items were encountered first or after the "least fun" product."* is not really supported (see Table 10) where the scores for Dobble, last two columns, do not change.

4 DISCUSSION

4.1 What has been learned?

From examining community use of the Smileyometer, albeit with a relatively small subset of papers, it appears that the Smileyometer is used in a variety of ways, some that are different from those initially envisaged where it was intended to be a tool to use before and after engagement alongside other metrics. The Smileyometer has inspired other visual scales, has been used for exploring new methods, and has been used as a response-catcher for questions and UX surveys (where it often sees small modifications). This differentiated appropriation of the tool is valued and it has enabled new questions to be asked and explored.

The community has used the Smileyometer to compare experiences across different systems and products and to compare before and after experiences. These two uses can be considered horizontal—for example with a control group and two experimental groups—and vertical—when used before and after. The end point of horizontal use is to show an effect of an interaction or a design or to show children's preferences for one product over the other. Critiques of the Smileyometer in the literature argue that it is not effective with some children and also that the skew and lack of differentiation are problematic.

The two case studies presented here provide some evidence that (a) young children can complete Smileyometers with help, (b) *Brilliant* all the way is not so common as expected, and (c) that children can and do attribute a range of scores, viz.

differentiate. These findings come with the usual caveats that the studies described here are small, that the results would ideally be validated with other scales and that there are many open questions around what is being chosen by children, especially young children, when they choose a smiley face.

4.2 Use and Validity

There is an ongoing debate on how to use data from Smileyometers and how to ascertain validity. For all children, and especially young children, discussion on whether their responses are representative of their experience, as noted in the comments in our own studies, is pertinent. However, it is important to also understand that participation is empowering for children (Van Mechelen *et al.*, 2021) and increases feelings of self-esteem and wellbeing (Gordon & Russo, 2009); this is a good reason to facilitate ways for young children to give feedback. Several papers looked at in this review validated the Smileyometer with favourable comparisons to results from the Again Again (Lochtefeld *et al.*, 2022, Sim & Cassidy, 2013, Zhang *et al.*, 2021) and the Fun Sorter (Dawidowsky *et al.*, 2021, Jurdi *et al.*, 2018) and others compared results to observed fun (Leite *et al.*, 2017, MacFarlane *et al.*, 2005). A potentially interesting area to explore going forward is to compare Smileyometer results with other user ratings; one paper compared the Smileyometer (M = 3) with Google (4) and Apple playstore (3.7) ratings (Tsoi *et al.*, 2021), which may have been adult generated so could be quite different.

There is still concern on how a child can report that something is better than expected, or better than a previous experience, when choosing *Brilliant* as a starting point. This was discussed by Hall *et al.* (2016) and was studied in our current paper from the perspective of establishing if there is any evidence to support the hypothesis that this is a problem. In our labs we have seen children occasionally putting two circles around *Brilliant* or placing two ticks, but in the case studies reported here, aside from two of the very young children's first attempts at the Smileyometer, this was not seen. With the large numbers of children and the scoring in Case Study 2, it appears the problem may be overstated; what is apparent is that discrimination is possible, as shown in papers in Table 4; and in Case Study 2 the Smileyometer did seem to identify a winning game.

Skew and limited differentiation are evident in the results in many of the studied academic papers. When looking at the situations being evaluated, it is clear that a skew towards happiness is potentially more a consequence of our unwillingness to subject children to bad experiences, as put eloquently in Leite *et al.* (2017), rather than an inherent problem with the scale—which, when used in tricky interactions, can gather "*awful*" ticks (Tsoi *et al.*, 2021). We need to constantly remind ourselves that skew is not necessarily bad—it is not wrong that children are having such a good time with our technology.

4.3 New Guidelines for using the Smileyometer

BEFORE: G1. Find your Reason: The first guideline is to consider the use of the Smileyometer mindfully, that is to say—ask why children's "self-reported" experience matters, and, if it does, ask whether this is the right tool for this situation. In many situations observations of children may be sufficient to ascertain that children are having fun or that they are having a good experience. In some situations other tools may be more appropriate for children to use including those described in the literature review in this paper. One reason to gather numeric data is to add rigour to a study, another is for ease of reporting—that is to get a snapshot of experience (Antle & Hourcade, 2022). In such cases where numerics are considered a good choice, rigour to the researcher, and value to the child, are both supplied by paying attention to making the Smileyometer (or other chosen scale) easy to use and valuable for its purpose. Questions should be asked as to whether a before and after score is needed and whether the objects being compared are likely to result in different scores. As an example, using a Smileyometer to "compare" children's experiences of two games, both expected to be equally fun, might be challenged—are we wasting the children's time or are we empowering them to give an opinion? The scales should really only be used to support a hypothesis or answer a question.

DURING: G2. Consider Completion: Having chosen the Smileyometer, the aim is to then get the best results possible from the tool. This may require additional words, different presentations, or practice sessions. As described in the four-week study with young children (Case Study 1), it may take a little bit of time to help small children get the hang of what is going on. If the Smileyometer hasn't been used before in this, or a similar, context or with similar children, pilot its use, or consider training. We refer the reader to the good practice found in Leite *et al.* (2017) where children initially rated their responses to "broccoli", "ice-cream", and "a stubbed toe" on the Smileyometer scales before using the scales to evaluate a robot experience. From Case Study 1 in this paper it may be that young children need practical assistance to complete the scales especially when they first meet them; this act of showing and then letting them do it is how children are taught in school so incorporating practice sessions on scale completion make sense. When using repeating Smileyometers, maybe for a before and after scale, consider presenting these on separate *"pages"* to reduce cross referencing and, when used with surveys to capture responses (which is a common use reported in the literature), which are often presented on a single page, mix the order of questions if possible to minimise presentation and order bias.

AFTER: G3. Report with Caution: Rigour in reporting, and using Smileyometer data to show an effect, requires cautious use of statistical tests and arithmetic means. With a single population comparing different things, arithmetic means can be used for illustration. When a large group from a reasonably homogenous population, assigned without bias or selection, is comparing things in a between-subjects situation, arithmetic means can also be applied for illustration. A mean can be reported as an average score in a summative one-time use of the scale, but it may be better in those cases to report percentages for each of the five scores. Getting the rigour right is essential to justify the child's effort.

5 CONCLUSION

A literature study showed the diverse ways in which the Smileyometer has been used in HCI. The survey highlighted that many variations are in place and that there is considerable variety in how the Smileyometer is used; from research studies with several conditions, as a response tool with other surveys, and as a summative tool to log the general happiness of children with an experience. Discrimination and skew, as well as use with younger children, were themes that emerged from the literature, and we explored these in two case studies. The studies are not conclusive; however, the first points to young children being able to complete the Smileyometer and the second illustrates concerns about skew and discrimination being partly unfounded. Considering how to improve the effectiveness of the Smileyometer in future studies, we provided three guidelines to encourage users to think carefully

about why they are asking children's opinions, to actively think about how the children are empowered to complete the Smiley-ometers, and to consider how to report such findings.

There are several limitations to this work, the first being the choice made to only look for insights from two venues, and within that, only from those papers that provided full text versions. This has naturally omitted several significant papers that will also contribute insights. Initially there was a plan to look at papers from IJCCI but on searching for those, given the relatively small size of the community, it was clear that a large percentage of these were reporting work that had previously been reported in the included conference papers. As the present work is not being presented as a systematic review, a pragmatic choice that was made to not extend the search into other journal venues. A second limitation is with the two case studies reported, each is a one-time study with all the limitations inherent in drawing conclusions from such works. Case Study 1, with small children had only a small population, and the findings from it may not transfer easily to other situations or contexts especially where formal schooling starts at different ages. Case Study 2, while having large numbers, compared games that were not digital, and it might be hypothesised that with digital games the ratings might have skewed differently; findings from this study cannot necessarily be translated to interactive experiences, interaction modalities, or educational contexts.

There is considerably more work that can be done in the future. This includes studying different populations and contexts and more work around the discrimination question with studies that are comparing things that are predicted to be quite different in order to see if children's responses do in fact match expectations.

Data Availability Statement

The data underlying this article cannot be shared publicly due to us not explicitly asking for that when gathering consent. Summary data will be shared on reasonable request to the corresponding author.

Acknowledgments

We acknowledge the fabulous children who helped make this paper happen and their teachers who facilitated the interaction.

References

Ahmad, M. I., Mubin, O., & Orlando, J. (2016). *Effect of different adaptations by a robot on children's long-term engagement: An exploratory study*.

Al-Dawsari, H. M., & Hendley, R. (2024). Exploring the impact of matching E-training material to Arabic dyslexia type on learners, Äô Reading performance, satisfaction and behaviour. *Interacting With Computers*, iwad050. https://doi.org/10.1093/iwc/iwad050.

Alghabban, W. G., & Hendley, R. (2020). *The impact of adaptation based on students' dyslexia type: An empirical evaluation of students' satisfaction*.

Alghabban, W. G., & Hendley, R. (2023). Adaptive E-learning and dyslexia: an empirical evaluation and recommendations for future work. *Interacting With Computers*, iwad036.

Antle, A. N., & Hourcade, J. P. (2022). Research in child–computer interaction: provocations and envisioning future directions. *International Journal of Child–Computer Interaction*, **32**, 100374.

Argyriadi, A., & Sotiropoulou-Zormpala, M. (2017). Engaging first-graders in language arts through 'arts-flow activities'. *Curriculum Perspectives*, **37**, 25–38.

Bartneck, C., Kulić, D., Croft, E., & Zoghbi, S. (2009). Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International Journal of Social Robotics*, **1**, 71–81.

Benton, L., Johnson, H., Ashwin, E., Brosnan, M., & Grawemeyer, B. (2012). *Developing ideas: supporting children with autism within a participatory design team*.

Bertou, E., & Shahid, S. (2014). *Lowfidelity prototyping tablet applications for children*.

Bhattacherjee, A., & Premkumar, G. (2004). Understanding changes in belief and attitude toward information technology usage: a theoretical model and longitudinal test. *Management Information Systems Quarterly*, 229–254.

Bonner, M., Wang, L., & Mynatt, E. D. (2012). *Activity-based interaction: designing with child life specialists in a children's hospital*.

Borgers, N., De Leeuw, E., & Hox, J. (2000). Children as respondents in survey research: cognitive development and response quality 1. *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique*, **66**, 60–75.

Borgers, N., & Hox, J. (2001). Item nonresponse in questionnaire research with children. *Journal of Official Statistics-Stockholm*, **17**, 321–335.

Borgers, N., Hox, J., & Sikkel, D. (2003). Response quality in survey research with children and adolescents: the effect of labeled response options and vague quantifiers. *International Journal of Public Opinion Research*, **15**, 83–94.

Carroll, J. M. (2004). Beyond fun. *Interactions*, **11**, 38–40.

Cesário, V., Radeta, M., Matos, S., & Nisi, V. (2017). *The ocean game: Assessing children's engagement and learning in a museum setting using a treasure-hunt game*.

Chen, Y., Lin, Y., Liu, L., Yao, C., & Ying, F. (2019a). *Shadower: applying shadows to children's outdoor interaction*.

Chen, Y., Lin, Y., Wang, J., Liu, L., Yao, C., & Ying, F. (2019b). *Ipanda: A playful hybrid product for facilitating children's wildlife conservation education*.

Chu, S. L., Quek, F., & Sridharamurthy, K. (2015). *Augmenting children's creative self-efficacy and performance through enactment-based animated storytelling*.

Cibrian, F. L., Gutierrez, O., & Escobedo, L. (2021). *Pensando: Developing a smartpen assessing handwriting skills for children*.

Cosentino, G., Gelsomini, M., Sharma, K., & Giannakos, M. (2023). *Interaction modalities and children's learning in multisensory environments: Challenges and trade-offs*.

Dawidowsky, K., Holz, H., Schwerter, J., Pieronczyk, I., & Meurers, D. (2021). *Development and evaluation of a tablet-based reading fluency test for primary school children*.

Delden, R. V., Bos, D. P.-O., With, A. J. V. D., Vogel, K., Klaassen, R., Zwart, N., Faber, J., Thio, B., & Kamp, M. V. D. (2020). *Spiroplay, A Suite of Breathing Games for Spirometry by Kids & Experts*. (Vivianne).

Deshmukh, A., Janarthanam, S., Hastie, H., Lim, M. Y., Aylett, R., & Castellano, G. (2016). *How expressiveness of a robotic tutor is perceived by children in a learning environment*.

Dietz, G., Pease, Z., McNally, B., & Foss, E. (2020). *Giggle gauge: a self-report instrument for evaluating children's engagement with technology*.

Dijk, E. M. V., Lingnau, A., & Kockelkorn, H. (2012). *Measuring enjoyment of an interactive museum experience*.

Draper, S. W. (1999). Analysing fun as a candidate software requirement. *Personnel Technology*, **3**, 117–122.

Duh, H. B.-L., Yee, S. L. C. Y., Gu, Y. X., & Chen, V. H.-H. (2010). *A narrative-driven design approach for casual games with children*.

Erfurt, G., Hornecker, E., Ehlers, J., & Plaschkies, S. (2019). *Hands-on math: A training system for children with dyscalculia*.

Ferraz, M., Câmara, A., & O'Neill, A. (2016). *Increasing children's physical activity levels through biosymtic robotic devices*.

Ferraz, M., Resta, P. E., & O'neill, A. (2017). *Whole-body interaction in natural environments benefits children's cognitive function compared to sedentary interaction indoors*.

Ferraz, M., Romão, T., & Câmara, A. (2010). *The "alberto's gravimente toys": children's fiction on technological design*.

Foster, M. E., Lim, M. Y., Deshmukh, A., Janarthanam, S., Hastie, H., & Aylett, R. (2014). *Affective feedback for a virtual robot in a real-world treasure hunt*.

Fowler, A. (2013). *Measuring learning and fun in video games for young children: a proposed method*.

Fowler, A. (2017). *Engaging young learners in making games: an exploratory study*.

Fowler, A. (2019). *Jamming with children: an experience report*.

Fowler, A., & Schreiber, I. (2017). *Engaging under-represented minorities in stem through game jams*.

Garcia-Sanjuan, F., Nacher, V., & Jaen, J. (2016). *Markairs: Are children ready for marker-based mid-air manipulations?*

Godinez, K. D., Gaytán-Lugo, L. S., Alcaraz-Valencia, P. A., & Arellano, R. M. (2017). *Evaluation of a low fidelity prototype of a serious game to encourage reading in elementary school children*.

Gordon, M., & Russo, K. (2009). Children's views matter too! A pilot project assessing children's and adolescents' experiences of clinical psychology services. *Child Care Pract.*, **15**, 39–48.

Graßl, I., & Fraser, G. (2023) The abc of pair programming: Gender-dependent attitude, behavior and code of young learners. In *Proceedings of the 45th International Conference on Software Engineering: Software Engineering Education and Training, ICSE-SEET '23*, pp. 115–127. IEEE Press.

Greifenstein, L., Graßl, I., Heuer, U., & Fraser, G. (2022). *Common problems and effects of feedback on fun when programming ozobots in primary school*.

Hall, L., Hume, C., & Tazzyman, S. (2016). *Five degrees of happiness: Effective smiley face likert scales for evaluating with children*.

Hastie, H., Lim, M. Y., Janarthanam, S., Deshmukh, A., Aylett, R., Foster, M. E., & Hall, L. (2016). *I remember you! interaction with memory for an empathic virtual robotic tutor*.

Hernandez-Lara, M., Martinez-Garcia, A. I., & Caro, K. (2023). Using Emotion4Down: evaluating the Design of a Serious Video Game for supporting emotional awareness with people with intellectual disabilities. *Interacting With Computers*, **35**, 363–386.

Hijkoop, V., Bekker, T., Skovbjerg, H. M., Lieberoth, A., & Jørgensen, H. H. (2020). *Designing playful, tangible self-report tools to give the child a voice*.

Holz, H., Beuttler, B., & Ninaus, M. (2018). *Design rationales of a mobile game-based intervention for german dyslexic children*.

Huisman, G., Hout, M. V., Dijk, E. V., Geest, T. V. D., & Heylen, D. (2013). *Lemtool: measuring emotions in visual interfaces*.

Hyde, J., Kiesler, S., Hodgins, J. K., & Carter, E. J. (2014). *Conversing with children: cartoon and video people elicit similar conversational behaviors*.

IJsselsteijn, W. A., De Kort, Y. A., & Poels, K. (2013). *The game experience questionnaire*.

Joly, A. V. (2007). *Evaluating interactive tv applications for and with preliterate children*.

Jung, D. H., Kim, J., Lee, J. G., Yang, H. J., & Ryu, H. (2019). *Lessons learned from an auditory-vibrotactile sensory experience in the museum*.

Jurdi, S., Garcia-Sanjuan, F., Nacher, V., & Jaen, J. (2018). Children, Äôs acceptance of a collaborative problem solving game based on physical versus digital learning spaces. *Interacting With Computers*, **30**, 187–206.

Keskinen, T., Heimonen, T., Turunen, M., Rajaniemi, J.-P., & Kauppinen, S. (2012). Symbolchat: a flexible picture-based communication platform for users with intellectual disabilities. *Interacting With Computers*, **24**, 374–386.

Klingberg, B., Hoeboer, J., Schranz, N., Barnett, L., De Vries, S. I., & Ferrar, K. (2019). Validity and feasibility of an obstacle course to assess fundamental movement skills in a pre-school setting. *Journal of Sports Sciences*, **37**, 1534–1542.

Kuhn, A., Quintana, C., & Soloway, E. (2009). *Storytime: a new way for children to write*.

Lagerstam, E., Olsson, T., & Harviainen, T. (2012). *Children and intuitiveness of interaction: a study on gesture-based interaction with augmented reality*.

Lehnert, F. K., Lallemand, C., Fischbach, A., & Koenig, V. (2020). *Experimenter effects in children using the smileyometer scale*.

Leite, I., & Lehman, J. F. (2016). *The robot who knew too much: Toward understanding the privacy/personalization trade-off in child-robot conversation*.

Leite, I., Pereira, A., & Lehman, J. F. (2017). *Persistent memory in repeated child-robot conversations*.

Lochtefeld, M., Milthers, A. D. B., & Merritt, T. (2022). *Staging constructionist learning about energy for children with electrochromic displays and low-cost materials*.

MacFarlane, S., Sim, G., & Horton, M. (2005). *Assessing usability and fun in educational software*.

Maqsood, S., & Chiasson, S. (2021). Design, development, and evaluation of a cybersecurity, privacy, and digital literacy game for tweens. *ACM Transactions on Transactions on Privacy and Security*, **24**, Article 28.

McAuley, E., Duncan, T., & Tammen, V. V. (1989). Psychometric properties of the intrinsic motivation inventory in a competitive sport setting: a confirmatory factor analysis. *Research Quarterly for Exercise and Sport*, **60**, 48–58.

Melniczuk, A., & Vrapi, E. (2023) Exploring feedback modality designs to improve young children's collaborative actions. In *Proceedings of the 25th International Conference on Multimodal Interaction, ICMI '23*, pp. 271–281. USA. Association for Computing Machinery, New York, NY.

Mostowfi, S., Mamaghani, N. K., & Khorramar, M. (2016). Designing playful learning by using educational board game for children in the age range of 7–12:(a case study: recycling and waste separation education board game). *International Journal of Environmental and Science Education*, **11**, 5453–5476.

Oberhuber, S., Kothe, T., Schneegass, S., & Alt, F. (2017a). *Augmented games: Exploring design opportunities in ar settings with children*.

Oberhuber, S., Kothe, T., Schneegass, S., & Alt, F. (2017b) Augmented games: Exploring design opportunities in ar settings with children. In *Proceedings of the 2017 Conference on Interaction Design and Children, IDC '17*, pp. 371–377. Association for Computing Machinery, New York, NY, USA.

Oliver, R. L. (1980). A cognitive model of the antecedents and consequences of satisfaction decisions. *Journal of Marketing Research*, **17**, 460–469.

Ooi, Y. P., Goh, D. H.-L., Mekler, E. D., Tuch, A. N., Boon, J., Ang, R. P., Fung, D., & Gaab, J. (2016). *Understanding player perceptions of regnatales, a mobile game for teaching social problem solving skills*.

Çorlu, D., Taşel, E., Turan, S. G., Gatos, A., & Yantaç, A. E. (2017). Involving autistics in user experience studies: a critical review.

Park, H. W., Gelsomini, M., Lee, J. J., & Breazeal, C. (2017). *Telling stories to robots: The effect of backchanneling on a child's storytelling*.

Price, S., & Pontual Falcão, T. (2011). Where the attention is: discovery learning in novel tangible environments. *Interacting With Computers*, **23**, 499–512.

Read, J. (2008). Validating the fun toolkit: an instrument for measuring children's opinions of technology. *Cognition, Technology & Work*, **10**, 119–128.

Read, J. (2012). *Evaluating artefacts with children: age and technology effects in the reporting of expected and experienced fun*.

Read, J., Horton, M., Fitton, D., King, J., Sim, G., Allen, J., Doumanis, I., Graham, T., Xu, D., Tierney, M., Lochrie, M., & MacKenzie, S. (2024) Inclusive child engagement in hci: Exploring ocean health with schoolchildren. In *Proceedings of the 23rd Annual ACM Interaction Design and Children Conference, IDC '24*, pp. 83–92. USA. Association for Computing Machinery, New York, NY.

Read, J., Horton, M., Fitton, D., Sim, G., Dick, R. A., Mazzone, E., & Forbes, R. (2023) Small cci–exploring app evaluation with preschoolers. In *Proceedings of the 22nd Annual ACM Interaction Design and Children Conference*, pp. 94–99. York. ACM, New.

Read, J., & MacFarlane, S. (2006) Using the fun toolkit and other survey methods to gather opinions in child computer interaction. In *Proceedings of the 2006 Conference on Interaction Design and Children, IDC '06*, pp. 81–88. Association for Computing Machinery, New York, NY, USA.

Read, J., MacFarlane, S., & Casey, C. (2002) Endurability, engagement and expectations: Measuring children's fun. In *Interaction Design and Children*, vol. **2**, pp. 1–23. Shaker Publishing, Eindhoven.

Salian, K., & Sim, G. (2014). *Simplifying heuristic evaluation for older children*.

Sargeant, B., & Mueller, F. F. (2018). *How far is up? bringing the counterpointed triad technique to digital storybook apps*.

Schafer, G. J., Green, K. E., Walker, I. D., Lewis, E., Fullerton, S. K., Soleimani, A., Norris, M., Fumagali, K., Zhao, J., Allport, R., Zheng, X., Gift, R., & Padmakumar, A. (2013). *Designing the lit kit, an interactive, environmental, cyber-physical artifact enhancing children's picture-book reading*.

Sim, G., & Cassidy, B. (2013) Investigating the fidelity effect when evaluating game prototypes with children. In *In 27th International BCS Human Computer Interaction Conference (HCI 2013) 27*, pp. 1–6. BCS Learning & Development Ltd.

Sim, G., Cassidy, B., & Read, J. (2013). *Understanding the fidelity effect when evaluating games with children*.

Sim, G., Horton, M., & McKnight, L. (2016a). Ipad vs paper prototypes: does form factor affect children's ratings of a game concept?.

Sim, G., Nouwen, M., Vissers, J., Horton, M., Slegers, K., & Zaman, B. (2016b). Using the memoline to capture changes in user experience over time with children. *International Journal of Child–Computer Interaction*, **8**, 1–14.

Sim, G., & Read, J. (2024). Using eye-tracking to demonstrate children, Äôs attention to detail when evaluating low-Fidelity prototypes. *Interacting With Computers*, iwad052.

Simpson, K., Imms, C., & Keen, D. (2022). The experience of participation: eliciting the views of children on the autism spectrum. *Disability and Rehabilitation*, **44**, 1700–1708.

Soleimani, A., Green, K. E., Herro, D., & Walker, I. D. (2016). *A tangible, story-construction process employing spatial, computational-thinking*.

Sylla, C. M., Arif, A. S., Segura, E. M., & Brooks, E. I. (2017). *Paper ladder: a rating scale to collect children's opinion in user studies*.

Tsoi, N., Connolly, J., Adéníran, E., Hansen, A., Pineda, K. T., Adamson, T., Thompson, S., Ramnauth, R., Vázquez, M., & Scassellati, B. (2021). *Challenges deploying robots during a pandemic: An effort to fight social isolation among children*.

Tzortzoglou, F. (2023) Exploring the usability of mobile augmented reality interactions in relation with primary school students' level of cognitive abilities. In *Proceedings of the 2nd International Conference of the ACM Greek SIGCHI Chapter, CHIGREECE '23*. Association for Computing Machinery, New York, NY, USA.

Van Mechelen, M., Have Musaeus, L., Iversen, O. S., Dindler, C., & Hjorth, A. (2021) A systematic review of empowerment in child–computer interaction research. In *Proceedings of the 20th Annual ACM Interaction Design and Children Conference, IDC '21*, pp. 119–130. USA. Association for Computing Machinery, New York, NY.

Vonach, E., Ternek, M., Gerstweiler, G., & Kaufmann, H. (2016). *Design of a health monitoring toy for children*.

Wang, P., Xu, J., & Wu, Y. (2019). Preschool children's preferences for library activities: laddering interviews in chinese public libraries. *Library & Information Science Research*, **41**, 132–138.

Wang, Y.-S. (2003). Assessment of learner satisfaction with asynchronous electronic learning systems. *Information & Management*, **41**, 75–86.

Xie, L., Antle, A. N., & Motamedi, N. (2008). *Are tangibles more fun? comparing children's enjoyment and engagement using physical, graphical and tangible user interfaces*.

Xu, W., Ma, J., Yao, J., Lin, W., Zhang, C., Xia, X., Zhuang, N., Weng, S., Xie, X., Feng, S., Ying, F., Hansen, P., & Yao, C. (2023). *Mathkingdom: Teaching children mathematical language through speaking at home via a voice-guided game*.

Yarosh, S., & Kwikkers, M. R. (2011). *Supporting pretend and narrative play over videochat*.

Yusoff, Y. M., Ruthven, I., & Landoni, M. (2011). *The fun semantic differential scales*.

Zaman, B., & Abeele, V. V. (2010). *Laddering with young children in user experience evaluations: theoretical groundings and a practical case*.

Zaman, B., Abeele, V. V., & De Grooff, D. (2013). Measuring product liking in preschool children: an evaluation of the smileyometer and this or that methods. *International Journal of Child–Computer Interaction*, **1**, 61–70.

Zaman, B., & Vanden Abeele, V. (2007). How to measure the likeability of tangible interaction with preschoolers. *CHI Nederland*, 57–59.

Zhang, C., Yao, C., Liu, J., Zhou, Z., Zhang, W., Liu, L., Ying, F., Zhao, Y., & Wang, G. (2021). *Storydrawer: A co-creative agent supporting children's storytelling through collaborative drawing*.

Zhang, F., Markopoulos, P., Bekker, T., Schüll, M., & Paule-Ruíz, M. (2019). *Emoform: Capturing children's emotions during design based learning*.

Zhang-Kennedy, L., & Chiasson, S. (2016). *Teaching with an interactive e-book to improve children's online privacy knowledge*.

Interacting with Computers, 2025, 1–16
https://doi.org/10.1093/iwc/iwaf016
Article