Measuring product liking in preschool children: An evaluation of the Smileyometer and This or That method

Bieke Zaman a,*, Vero Vanden Abeele b, Dirk De Grooff a

a CUO | Social Spaces, iMinds - KU Leuven, Parkstraat 45 Bus 3605, 3000 Leuven, Belgium
b Corresponding author. E-mail address: Bieke.Zaman@soc.kuleuven.be

this is the authors’ post-print (i.e., final draft post-refereeing) version. Please cite as: Zaman, B., Vanden Abeele, V., De Grooff, D. (2012). Measuring product liking in preschool children: An evaluation of the Smileyometer and This or That method. International Journal of Child-Computer Interaction, 1(2), 61-70.

Abstract
This article provides a detailed evaluation of two summative evaluation methods, This or That and Smileyometer. The methods’ reliability and validity were examined, using a sample of 113 preschoolers (ages from 33 to 90 months). The results show considerable promise for This or That for measuring preferences across entertainment products. This or That resulted in reliable and valid responses in preschool children aged four and older. Nevertheless, low reliability and validity scores were obtained for This or That in preschoolers younger than four years old. As for the Smileyometer results, preschoolers’ responses clearly indicated an overrepresentation of extreme positive scores that were inconsistent with their actual product preferences. Consequently, the validity of these young children’s responses evoked through the Smileyometer was not supported.

Keywords: This or That; Smileyometer; preschoolers; evaluation; methods; user experience

1. Introduction
This research was conducted with the ambition to participate in the methodological quest for evaluation methods to be used in young, preliterate children for measuring product liking. Since the eighties, adults have started to acknowledge children’s capability to provide valuable contributions to research and their right to be directly approached when their behaviour and opinions are being studied (Buckingham, 1993; Darbyshire, 2005; Greig & Taylor, 1999; Mahon, Glendinning, Clarke, & Craig, 1996; Morgan, Gibbs, Maxwell, & Britten, 2002; Morrow & Richards, 1996; Neill, 2005). This new approach to research with children has been referred to as the ‘new sociology of childhood’, initiated by the Children’s Rights movement and inspired by the 1989 UN Charter that highlighted that “children do have valid and valuable views and opinions that deserve to be elicited and taken seriously” (Darbyshire, Schiller, & MacDougall, 2005, p. 467). This competence and right has to be considered for all children, young as they may be (Irwin & Johnson, 2005, p. 822). It is now recognized as problematic, even though convenient, to work with adult-originated accounts of children’s experiences without complementing this with alternative methods that include children’s voices directly (Darbyshire et al., 2005, p. 467; Lobe, Livingstone, & Haddon, 2007).

Recent developments in researching children have been looking towards ways of locating children’s lives in the centre of the research process, by strengthening the position of the child as an ‘active research participant’ as opposed to a ‘passive research object’ (Borgers, De Leeuw, & Hox, 2000; Druin, 1998; Lahman, 2008; Pole, 2007, p. 70). The child-centred approach regards children as competent in reporting their own experiences (Mayall, 1996), and consequently it has been marked by the prepositional shift from working ‘on’ to working ‘with’ children at all levels of the research process – from designing the study to data collection and data reporting (Lahman, 2008; Lobe, 2008).

In this context, child-computer interaction (CCI) researchers have been challenged to find the right methods for involving children in a child-centred design process. Many of these research
efforts have been put into finding the right evaluation methods that yield valuable results, on which informed product design decisions can be taken. In the middle and late design phases, empirical evaluative methods have become most prominent, of which verbalization methods, survey methods and observation (in combination with recording and logging of data) are very popular (Markopoulos, Read, MacFarlane, & Höysniemi, 2008). Verbalization methods and observation are usually of a formative kind whereas survey methods are typically part of the class of summative evaluation methods.

Although nowadays, surveys are mainly administered via paper-and-pencil or electronically in research with adults (Ozok, 2008, p. 1153), a face-to-face approach is more appropriate in the case of research with child respondents; and this is especially true for user experience evaluations (Read, 2008). The quality of such a face-to-face survey approach depends on the child-friendliness of the scales, which relates to the appropriateness of the content of the questions and response items as well as the type of scale.

Overall, summative product evaluation methods can be categorized in acceptance evaluation methods on the one hand and preference evaluation methods on the other (Hein, Jaeger, Tom Carr, & Delahunty, 2008). During acceptance testing, liking for a single product is measured without direct comparison to another product. In contrast, during preference testing, product preferences are measured by comparison of two or more products with each other.

This study evaluates two types of survey scales that have been used for summative, empirical product liking evaluations in the CCI community, namely smiley-based ordinal rating scales and dichotomous, nominal scales. Within these two categories, two empirical product evaluation methods are selected and investigated for their use with preschoolers, namely the Smileyometer (Read, MacFarlane, & Casey, 2002; Read, 2008) as an acceptance evaluation method and This or That (Zaman, 2009) as a preference evaluation method.

This article investigates the reliability and validity of This or That and Smileyometer in preschool children. Firstly, the main characteristics of acceptance and preference evaluation methods are described, together with their advantages and disadvantages. Then, the Smileyometer and This or That are introduced, followed by an overview of related work in which these methods have been evaluated. Secondly, the method of our empirical study is elaborated on. A presentation and discussion of the results follow, with special attention to reliability and validity testing. The article ends by summarizing the results and suggesting a concrete research agenda for further work.

1. Measuring product liking

Two general classes of summative, empirical evaluation methods can be distinguished for the measurement of product liking: acceptance evaluation methods and preference evaluation methods (Hein et al., 2008). In the next paragraphs, the main characteristics of these classes of methods will be discussed. Then, for each class, one method from the CCI domain will be selected for a more detailed discussion, namely Smileyometer and This or That respectively.

1.1 Acceptance Evaluation Methods

In acceptance evaluation methods, participants are presented with products individually, for which an evaluative judgment is elicited without direct comparison to other products. This class of methods typically relies on some kind of category scaling technique to quantify overall product liking and measure the degree of liking/disliking. In other words, these methods allow answering the question how well a product is (dis)liked. For instance, a 5-point Likert scale ranging from ‘dislike a lot’ to ‘like a lot’ can be employed to measure overall product liking. In both industry and academy, acceptance evaluation methods have typically been designed with a smiling face scale for research with young children, because they are considered to be more capable of and more motivated by using faces than words or numbers (Macklin & Machleit, 1989; Popper & Kroll, 2005).

Surveys based on face or smiley scales have a long tradition in medical research for measuring pain or anxiety (e.g. Bosenberg, Thomas, Lopez, Kōinsky, & Larsson, 2003; Wong & Baker, 1988), in psycho-social investigations for relationship assessment (e.g. Denham & Auerbach, 1995) and in marketing research (e.g. Macklin & Machleit, 1989). In 2001, Read and her colleagues (Read, MacFarlane, & Casey, 2001) were inspired by these smiley scales, and introduced the Smileyometer in the domain of child-computer interaction. The Smileyometer
measures how good the product experience was on a simple scale with smiley faces, as illustrated in Figure 1 (Read et al., 2001; Read & MacFarlane, 2006). The Smileyometer belongs to the family of 5-point Likert scales. Characterizing is that the typical text-based response items are replaced by smiley faces representing the child’s perceived level of product liking with the interactive product(s) under investigation (Read & MacFarlane, 2006). The ordinal response items are scaled with higher values reflecting a more positive evaluative judgment. More particularly, the Smileyometer’s scale ranges from ‘awful’ to ‘brilliant’. When using the Smileyometer in comparative product evaluations, a separate Smileyometer scale should be presented for each alternative that is evaluated (Read et al., 2002).

![Smileyometer instrument](chart)

**Figure 1** The Smileyometer instrument

As the Smileyometer has been devised as one particular method within a ‘Fun Toolkit’ of fun measures, it has been extensively compared with those Fun Toolkit instruments. In 2002, the results obtained with the Smileyometer, Funometer, Fun Sorter and Again-Again Table have been evaluated and compared in three studies with children within the age range of six to ten years old (Read et al., 2002). The descriptive statistical analyses suggested that young children scored most things as ‘Brilliant’ on the Smileyometer in comparison to the older children. Further, a correlation across the Funometer and Smileyometer was found, but not across the other measures. Additionally, previous work has shown little correlation between the fun ratings of seven- and eight-year-olds obtained via ranking and Smileyometer (MacFarlane, Sim, & Horton, 2005). The latter authors gave two possible explanations, one related to children’s assumed tendency to select the most positive options on the Smileyometer, another reason related to the fact that the questioning was not consistent over the measures (MacFarlane et al., 2005, p. 108).

In 2006, the degree of variance in the Smileyometer results was further investigated in two studies involving children within the age category of seven to thirteen years old (Read & MacFarlane, 2006). It was shown that the children aged twelve and thirteen showed a much higher degree of variance than the children of seven to nine years old.

The studies reviewed above have described the results obtained in children within the age spectrum of six to thirteen years old. Overall, these findings have indicated that children between nine and thirteen years old showed a higher degree of variance across the Smileyometer as compared to six- to nine-year-olds (Read & MacFarlane, 2006; Read, 2008). To date, the information available within the domain of CCI on how children younger than six perform in using summative evaluation measures is scarce. Outside the CCI domain, though, it has been found that young children respond in an extreme manner—which may relate to both the positive or negative pole- when rating emotion-based tasks (Chambers & Johnston, 2002).

Although the Smileyometer has been analysed in comparative studies, none of the studies described above have focused on an evaluation of the method within the age category of preschoolers. Overall, the Smileyometer and smiling face scales in general need further assessment to clarify their appropriateness and validity when used with preliterate children (Macklin & Machleit, 1989; Read, 2008). Moreover, the Fun Toolkit measures have mainly been validated by measuring correlations among the measures and by focusing on the variability obtained in the answers. It would be interesting to focus on other measures to judge the quality of the Fun Toolkit methods, such as internal, test-retest or inter-method reliability and several types of validity measures.

### 1.2 Preference Evaluation Methods

The second class of summative product liking methods consists of preference evaluation methods. In preference testing, participants are required to choose the product that is preferred over another, using the criteria imposed by the researcher. Product liking is then measured relatively, in terms of the explicit preference of one product over another. Preference testing measures can include paired-preference testing (one product is chosen over another, either
revealed through testing single pairs or multiple pairs) or ranked-preference testing in which three or more products are ranked in relative order from least to most preferred.

In 2004, Hanna et al. suggested the potential of pairwise comparisons for evaluations of interactive technologies with children (Hanna, Neapolitan, & Risden, 2004). One preference evaluation method in CCI is This or That (Zaman, 2009). This or That requires a child to choose his or her preferred product among two; This or That therefore belongs to the category of paired-preference testing, either as single or multiple paired testing. The pairs are presented in random order and counterbalanced across the participants to minimize order bias. Basically, the child is instructed as follows: “Each time [in the case of multiple pairs] I will give you two products. I want you to try them out and show me which one you prefer”. For (each) product pair, the child is asked to indicate his or her preference via six questions. More particularly, after having tried out the products, the child is encouraged to show which product was 1) “most fun?”, 2) “a little bit stupid?” and 3) “a little bit boring?” to directly question product preferences. As an indirect measure of preference, self-reported intentions are measured by the item 4) “Show me which product you would like to play again?” as indicative for attitude-behaviour consistency (Fishbein & Ajzen, 2010). Finally, the child’s desirability to obtain one of the products, as an indicator of the perceived favourableness (Hazlett & Benedek, 2007), is questioned by two items, namely 5) “Show me which product you would like to receive as a gift?” and 6) “which product you would like to take home?”. During this questioning, the product alternatives (or their pictorial representations) are preferably presented simultaneously to the child, who is then given the freedom to choose the preferred product, simply by pointing, and thus without having to verbalize the answers. The response options are hereby introduced by the researcher who ends each question item by asking “[question item]: this one or that one? That one or this one?”. For instance, the response options to the question “Show me which game was most fun?” are then formulated as follows: “this [game] or that [one], that [game] or this [one]?”. The alternation of ‘this or that’ should prevent one-sided perception of the answer. The researcher should thus ensure balancing the option he or she is referring to first, and allow “don’t know” and “(dis)like both equally” answers. Explicitly providing the latter answer categories avoids children mentioning random information to please the adult (Christensen & James, 2000) and relieves the child from possible internal conflicts (Sudman & Bradburn, 1974, p. 10).

This or That has already been included in empirical studies with the aim of analysing the quality of its results obtained in children. In 2009, This or That was empirically validated in three experiments which showed evidence for the uni-dimensionality, validity and internal consistency of its pairwise comparison scale when used with preschoolers (Zaman, 2009). Further, This or That has also been compared against other measures. In 2012, Sim & Horton compared it with the Smileyometer, Fun Sorter and Again-Again Table in children aged seven and eight years old (Sim & Horton, 2012). The latter study showed that This or That was an internally reliable instrument. Both This or That and the Fun Toolkit allowed for preference testing in a consistent manner. When comparing This or That and the Again-Again Table, the results were rather consistent, pointing towards the same product preferences, even though the Again-Again Table provided an explicit ‘maybe’ response option, which was not foreseen in Sim et al.’s This or That research protocol (Sim & Horton, 2012). The results of the Smileyometer that was administered after product experience, -although suggesting a preference-, did not significantly correlate with the This or That results.

Overall, some tendencies with regard to the use of This or That with children have been revealed in previous research. In Zaman (2009) This or That was tested in preschoolers, but in isolation from other related measures. In contrast, Sim and Horton (2012) did evaluate This or That in comparison to the results obtained via other related measures. However, the latter authors did not involve preschoolers in their study but seven- to eight-year-olds. As it is insightful to assess the usefulness and validity of a survey instrument and its scale in the light of the results obtained with (an)other relevant measure(s) (Hourcade, 2007; Markopoulos et al., 2008), and because the This or That method is still young, the method would benefit from a more complete validation.

---

1 The question “a little bit boring?” was added to the scale in order to consider the option to compensate for possible language issues in the interpretation of the question “a little bit stupid”. The This or That questions items were originally designed in Dutch where ‘stupid’ only has a negative connotation (Zaman, 2009) in contrast to the English term ‘stupid’ which may be interpreted as both a positive or negative construct (Sim & Horton, 2012).
and a broader discussion of its reliability and validity when administered with preschool-aged children.

1.3 Pros and Cons: Acceptance versus Preference Testing

Both acceptance and preference evaluation methods have advantages and disadvantages. In essence, preference testing is relatively simple and easy to organize and conduct. Considering the method’s complexity from the participant’s perspective, it presents a rather familiar and unambiguous design. When people are asked to form an opinion about a product, they are likely to do so by comparing the product on a binary scale of attributes (e.g. bad-good) against a set of about two alternatives (Fishbein & Ajzen, 2010; Peter & Olsen, 1999). In that perspective, preference testing presents a natural choice situation. This is particularly important for research with young children. Previous research has confirmed that for young child participants preference testing is less complex compared to acceptance testing (Guinard, 2000; Leon, Couronne, Marcuz, & Koster, 1999; Popper & Kroll, 2005).

To account for the complexity of acceptance testing in young respondents, smiley face scales have been designed as a more entertaining and simpler scale alternative, and therefore considered to keep the child’s attention closer to the task (Macklin & Machleit, 1989; Popper & Kroll, 2005). It should be acknowledged, though, that young respondents might not correctly understand what these scales represent, when they rely on what the faces show (a happy person) rather than on what these are supposed to represent (how the product makes you feel). Misinterpretation of the faces intended to show a degree of ‘dislike’ may then occur as if they were expressing feelings of angeriness (Popper & Kroll, 2005).

One of the main advantages of acceptance testing is that products are evaluated individually, so that data can be compared across studies. In contrast, preference evaluation methods measure sample preference within the test set and exclude comparisons with products that are not included in the test (Hein et al., 2008).

Further, acceptance testing allows measuring the degree of product liking, data from which preferences can be inferred. This is both an advantage and a disadvantage. The advantage is that acceptance testing provides a direct indication of the magnitude of the measured preferences. To illustrate, when a child rates his/her interaction with product ‘A’ as ‘brilliant’ via the Smileyometer instrument, then the researcher can link this directly with the numerical ordinal value of 5 on the underlying Likert-scale that ranges from 1 to 5. This contrasts preference testing, in which preference strength for one product can only be measured indirectly, by relating it to the concrete choice situation of alternative products. For instance, the repeated administration of the This or That questions reveals the relative strength of one’s preference for one or the other product in a choice context of two products that are being evaluated. The researcher hereby relies on the ambivalence of the answers to calculate the relative preference strength. This means that when a respondent has expressed his or her preference for product A for five of the six question items and for product B for one item only, this results in a relative preference of .83 for product A and .17 for product B. This indirect measure of preference strength is based on the assumption that people who have a strong preference for one particular product will significantly link that product more to the positive pole of the preference construct than the alternative(s). Evaluative judgments formed through a mix of positive and negative evaluations are weaker than those built up from merely positive attitudes (Fishbein & Ajzen, 2010).

The direct measurement of the degree of product liking in acceptance testing has its downside as well. The rating scales used in acceptance testing require the respondents to provide fine-grained intensity judgments, which brings an additional level of complexity (Chambers & Johnston, 2002). Moreover, in acceptance testing, the intervals between the scale parts are not necessarily perceptually equal. In user experience testing, in which the user experience construct is typically operationalized with positive aspects, one may wonder whether the ‘neutral’ option reflects a true ‘zero’ point in the measurement. Additionally, although preferences can be inferred from acceptance testing data, one should realize that these preferences are only indirectly defined. Participants are not asked to compare the alternatives directly, which contrasts preference testing as the latter does measure preferences based upon direct comparisons of products.
Although one can consider the simplicity of preference testing as its main advantage, the research design becomes more complex and cumbersome in the case of multiple paired comparisons in a product set of more than two alternatives. For instance, in a product set of four products, six pairs are formed; this increases to ten pairs in the case of a product set of five and 15 pairs in a product set of six. Consequently, the time required for the test greatly increases with the number of product alternatives considered in multiple-paired comparison testing. If this is the case, ranked-preference testing in which all products are compared simultaneously instead of pairwise can provide a less extensive alternative for preference testing.

Additionally, another disadvantage in preference testing is that the respondents' preference judgments greatly depend on the (meaningfulness of the) alternatives that are given; hence another choice context can elicit different results.

To conclude this discussion on the pros and cons of these two classes of product liking methods, we refer to Guinard (2000) who has provided an overview of these methods’ appropriateness for sensory evaluations in children. His literature overview has shown that for the age group of two to three-year-olds, paired-preference testing is the only reliable method. For preschoolers older than four years old, preference ranking extending over two or more products, and acceptance scales (3- to 9-points) can also be considered (Guinard, 2000). It is not clear, however, whether these results that have proven valid for sensory evaluations such as measuring food preferences and perceived sweetness intensity, also hold for preference evaluations related to interactive, entertainment product experiences.

2. Method

2.1 Purpose
The aim of this study is to compare This or That with Smileyometer for measuring product liking in preschoolers and more particularly to:

- determine the reliability of these methods in preschool children. This objective will provide an indication of the internal consistency of the results obtained by the measures.
- determine how well these methods compare with each other with regard to their validity in order to assess whether these methods provide a good test of predictability of overt behaviour.

2.2 Participants
As for the definition of a preschooler, we relied on two defining characteristics that have been put forward as relevant in the definition of a preschooler-based sample. More particularly, we recruited children along the age spectrum of two to seven years old, an age category that in literature has been referred to as the ‘preoperational phase of children’ and further divided in a ‘preconceptual sub stage’ for children between two and four years old, and an ‘intuitive sub stage’ for children aged four to seven years (Mussen & Piaget, 1983; Piaget, 1970). Our study aimed at involving children from both stages, in the remaining parts of this article simply referred to as respectively the youngest and oldest preschoolers. Moreover, in accordance to Miller and Aloise, we also followed the literacy criterion by defining a preschooler as a child who has not yet received primary education, and thus cannot yet read or write (Miller & Aloise, 1989, p. 258).

In total, written parental informed consent was obtained for 113 preschoolers, half boys (n = 56) and half girls (n = 57 girls), ranging in age from 33 to 90 months (M = 58.4, SD = 14), see Table 1. The majority of the child participants belonged to the intuitive sub stage (73%), only one quarter of the children belonged to the preconceptual sub stage (27%).

Table 1 Age distribution of participants

<table>
<thead>
<tr>
<th>Age (months)</th>
<th>33-35</th>
<th>36-47</th>
<th>48-59</th>
<th>60-71</th>
<th>72-83</th>
<th>84-90</th>
</tr>
</thead>
<tbody>
<tr>
<td>n</td>
<td>8</td>
<td>16</td>
<td>39</td>
<td>23</td>
<td>25</td>
<td>2</td>
</tr>
</tbody>
</table>

In our sample, all preschoolers were illiterate. The study took place during summer holidays, only a few weeks away from the start of the school year in which the oldest children would start learning to read. In Belgium, where the study was organized, children start reading and enter
primary education in the year that they turn seven years old. The study took place at two recreation areas; it concerned local initiatives that take care of children during holidays at (almost) no fee, bringing together children from a variety of backgrounds.

2.3 Product Set
This or That and Smileyometer were surveyed in the context of an empirical evaluation study of two keyboard interaction games that were designed for a preschool-aged target group: the Toewie game (see Figure 2, left) (Husson, Jelle, 2006) and the MamaMoe en de kraai game (see Figure 2, right) (Mindscape Northern Europe, Wieslander, Nordqvist, Grammofon Multimedia AB- part of Norstedts Barnmedia, & Published by Lannoo, 2007). None of the children had ever played one of these games before.

![Toewie game](a) ![MamaMoe game](b)

Figure 2 Toewie game on the left (a), MamaMoe game on the right (b).

2.4 Procedure
The study followed a within-subject research design. The main procedure consisted of the following steps: a) rapport building, b) game play, c) Smileyometer, d) the repetition of step b and c for the evaluation of the second game, e) This or That, f) reward and g) natural choice game selection (see also Table 2).

Table 2 Overview procedure within-subject study, including eight phases.

<table>
<thead>
<tr>
<th>Step 1</th>
<th>Step 2</th>
<th>Step 3</th>
<th>Step 4</th>
<th>Step 5</th>
<th>Step 6</th>
<th>Step 7</th>
<th>Step 8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rapport building</td>
<td>Game play (1st game)</td>
<td>Smileyometer (1st game)</td>
<td>Game play (2nd game)</td>
<td>Smileyometer (2nd game)</td>
<td>This or That</td>
<td>Reward</td>
<td>Natural choice selection</td>
</tr>
</tbody>
</table>

a) **Rapport building**: The researcher led each child individually to the test room and engaged in small talk as a means to build rapport.

b) **Game play**: The actual study started by inviting the child to explore and play one of the two games until he or she had finished the first game level. Finishing the first level took about five minutes for each game. Half of the children started playing the Toewie game first, the other half began with the MamaMoe game.

c) **Smileyometer**: After this first game was played, the Smileyometer was presented to the child. The game was not shut down and the laptop remained on the table while the child was handed over the Smileyometer instrument. In an attempt to avoid misinterpretation of the faces, the child was asked to point to the location on the scale that represented how (s)he felt while playing the game. Explicitly asking the child to express his or her emotional user experience judgment is important, as children are likely to rely on what the smiley faces show (a happy person) rather than on what these are supposed to represent (how the product made them feel) (Popper & Kroll, 2005).

**Repetition steps b) and c)**: The previous steps, i.e. the game play and the Smileyometer evaluation, were repeated for the evaluation of the second game. Note that the order in which the games were introduced to the child participants was counterbalanced. Both games were presented on identical laptops (see Figure 3). Only one researcher was
present, sitting next to the child in child-size chairs to give support and make the child feel at ease. The researcher took quick notes with pen and paper; for more detailed notes, the video footages were relied upon.

Figure 3 Research set up: child and researcher sitting in child-size chairs during game play and game evaluations. The games were installed on a separate laptop. Both laptops had similar hard- and software characteristics.

d) This or That: After the game explorations and the Smileyometer evaluations, the This or That instrument was administered. Children had to express or point to their favoured game as a response to each of the six This or That questions.

e) Reward: The child was then given a biscuit as a reward for participation. This reward was not mentioned before the study in order not to reinforce the idea that there were good or bad answers or behaviours that were to be rewarded, which could make the children less critical and more prone to social desirability (Markopoulos et al., 2008).

f) Natural choice game selection: Finally, children were told that the researcher ‘happened to have some time left’ so that they could select one of the two games to play a second time if they wanted to ‘because they did so well during the study’. This way, the option – they were not forced – to play one of the two games again was presented as some kind of a reward to elicit a natural choice situation. In the time between This or That and the natural choice game selection, children had to go to the opposite corner of the room to receive their biscuit. It would prevent children from staying at their place, and hence, possibly selecting the last played game to be played again as for convenience sake.

2.5 Data treatment
In the following section, the data treatment for the reliability and validity testing will be explained. Then, the treatment of missing values will be reported upon. The section ends by an overview of the results of exploratory data analyses.

2.5.1 Data treatment for reliability testing
In the data treatment for reliability testing, the This or That values for the variables most fun, gift, home and again obtained the value ‘1’ when the child preferred Toewie over MamaMoe on the positive This or That items, and ‘0’ when MamaMoe was preferred over Toewie. For the This or That items that were negatively formulated (i.e., the question items ‘a little bit stupid’ and ‘a little bit boring’), the scores were reversed (stupid_rev and boring_rev).

The Smileyometer was administered twice, once after the first game was played (Smile_first), and another time after the second game was played (Smile_sec). The corresponding variables, Smile_first and Smile_sec, were coded with values in the range of -2 to 2. The scores on the 5-point Likert scale obtained a value of 1 and 2 in the case of liking for a product (‘really good’ and ‘brilliant’ smiley, respectively), a value of -1 and -2 in the case of disliking (‘not very good’ en ‘awful’ smiley, respectively); neutral (‘good’ smiley) was scored 0.

2.5.2 Data treatment for validity testing
The natural choice game selection was essential in assessing the validity of the research methods as it provided a test of the methods’ predictability of overt behaviour. The natural choice game selection foresaw in a manifest response of game preference, revealing the coherence between the preference assessed by each method on the one hand, and a more natural choice on the other. Therefore, validity of the This or That method is supported if b) the overall relative This or That score for the selected game \( Tsg \) is greater than the This or That score for the not-selected game \( Tng \). Validity of the Smileyometer would be supported if a) the Smileyometer score for the selected game (i.e. game selected by natural choice) \( Ssg \) is greater than the Smileyometer score for the not-selected game \( Sng \).

Therefore, post hoc, the results for each This or That question item were binary coded with ‘1’ in the case of a match with the selected game and ‘0’ in the case of a mismatch (the values for the items ‘boring’ and ‘stupid’ were reversed). For instance, when a child indicated that Toewie was the most fun game on This or That, the question item ‘show me which game was most fun’ was coded with value 1 in the case that the child also selected Toewie as the natural choice game selection. In this manner, 12 additional variables were computed, 6 representing the This or That values for the selected game \( Tsg \): \( Tsg\text{-}fun, Tsg\text{-}StupidRev, Tsg\text{-}BoringRev, Tsg\text{-}Again, Tsg\text{-}Gift\text{ and } Tsg\text{-}home; \) the other 6 representing the This or That values for the game that was not selected to play again \( Tng \): \( Tng\text{-}fun, Tng\text{-}StupidRev, Tng\text{-}BoringRev, Tng\text{-}Again, Tng\text{-}Gift \text{ and } Tng\text{-}home. Finally, from these the overall relative This or That scores for the selected game \( Tsg \) and not-selected game \( Tng \) were calculated. Obviously, from the above follows that \( Tng = 6 - Tsg \).

For the validity testing, the Smileyometer values were also considered with respect to the game that was chosen as the natural choice game selection. This resulted in the computation of two additional variables, one variable that represented the Smileyometer values for the selected game \( Ssg \), and the other variable represented the Smileyometer values for the game that was not selected to play again \( Sng \).

2.5.3 Treatment of missing values

The values of \( Tsg\text{-}fun, Tsg\text{-}StupidRev, Tsg\text{-}BoringRev, Tsg\text{-}Again, Tsg\text{-}Gift \text{ and } Tsg\text{-}home \) as well as the values of the variables \( Tsg, Tng, Ssg \text{ and } Sng \text{-}values \) depended on the selection of the final game at the end of the study. We could not include the children who did not select a final game in the analyses on these variables, and consequently nine children, of which three girls and six boys, were not included. Two of them belonged to the preconceptual sub stage, seven to the intuitive sub stage.

Moreover, in preference testing in which the options of ‘(dis)like both equally’ and ‘don’t know’ are included, the data analysis should be performed only on those responses that show a definite preference for one of the two products (Moskowitz, Beckley, & Resurreccion, 2012). Hence, responses that corresponded with one of these three responses were withdrawn for further analysis in our study. This resulted in missing values for This or That in five boys and two girls, two of them belonged to the preconceptual sub stage, five to the intuitive sub stage.

As for the missing values in Smileyometer, there was one boy from the intuitive sub stage and one girl from the preconceptual sub stage who did not provide a rating for the Smileyometer on the game that was played first \( (Smile\text{ first}) \).

2.5.4 Exploratory data analyses

Exploratory data analyses were performed to investigate the normality distribution of the data in order to select the right statistical tests for our analyses. As for the This or That items, the Shapiro–Wilk normality test values were significant, indicating non-normal data, and this in both age groups. The test statistic for the items were \( W\text{-}fun(107) = .475; W\text{-}gift(107) = .485; W\text{-}stupidRev(107) = .575; W\text{-}home(107) = .524; W\text{-}again(107) = .515; W\text{-}boringRev(107) = .555; \) all at \( p < .001 \).

A Shapiro–Wilk normality test was also calculated on the Smileyometer variables, resulting in significant values \( p < .001 \), clearly indicating that the distributions for the \( Ssg \text{-}values \) and \( Sng \text{-}values \) were not normal (for the youngest preschoolers: \( W(27) = .707, p < .001 \text{ and } W(27) = .773, p < .001 \), respectively; for the oldest preschoolers \( W(80) = .510, p < .001 \text{ and } W(80) = .621, p < .001 \), respectively).

This or That has been introduced as a uni-dimensional scale consisting of various items that all refer to one aspect of a higher ordered construct (Zaman, 2009). In order to reveal whether there
was any latent variable, exploratory factor analysis (EFA) was computed. The principal axis factoring method was selected because our data were significantly non normally-distributed (Costello & Osborne, 2005). The calculated scree plot indicated a point of inflexion on the curve at two, which suggests a uni-dimensional scale. Moreover, the principal axis factoring extraction method confirmed that no more than one factor could be extracted using the eigenvalue > 1 rule. The total variance that could be explained by the unique factor was 54% of variance for initial eigenvalues (eigenvalue = 3.212).

3. Results
In what follows, the results of the reliability and validity testing are discussed in detail.

3.1 Reliability
Firstly, the reliability test results of This or That will be presented, followed by a discussion on the Smileyometer’s reliability.

3.1.1 This or That
Coefficient alpha was computed for most fun, gift, home, again, stupid_rev and boring_rev. The overall Cronbach’s alpha of the 6-item dichotomous This or That scale yielded $\alpha=.825$ ($n = 107$). Each item correlated with the overall score from the scale. There were no corrected item-total correlations below .3. Furthermore, none of the items substantially affected reliability if they were deleted, since no alpha value for deleted items was greater than the overall Cronbach’s alpha. Additionally, we split the sample by age, into the young and older preschoolers, to investigate age differences. We found remarkable lower reliabilities for the 6-point scale in the youngest subsample -preconceptual sub stage- ($\alpha=.654$; $n = 28$), compared to the $\alpha$-value in the oldest preschoolers of the intuitive sub stage subsample ($\alpha=.850$; $n = 79$).

Table 3 shows the extraction communalities for each This or That item ($M = .447, SD = .124$), indicating the proportion of common variance present in each variable accounted for by our unique factor. Splitting the data by age, the extraction communalities were remarkably lower for the youngest preschoolers ($M = .455, SD = .234$) in comparison with the extraction communalities for the oldest preschoolers ($M = .509, SD = .235$). The results further suggested a uni-dimensional scale in the oldest preschoolers, with the eigenvalue of 3.472 that could explain the total variance of 58%. In contrast, in the youngest children there were two factors revealed from the analysis with an eigenvalue of 2.311 and 1.317 with the cumulative explained variance reaching 60.452%.

There are two issues that impede a conclusive interpretation of the results, namely the emergence of a second factor in the youngest preschoolers and the low communalities across all preschoolers. Given that in social sciences common magnitudes of communalities are between .40 and .70 (Costello & Osborne, 2005), our data score rather poor with an average communality of .447 ($SD = .124$). Variables with low extraction communalities are normally potential candidates for omission from the final scale. Nevertheless, caution is required in interpreting these results. In studies where low communalities emerge in the data, it is advisable to select a larger sample to determine whether or not the factor structure and individual items are valid (Costello & Osborne, 2005). One assumption for factor analysis is that it needs large samples for stable estimates (Comrey & Lee, 1992). In addition to the large-sample requirement, it is also questionable what common magnitudes of communalities are for verbalizations in surveys with preschool children. To our knowledge, there are no similar results in the CCI domain on product liking scales for young children to compare our results to.

3.1.2 Smileyometer
As the Smileyometer was administered as a single item measure, the Cronbach’s alpha internal consistency reliability test statistic could not be calculated. Nevertheless, it is worthwhile to reflect upon the appropriateness of the scale to produce reliable results that are consistent from one test administration to the next.

A Shapiro-Wilk normality test was calculated on the Smileyometer scores for the first administration (Smile_first) and second administration (Smile_sec), all resulting in significant values (p < .001), clearly indicating that the distributions were not normal (in the youngest preschoolers: W_Smile_first(29) = .590, p < .001 and W_Smile_sec(29) = .833, p < .001; for the oldest preschoolers W_Smile_first(82) = .518, p < .001 and W_Smile_sec(82) = .611, p < .001).

The data distribution shows a negative skew due to the overrepresentation of the extreme positive rankings (see Figure 4). For the Smileyometer administered after the first game was played (Smile_first), the median value was 2 (M = 1.56, SD = .96). The median value for the Smileyometer administered after the second game (Smile_sec) was played was identical, 2 (M = 1.25, SD = 1.12). The category 'really good' (16 %) and 'brilliant' (75 %) constituted 91 % of the answers on the first Smileyometer (Smile_first); these categories represented respectively 29 % and 56 % of the answers on the second Smileyometer (Smile_sec), good for 85 % of all the answers (see Figure 4).

![Figure 4](image)

**Figure 4** Proportion Smileyometer answers (in percentages) with regard to the game that was played first (Smile_first) and the second game (Smile_sec)

It is not only interesting to reflect on the data distribution, but also to verify whether these extreme responses reflect a true judgment of children’s preferences. Although this will be analysed in detail in the next section, it is interesting to already report on the descriptive statistics of the Smileyometer scores for the game that has been selected (Ssg) and the game that has not been selected (Sng) in the natural choice situation. At least for half of the children the 'brilliant' smiley, - the most extreme positive value on the Smileyometer Likert scale- represented the Ssg-value (52 % of the youngest preschoolers and 74 % of the oldest preschoolers (see Figure 5). This implies that the game that was selected in the natural choice situation received very high ratings.

However, considering the ratings for the game that was not selected to play again at the end of the study (Sng), our results showed that many children gave it very high score too, the brilliant smiley was chosen by 48 % of the youngest preschoolers and 68 % of the oldest preschoolers –see Figure 5.
3.2 Validity

In what follows, the validity test results of This or That will be presented, followed by a discussion on the Smileyometer's validity.

3.2.1 This or That

The validity of the This or That instrument was assessed by investigating whether the preferences revealed through the verbal This or That reports would predict the overt behaviour in the final game selection choice. We would expect higher overall This or That scores for the selected game ($T_{sg}$) than for the not-selected game ($T_{ng}$). For $T_{sg}$, 6 was the median value which is the highest score on the $T_{sg}$-variable. Shapiro-Wilk test showed that there were significantly more positive rankings in which $T_{sg}$ was greater than $T_{ng}$ ($z = -7.896, p < .001$). More particularly, the Wilcoxon Signed Ranks test specified that in 81% of the cases the $T_{sg}$-value was greater than the $T_{ng}$-value. Only for 8% of the cases was $T_{sg}$ lower than $T_{ng}$. In 11% of the cases, the values were identical. The test statistic was significant, confirming that our preschoolers significantly rated the selected game more often positively on the This or That scale than the game that was not selected to play again.

When splitting by age, the tendencies identified above were also present in the oldest preschoolers ($z = -7.386, p < .001$). In 86% of the cases, the $T_{sg}$-values were greater than the $T_{ng}$-values. Although there were more cases in which the $T_{sg}$-values were greater than the $T_{ng}$-values in the youngest preschoolers (16 against 4 respectively, and 6 ties), the Wilcoxon Signed Ranks test did not show any significant difference ($z = -2.692, p > .05$).

3.2.2 Smileyometer

The non-parametric Wilcoxon Signed Ranks t-test was calculated to investigate the Smileyometer's instrument predictive validity. In the previous section, the descriptive statistics on the Smileyometer results reflected mainly positive scores. To test the predictive validity, we would expect the Smileyometer scores for the selected game ($S_{sg}$) to be significantly greater than the $S_{ng}$-scores, even though the scores are all mainly positive. Overall, the median score for both $S_{sg}$ and $S_{ng}$ was 2, which represented the highest value on the 5-point Likert scale with range -2 to 2. There were 27 cases in which the $S_{sg}$-score on the Smileyometer was greater than the $S_{ng}$-score; for 25 cases the $S_{sg}$-score was lower than the $S_{ng}$-score and in 55 out of the 107 cases the scores were equal. The test statistic revealed that there was no significant difference between the scores ($z = -0.235, p > .05$).

Splitting the data by age, we found a median $S_{sg}$-score of 2 and a median $S_{ng}$-score of 1 in the youngest preschoolers. The results of the Wilcoxon t-test indicated that it did not concern any significant difference ($z = -1.097, p > .05$). There was also no significant difference in the $S_{sg}$ and $S_{ng}$-scores in the oldest preschoolers, with a median score of 2 for both $S_{sg}$ and $S_{ng}$-values ($z = -0.629, p > .05$).
4. Discussion

Both This or That and Smileyometer have been designed to evaluate interactive products with children. In this study, the measures were tested in preschoolers from the preoperational and intuitive substage. Overall, in doing research with children, it is important to critically analyse what information is desired to obtain and how, compared to what information children are capable of dealing with and what information they can provide. The younger the age group, the more challenging it is to devise valid, reliable and feasible test methods because of young children's cognitive and verbal limitations to reliably perform as research participants.

Hence, in interpreting our results, it may not surprise that we did not find the same reliability and validity scores as in adults. Preschoolers have shorter attention span and cognitive capacities that are not yet fully developed. Consequently, extreme answers, random answers, low repeatability and inconsistency may occur during research in young child participants. Taking this into account, our results are in line with previous research (Guinard, 2000), suggesting that when preschoolers are involved in summative product liking testing, preference testing is more appropriate than acceptance testing. The first method is less complex and provides a better predictability of behavioural preferences.

Discussing children's use of the Smileyometer in more detail, it appears that when preschoolers were evaluating their game experiences via the Smileyometer, their responses did not reflect a pattern in which the preferred game was receiving significant higher ratings than the least preferred option. When interpreting and using the Smileyometer, preschoolers might have been evaluating the games individually as a measure of single fun instead. As the underlying goal of games is to generate fun and pleasurable experiences (Koster, 2004; Sherry, Lucas, Greenberg, & Lachlan, 2006; Vorderer, Klimmt, & Ritterfeld, 2004), and especially games for children, it is not unlikely that preschoolers evaluate their single game experiences positively.

Moreover, there might have been a mismatch between the emotions pictured on the Smileyometer scale and the emotions that were likely to be elicited by an entertaining game. None of the games was likely to make a young child really sad, which is what the negative smiley faces might symbolize to the children. In that respect, it would have been interesting to introduce the Smileyometer as a unipolar scale offering options to differentiate on the positive spectrum without showing (extreme) negative smileys. In contrast, for research on pain, relationship assessments, or food sensory perceptions, domains in which for years face scales have successfully been used with children, it is more likely that the complete scale spectrum is perceived as relevant. To illustrate, in reporting on their health, young children aged six and seven years old have been found to use more extreme responses on Likert scales. Contrary to user experience judgments, these health-related judgments did not only tend to reflect more extreme opinions in the positive sense, but also in the negative sense (Rebok et al., 2001). Hence, the characteristics of the object or experience that is being evaluated might greatly affect the Smileyometer's appropriateness.

A limitation that urges us to treat the findings in the present study with caution, deals with our sample. There was an underrepresentation of preschoolers from the intuitive sub stage, which may have impacted the statistical solidity of the results. More particularly, this made it difficult to provide a conclusive answer as to how the methods perform within the age groups. A larger sample would have allowed us to draw more solid conclusions with respect to validity of the results in the youngest preschoolers and the variance that could be explained by the factors of the verbal scales. In addition, it would have been informative to have more details on the children themselves (e.g. ethnicity, socio-economic status, temperament...) and the possible impact of these demographic variables on the results. Further, the impact of the demographics of the researcher should have deserved more attention as well (Markopoulos et al., 2008; Popper & Kroll, 2005).

Additionally, we suggest follow-up studies to compare the data from This or That with the Fun Sorter's results in order to account for the compatibility in the measures. Indeed, the This or That technique resembles Read et al.'s Fun Sorter (Read et al., 2002) in that it also allows for evaluating product preferences on several comparative fun items.

Finally, another line of thinking concerns the deployment of Smileyometer as a multi-item instrument by defining a list of items that are more directly related to the sub constructs of user experience as employed in This or That. This would allow a better comparison with This or That,
as well as more nuanced hedonic measurements. Besides, previous research on multi-item pictorial scales for preschoolers used in marketing research has shown considerable promise (Macklin & Machleit, 1989).

**CONCLUSION**

New methods are emerging within the domain of child-computer interaction. In order to move the field forward, it is imperative to critique these methods and establish their potential and limitations, which ultimately aids the community’s understanding of the reliability and validity of the evaluation methods. While older children may effectively use verbal scales, even if some interviewer assistance is required, there is still no consensus on how best to test preschoolers. The present study examined the validity and reliability of Smileyometer and This or That for evaluating interactive entertainment products with preschoolers.

The results provide evidence for the internal reliability and uni-dimensionality of This or That’s pairwise comparison scale. Furthermore, the scale’s predictive validity was also supported. It should be noted though, that these findings only held true for the oldest preschoolers, aged four and older. More work is needed to explore the feasibility of this kind of paired-preference testing when employed with the youngest preschoolers. With regard to the Smileyometer, our results did not provide evidence of its predictive validity. There was an overrepresentation of the most extreme positive scores for both the preferred and non-preferred game in the youngest and oldest preschoolers. Consequently, when used with preschoolers, researchers should guard themselves in interpreting the Smileyometer results as a prediction of overt behaviour and preferences or as an indication of preference strength.

To conclude, more research is needed on both This or That and Smileyometer. To us, the avenues for further work should concentrate on:

   a) *deciding upon the final list of This or That question items*: Special attention should be paid to the validity of the This or That instrument for preference evaluations in preschoolers already from the age of two onwards.

   b) *investigating the potential of multi item pictorial scales*: If the Smileyometer is used in isolation of other methods, one cannot measure subtler user experience aspects other than ‘fun’. Hence, it is interesting to analyse how the current use of the Smileyometer as a single item measure of fun can be extended towards a multi item measure, accounting for aspects such as challenge, fantasy or creativity. The importance of designing good, unambiguous pictorial representations for each of these aspects may not be underestimated, though.

**ACKNOWLEDGEMENTS**

We would like to thank all children who participated in our study as well as the animators of the summer day play initiative for having given us the possibility to conduct our research on their playgrounds. We are also grateful to the TOEWIE developers' team because we could use their game for our research project. Finally, we express our gratitude to the KU Leuven, GroepT and iMinds (Interdisciplinary Institute for Technology), a research institute founded by the Flemish Government, for supporting our work.

**REFERENCES**


Lobe, B. (2008). Integration of Online Research Methods. Faculty of Social Sciences, University of Ljubljana.


