Exploring Verbalization and Collaboration of Constructive Interaction with Children

Benedikte S. Als, Janne J. Jensen, and Mikael B. Skov

Department of Computer Science, Aalborg University,
Fredrik Bajers Vej 7, DK-9220 Aalborg East, Denmark
{als, missj, dubois}@cs.aau.dk

Abstract. Constructive interaction provides natural thinking-aloud as test subjects collaborate in pairs to solve tasks. Since children may face difficulties in following instructions for a standard think-aloud test, constructive interaction has been suggested as evaluation method when usability testing with children. However, the relationship between think-aloud and constructive interaction is still poorly understood. We present an experiment that compares think-aloud and constructive interaction. The experiment involves 60 children with three setups where children apply think-aloud or constructive interaction in acquainted and non-acquainted pairs. Our results show that the pairing of children had impact on how the children collaborated in pairs and how they would afterward assess the testing sessions. In some cases, we found that acquainted dyads would perform well as they would more naturally interact and collaborate while in other cases they would have problems in controlling the evaluations.

1 Introduction

Children have been characterized as not just short adults, but as independent individuals with their own strong opinions, needs, likes, and dislikes, and they should be treated as such. The design and evaluation of children's technologies have received increased attention during the last several years [7, 8]. Druin [9] provides a classification of involvement where children play the roles of users, testers, informants, or design partners. The four roles encompass different levels of engagement and impose different opportunities and limitations. All roles involve different kinds of usability tests where children participate as subjects, for example user [29], tester [19], informant [10], and design partner [6].

Some research studies have started to investigate the roles of children in usability tests, cf. [18, 21]. Nielsen [26] suggests that evaluators should use a variation of think-aloud called constructive interaction [16, 23] (also known as co-discovery learning), since it may be difficult to get children to follow the instructions for a standard thinking-aloud test. Constructive interaction involves two test subjects collaborating in trying to solve tasks while using a computer system [27]. Even though constructive interaction with children seems appropriate, the relationship between think-aloud and constructive interaction in usability testing with children is poorly understood. A number of questions still need to be addressed and answered: 1) How do children think-aloud and collaborate in constructive interaction 2) How should pairs of chil-

M.F. Costabile and F. Paternò (Eds.): INTERACT 2005, LNCS 3585, pp. 443 – 456, 2005.

dren be configured in constructive interaction? 3) How do children perceive the testing situation during constructive interaction?

In this paper, we investigate and address the above stated questions by looking at how children perform and behave in constructive interaction during usability testing. Our particular focus is on how the children behave and perceive a testing situation when involved a traditional think-aloud test compared to constructive interaction tests. First, we present an experimental design involving 60 children participating in two different configurations of constructive interaction and a traditional think-aloud. Secondly, we present results from the evaluations by illustrating how the children applied the think-aloud protocol and collaborated and further how they perceived the situation. Finally, we outline three lessons on involving children in usability testing.

2 Constructive Interaction in Usability Testing with Children

Nielsen [26] claims that constructive interaction is preferable over think-aloud when conducting usability evaluations with children. Where children face difficulties in following the instructions for a think-aloud test, constructive interaction comes closer to their natural behaviour, since the children work in pairs and collaborate in solving the tasks. Due to the fact that different the children's ability to verbalize their thoughts and feelings during a test, Hanna et al. [13] propose some adjusted guidelines where they reflect upon common target age ranges. Jensen and Skov [15] found that 67% of the research on interaction design and children applied some sort of systematic field or laboratory evaluations. Furthermore, some studies have explored different methods for conducting usability evaluations with children; one studied the effectiveness of co-operative evaluations (think-aloud) and co-discovery evaluations (constructive interaction) [1, 21], where another studied different method's effectiveness to elicit verbal comments from children [18]. The first compared the difference in total number of identifies usability problems identified by four subjects or four pairs, and found only negligible differences between the two methods.

Miyake [23] states that constructive interaction inherently integrates a number of opportunities and limitations. An advantage is that the test subjects naturally use think-aloud in their collaboration, one of the disadvantages is that the might aim for different strategies for learning and using computers. Furthermore, since constructive interaction requires twice as many test subjects as think-aloud, in order to conduct the same number of usability sessions, it is typically more expensive [26]. Configuring pairs for construction interaction includes two important steps [16]. First, test subjects must be selected and acquired for the usability test [27]. Secondly, usability evaluators are further faced with challenges of pairing subjects when adapting constructive interaction as evaluation technique. A number of challenges seem to influence the configuration of subjects in constructive interaction.

First, one challenge concerns the level of expertise. The level of expertise is important, as argued by O'Malley et al. [27], since the test subjects' knowledge of specific work tasks is quite often corresponding to their level of expertise. Nielsen [26] recommends that the test subjects have the same level of experience, whereas having one of the test subjects enabled to guide the interaction, is an argument used by Kahler [16] when stating advantages by pairing test subjects with different levels of

experience. Usually children do not posses expertise of work that might influence the outcome of the usability test, which makes the issue of expertise subtler when working with children. Most studies involving children do not explicitly consider the level of expertise [19, 25], one of the exceptions is a study where the participating children are profiled according to their scripting level [28]. Where age does not seem to matter when testing with adults, it has a more eloquent impact when conducting tests with children, since the children's level of maturity changes more quickly than adults. Most studies equalize the children's age, with their level of expertise. It is not obvious how children's ages influence results of a usability test.

Secondly, level of acquaintance is another important aspect in constructive interaction. Previous studies have indicated that children behave quite differently according to how well they know each other. In a study where adult test subjects were asked to bring a friend, co-worker, or family member to the usability test provided a positive experience [16] while other studies stress the importance of using non-acquainted test subjects [17]. Most studies involving children seems to prefer acquainted pairs of children; this is often achieved through involvement of children attending same school classes or kindergartens [10, 25, 28]. In the Eco-I project [30], the pairing goes beyond acquaintance, since a participating teacher had configured the pairs of children according to how well they worked together. Few studies indicate that the pairs of children were unacquainted, but this might have been the case in the StoryMat project [5] since the children attended different schools.

Thirdly, gender is potentially important when working with children; for example illustrated by girls and boys preferring different types of computer games [12]. Gender can also play a subjective role with children's preferences and attitudes towards technologies [4, 14]. But it is not apparent if and how gender influences other other issues of usability testing, such as effectiveness, efficiency, or number of identified usability problems. Several studies involve both genders in the design processes [3, 6, 19, 20, 30, 32, 33]. Some studies adapted imbalanced numbers of girls and boys [2, 25], while others deliberately chose an equal number of boys and girls [19]. Furthermore, some studies intentionally use same-sex pairs [10, 24].

Analyzing previous research on interaction design and children, we found several studies in which children participated as test subjects applying think-aloud [7, 8, 9, 28, 33], constructive interaction [24, 25, 30, 31], or both approaches [2, 6]. However, none of these studies present results related to how well the children adapted to think-aloud or constructive interaction. Summarized, we need a deeper understanding of involving children in the evaluation of software products to assess some of the opportunities and limitations related the different evaluation methods.

3 Experimental Method

The purpose of our experiment was to explore the impact of involving children in the evaluation of a software product. The idea was to place children in different settings or conditions to see how this affects their performance. Thus, in this paper we do not measure the performance of the different setups in terms of usability problem identification (please refer to [1] for this aspect of our study).

Table 1. 60 children participated in our experiment in three different setups: constructive interaction as acquainted dyads or non-acquainted dyads and think-aloud as individual testers

	Constructive Interaction		Think-Aloud
	Acquainted Dyads (N=24)	**Non-Acquainted Dyads (N=24)**	**Individual Testers (N=12)**
Girls	6x2	6x2	6
Boys	6x2	6x2	6
Total	12x2	12x2	12

We designed the experiment as a 3x2 matrix consisting of three types of sessions: individual testers using think-aloud, acquainted dyads (pairs) using constructive interaction, and non-acquainted dyads using constructive interaction. Furthermore, we configured the usability test sessions with same-sex dyads having sessions with girls and boys for each of the three setups. This is illustrated in table 1.

3.1 Participants

60 children (30 girls and 30 boys) at the age of 13 and 14 years old (M=13.35, SD=0.48) participated as test subjects in the experiment. The children were all 7th grade pupils from five different elementary schools in the greater Aalborg area. The children did not receive compensation for their involvement in the experiment.

The children were assigned as test subjects to one of the three test setups e.g. individual testers, acquainted dyads, or non-acquainted dyads. Each setup had twelve individual testers (six girls and six boys), twelve acquainted dyads (six pairs of girls and six pairs of boys), and twelve non-acquainted dyads (six pairs of girls and six pairs of boys). Assignment of the children to the three test setups was done randomly under two conditions 1) all acquainted dyads attended the same school class and 2) all non-acquainted dyads attended different schools. The acquainted pairs had known each other for at least five years except for one pair of girls and one pair of boys who had been acquainted for one year (M=6.25, SD=2.5). None of the non-acquainted dyads knew each other in advance.

3.2 System

The selected system for our experiment was an inno-100 mobile phone by innostream. This particular mobile phone was selected since it had not been released on the European market at the time of our experiment. Thus, all children would have to learn to use the mobile phone.

The inno-100 integrates a range of standard mobile phone features, such as making and receiving phone calls and short text messages, and more advanced features, including speed dial functions and options for creating personalized ring tones. The inno-100 has two separate screens with a main 128x144 pixel 16 bit colour screen and 64x80 pixel sub screen on the cover. The navigation is primarily based on icons in the two upper menu levels. The lower levels are textual based including choice menus for setting values. Furthermore, the inno-100 integrates a number of games.

3.3 Procedure

Children from five schools in Aalborg, Denmark were introduced to the experiment by two of the participating researchers. The researchers explained the children's roles in the experiment and how their participation would contribute to our research. Participation in the experiment was voluntarily and interested children got an information sheet describing the experiment in detail and a consent form that had to be signed by a parent or a guardian. After receiving signed consent forms from a total of 60 children, we scheduled the usability evaluation sessions.

The sessions were held at the usability laboratory at Aalborg University. We adapted the guidelines for usability testing with children proposed by Hanna et al. [13]. Particularly, we focused on greeting the children, stressing the importance of the participation, and stressing that they were not the object of the test. The purpose of the evaluation was explained in detail to the children and they were shown the facilities of the usability lab. Test subjects intended for roles as non-acquainted dyads were kept separate before the test sessions. The children received questionnaires on which they had to provide answers to such as age, name, school, and mobile phone experience. The usability test sessions were conducted in a specialized usability laboratory. The laboratory integrated two rooms; an observation room in which the evaluations took place and a control room where one of the researchers would handle electronic equipment for recording the sessions. The two rooms were separated with a one-way mirror allowing people in the control room to see what was going on in the observation room. All sessions were recorded on video tapes for later analyses including perspectives of the children and of their interactions with the mobile phone.

The children were asked to solve twelve tasks one at a time addressing standard and advanced functionalities in the inno-100 mobile phone. This included making a phone call, sending a short text message, adjusting the volume of ring tones, and editing entries in the address book. We did not specify any time limits for the tasks, but required the participants to try to solve all tasks. All children were able to solve all specified tasks. On average, the children spent 26:45 minutes (*SD*=06:39) on the twelve tasks. The individual testers were asked to think-aloud while solving the tasks. We explained think-aloud to the individual testers in terms of the descriptions in [26, p. 195-198]. The acquainted and non-acquainted dyads were asked to solve the tasks by constructive interaction where they should collaborate with each other in order to solve the tasks. We explained constructive interaction to the dyads in terms of the descriptions in [26, p. 198].

After the usability sessions, the children completed a subjective workload test (NASA-TLX) [22]. The children filled in the test individually even though the participated in pairs. This was done to evaluate the workload as experienced by the children in order to compare the different setups. We translated the test into the children's native language, Danish.

3.4 Data Analysis

After conducting all 36 sessions, the sessions were analyzed in a collaborative effort between two of the authors of this paper. The sessions were picked randomly for the analysis to avoid bias in the analysis. We analyzed the sessions according to how well the children collaborated (in constructive interaction sessions) and recorded their

verbal interaction and comments. The six different aspects of our analysis were: 1) Level of verbalization, 2) quality of verbalization, 3) interaction between test subject(s) and test monitor, and 4) influence of test monitor on the solving of tasks. The two setups of constructive interaction were additionally analyzed according to: 5) Level of collaboration between the dyads and 6) quality of the collaboration between the dyads. We analyzed and marked each of the six aspects on a scale from 1 to 5 where 1 being the lowest score and 5 being the highest score. For example, for the level of verbalization, a session was marked 1 if the children made none or very few verbalizations during their interaction with the system, and a sessions was marked 5 if the children constantly or almost all time made verbalization during interaction.

The NASA-TLX tests were further analyzed. 55 tests were answered correctly by the children while 5 were incomplete answered. Data from our assessment of think-aloud and collaboration and the NASA-TLX tests were analyzed with one-way ANOVAs, followed by post hoc comparisons using Tukey tests.

4 Results

The 60 children in the 36 usability test sessions solved all 12 assigned tasks. Even though the constructive interaction sessions with acquainted dyads (M=29:54, SD=06:57) spent most time on the assignments in our experiment; the individual testers (M=25:34, SD=03:44), and the non-acquainted dyads (M=24:48, SD=07:53), we found no significant differences for the task completion times. The children performed and behaved differently in the three setups and the following sections present our assessment of their interaction and collaboration and the NASA-TLX test.

4.1 Assessment of Think-Aloud and Collaboration

As a part of our assessment of the three setups, we applied six different aspects of the verbalization and collaboration in usability tests. These six aspects are illustrated in table 2. Not surprisingly, we found that the level of verbalization was considerably higher for the constructive interaction sessions compared to the think-aloud sessions. The acquainted dyads scored rather high (M=4.58, SD=0.90) especially compared the individual testers who scored rather low (M=2.17, SD=1.19). An analysis of variance shows significant differences between the three setups on level of verbalization $F_{(2,33)}$=13.421, p=0.001. A post-hoc test showed significant difference at the 0.1% level between the acquainted dyads and the individual testers and at the 5% level between the non-acquainted dyads and the individual testers. Furthermore, we found a tendency towards a higher level of verbalization for the acquainted dyads compared the non-acquainted dyads, but this difference is not significant (p=0.090).

Looking further at verbalization in the test sessions, we analyzed the quality of the verbalization primarily defined as the ability of the verbal comments to facilitate the identification and classification of usability problems. Considering the quality of the verbalization the differences between the setups are less apparent than for the level of verbalization. For the acquainted dyads (but also for some non-acquainted dyads), several verbal comments did not concern the actual test; a lot of the verbal comments did not facilitate the identification of usability problems. Summarized, the differences between the setups on quality of verbalization were not significant $F_{(2,33)}$=2.171, p=0.130.

Table 2. Assessment of verbalization and collaboration in the three setups. A plus indicates a significant difference to the setup marked with a minus according to an ANOVA test.

	Constructive Interaction		Think-Aloud
	Acquainted Dyads (N=12)	**Non-Acquainted Dyads (N=12)**	**Individual Testers (N=12)**
Level of verbalization	4.58 (0.90) +	3.58 (1.31) +	2.17 (1.19) -
Quality of verbalization	3.58 (1.00)	3.25 (1.48)	2.50 (1.38)
Interaction between test subject(s) and monitor	2.75 (0.87)	3.08 (0.79)	3.25 (0.87)
Influence of test monitor on the solving of tasks	2.17 (0.39)	1.67 (0.65)	1.83 (0.58)
Level of the collaboration between the dyads	4.75 (0.62)	3.83 (1.47)	N/A
Quality of the collaboration between the dyads	3.67 (1.56)	3.58 (1.56)	N/A

We further analyzed the influenced of and interaction with the test monitor. Constructive interaction provides potentially natural thinking-aloud as test subjects collaborate in pairs to solve tasks and therefore, one could expect less influence and interaction with a test monitor. We found that the test monitor has slightly more interaction with the think-aloud subjects compared the constructive interaction subjects, but the difference is not significant $F_{(2,33)}=0.134$, $p=0.875$. On the other hand, we identified a higher influence form the test monitor on the solving of tasks for the acquainted dyads compared both non-acquainted dyads and individual testers, but again this difference is not significant $F_{(2,33)}=0.282$, $p=0.756$.

As constructive interaction have test subjects collaborate in pairs to solve tasks, we finally assessed the level and quality of collaboration. Most of the acquainted dyads collaborated during the entire sessions (M=4.75, SD=0.62) and we identified a tendency towards a higher collaboration between them than the non-acquainted dyads (M=3.83, SD=1.47), but this difference is not significant according to a Student's t-test $t_{(22)}=1.993$, $p=0.059$. Considering the quality of the collaboration, we found no difference between the two setups $t_{(22)}=0.131$, $p=0.897$.

4.2 Assessment of Subjective Workload

Table 3 summarizes mean values for the six factors of the NASA-TLX test as assessed by the 60 children in the three setups. As the table illustrates, minor differences could be observed between the different setups, however we found no significant differences between them. Even though not significant, we can however see that the individual testers found the effort factor more important than the dyads, but large variances were identified for the individual testers on this factor.

On the other hand, more factors were assessed to almost the same mean values for the three setups e.g. frustration and mental demand. While the absolute values of the

Table 3. Subjective workload (NASA-TLX test) for think-aloud and constructive interaction illustrating the mean values for the six factors as assessed by children

	Constructive Interaction		Think-Aloud
	Acquainted Dyads (N=20)	Non-Acquainted Dyads (N=24)	Individual Testers (N=11)
Effort	38.5 (19.7)	41.9 (20.3)	52.7 (23.8)
Frustration	34.3 (25.4)	35.8 (22.4)	39.5 (23.4)
Mental	43.5 (16.2)	42.1 (19.3)	50.0 (12.2)
Performance	27.0 (21.7)	25.8 (17.7)	35.0 (24.5)
Physical	41.0 (25.8)	39.4 (25.9)	27.3 (13.8)
Temporal	38.5 (20.1)	27.5 (18.9)	37.7 (25.7)

factors provided no significant differences between the three setups, we analyzed the inter-relative importance of the factors.

The assessment of the relative importance of the factors (table 4) showed significant difference between the three setups on the effort factor $F_{(2,52)}$=5.693, p=0.006. A post-hoc comparison showed significant difference at the 1% level between the acquainted dyads and non-acquainted dyads and at the 5% level between the acquainted dyads and the individual testers. Additionally, sitting with an acquainted influenced the importance of performance as acquainted dyads found this significantly more important than the individual testers and the non-acquainted dyads $F_{(2,52)}$=3.775, p=0.029. A post-hoc test showed significant difference at the 5% level between the acquainted and non-acquainted dyads.

Table 4. Inter-relative assessment of workload factors for the three setups. A plus indicates a significant difference to the setup marked with a minus according to an ANOVA test.

	Constructive Interaction		Think-Aloud
	Acquainted Dyads (N=20)	Non-Acquainted Dyads (N=24)	Individual Testers (N=11)
Effort	2.30 (1.17) -	3.38 (1.10) +	3.45 (1.29) +
Frustration	1.75 (1.25)	2.54 (1.59)	2.64 (0.92)
Mental	2.90 (1.48)	3.38 (1.28)	3.73 (1.49)
Performance	3.15 (1.60) +	2.08 (1.18) -	2.09 (1.38)
Physical	2.35 (1.66)	2.08 (1.79)	1.36 (1.75)
Temporal	2.55 (1.50)	1.54 (1.47)	1.73 (1.27)

We found that the acquainted dyads assessed frustration as the least important factor while both individual testers and non-acquainted dyads rated it as the third most important factor, but this difference was not significant $F_{(2,52)}$=2.337, p=0.107. For the

remaining three factors, we found only minor differences between the three setups and no significant differences, mental demand $F_{(2,52)}$=1.357, p=0.266, physical demand $F_{(2,52)}$=1.160, p=0.322, while we identified a tendency for temporal issues $F_{(2,52)}$=2.800, p=0.070.

Table 5. Calculated workload for the three setups. A plus indicates a significant difference to the setup marked with a minus according to an ANOVA test.

	Constructive Interaction		Think-Aloud
	Acquainted Dyads (N=20)	Non-Acquainted Dyads (N=24)	Individual Testers (N=11)
Effort	99.8 (80.0) -	148.7 (90.7)	190.9 (126.3) +
Frustration	61.0 (63.1)	108.8 (97.5)	116.8 (91.5)
Mental	120.8 (65.1)	132.5 (69.5)	186.8 (90.7)
Performance	83.8 (88.4)	51.7 (56.0)	65.0 (50.0)
Physical	118.8 (113.8)	80.0 (104.4)	40.5 (61. 6)
Temporal	90.0 (60.7) +	42.1 (55.6) -	58.2 (59.3)

Combining the two measures, we calculated the overall score for the workload for the participating children. As discovered above, we found that the individual testers had to put much more effort into the testing situation and an ANOVA test showed a significant difference between the three setups $F_{(2,52)}$=3.464, p=0.039. A post-hoc comparison showed significant difference at the 5% level between the individual testers and the acquainted dyads. On the other hand, the acquainted dyads in total assessed temporal demand rather high compared to the two other setups and we found a significant difference between the three setups $F_{(2,52)}$=3.737, p=0.030. A post-hoc test showed significant difference at the 5% level between the acquainted dyads and the non-acquainted dyads.

We identified a tendency for mental demand as the individual testers in general assessed this factor higher than both constructive interaction setups, however the difference was not significant for our test $F_{(2,52)}$=3.114, p=0.057. Again and as above, we found that the level of frustration is much lower for the acquainted dyads compared the two other setups, however the difference is not significant $F_{(2,52)}$=2.247, p=0.116. Furthermore, we found no significant differences for the other calculated values; physical demand $F_{(2,52)}$=2.198, p=0.121 and performance $F_{(2,52)}$=1.190, p=0.312.

5 Discussion

This section provides qualitative results from the study. We have identified a number interesting lessons related usability testing with children.

Lesson 1: *Constructive interaction did not necessarily facilitate natural think-aloud as the dyads tended to talk-aloud and not think-aloud.* Constructive interaction in usability testing with children potentially provides natural thinking-aloud as the

children collaborate in pairs to solve tasks. Our study illustrated that children in pairs using constructive interaction had a much higher level of verbalization, but often they were more talking-aloud than actually thinking-aloud. We further experienced that the individual testers applying think-aloud tended to be quieter during the sessions compared to the dyads; they expressed themselves noticeably fewer times than the dyads. When asked about their choices, more of them would mostly answer our questions in very few words without giving further insight into their behaviour and choices. On the other hand, the non-acquainted dyads had less interaction with each other compared to the acquainted dyads; they mainly kept focus on the task they were solving. The interaction of the acquainted dyads was partially related to the task, but we identified some interaction as noise as this was irrelevant to the solving of the task, for example some would have long discussions on what to name the melody they had just composed. These observations resemble the discussion by Ericsson and Simon of think-aloud and talk-aloud [11]. It was very difficult to get the children to explain their interaction and motivation even though they had been carefully instructed before the session. Thus, this can be seen as a contradiction to benefits of constructive interaction as stated by Nielsen [26] as we found only minor differences between the think-aloud sessions and constructive interaction sessions.

Lesson 2: *Dyad configuration in constructive interaction influenced the children's behaviour and assessment of the testing situation according to their acquaintance*. Our study indicated that there were a significant difference between how the acquainted and the non-acquainted dyads experienced the assessment of effort and performance. Our results showed that the acquainted dyads were significantly more satisfied with their own performance and they did not feel it demanded a lot of effort from them. It was just the opposite for the non-acquainted dyads. Even though the acquainted dyads sometimes would try to pull the phone out of the hands of their co-solver, they rated performance of minor importance compared to the non-acquainted dyads. From our study, we also found that the non-acquainted dyads acted rather polite against each other and in general they were more polite to each other than the acquainted dyads. Consequently, they collaborated quite differently compared to the acquainted dyads and they did not argue explicitly for the control of the tested phone. This is also indicated in our results as we found a tendency, however not significant, towards better collaboration between the non-acquainted dyads. Further, the non-acquainted dyads separated the roles between them during the test. Even in the cases were they did not collaborate very well, they would some times read the task aloud, or they would take turns by shifting in between tasks. The acquainted dyads' interaction were influenced by the fact that the children knew each other in advance, they referred to each others by nick-names, remarked their co-solvers intelligence etc. They would also physically try to grab the phone and thereby preventing their co-solver from helping to solve the task. The acquainted dyads would easily get distracted from the task they were solving, they would discover something interesting in the menu, and would spend time discovering such aspects. Some of the non-acquainted dyads did not collaborate very well while solving the task; we found no significant differences between the girls and the boys in this issue. The children took turns in operating the system and the child who was not in control of the interaction had sometimes difficulties in seeing what was going on the screen of the phone.

Lesson 3: *Gender issues might play important roles in the configuration of dyads in constructive interaction.* Our study utilized pairs of same sex dyads as adapted in several studies with children [10, 24]. Even though we haven't summarized the results gender wise, our study showed a tendency towards that the boys collaborated better than the girls. Especially the acquainted dyads of boys collaborated rather well and had a fruitful and successful collaboration whereas the acquainted dyads of girls experienced several situations where their collaboration was rather poor. Thus, while it seems to be of less importance if the boys tested in acquainted or non-acquainted dyads, the girls should test in non-acquainted dyads. For some of the specified tasks, we observed that the acquainted dyads of girls would more easily get distracted from the task they were solving, they would discover something interesting in the menu, and would spend time discovering what it was, for example acquainted dyads quite often used several minutes to compose a melody, for example "Itsy Bitsy Spider".

6 Conclusion

In this paper, we investigate and address the above stated questions by looking at how children perform and behave in constructive interaction during usability testing. Our particular focus is on how the children behave and perceive a testing situation when involved a traditional think-aloud test compared to constructive interaction tests. Thus, we did not treat the performance of the different setups in terms of usability problem identification (please refer to [1] for this aspect of our study).

Our results show that the pairing of children had impact on how the children verbalized and collaborated in pairs during the testing sessions. First, we found that constructive interaction did not necessarily facilitate natural think-aloud as the dyads tended to talk-aloud and not think-aloud. Our children in pairs had a high level of verbalization, but often they were more talking-aloud than actually thinking-aloud. This issue resembles some of the discussions by Ericsson and Simon of think-aloud and talk-aloud [11]. Secondly, dyad configuration in constructive interaction influenced the children's behaviour and assessment of the testing situation according to their acquaintance. The acquainted dyads were significantly more satisfied with their own performance and they did not feel it demanded a lot of effort from them. It was just the opposite for the non-acquainted dyads. Thirdly, gender issues might play important roles in the configuration of dyads in constructive interaction. Our study showed a tendency towards that the boys collaborated better than the girls. Especially the acquainted dyads of boys collaborated rather well and had a fruitful and successful collaboration whereas the acquainted dyads of girls experienced several situations where their collaboration was rather poor. Thus, while it seems to be of less importance if the boys tested in acquainted or non-acquainted dyads, the girls should test in non-acquainted dyads.

Our study suffers from a number of limitations which could form further research with children. First, our results of our experiment cannot simply be generalized for all ages of children. Thus, replicating the experiment with younger children may show a different kind of relationship between think-aloud and constructive interaction. Secondly, we recorded that the non-acquainted dyads continuously took turns with the

mobile phone making it difficult for the other child to see what was going on at the interface. This could probably be different for desktop-based applications.

Acknowledgements

The work behind this paper received financial support from the Danish Research Agency (grant no. 2106-04-0022). We would especially like to thank all the participating children and their parents. Furthermore, we want to thank several anonymous reviewers for comments on drafts of this paper.

References

1. Als, B. S., Jensen, J. J., and Skov, M. B. (2005) Comparison of Think-Aloud and Constructive Interaction in Usability Testing with Children. In *Proceedings of the 4th International Conference on Interaction Design and Children (IDC'05)*, ACM Press
2. Benford, S., Bederson, B. B., Åkesson, K-P, Bayon, V., Druin, A., Hansson, P., Hourcade, J. P., Ingram, R., Neale, H., O'Malley, C., Simsarian, K. T., Stanton, D., Sundblad, Y., and Taxén, G. (2000) Designing storytelling technologies to encouraging collaboration between young children. In *Proceedings of the Human Factors and Computing Systems CHI'00*, ACM Press, pp. 556 - 563
3. Bers, M. U., Gonzalez-Heydrich, J., and DeMaso, D. R. (2001) Identity Construction Environments: Supporting a Virtual Therapeutic Community of Pediatric Patients Undergoing Dialysis. In *Proceedings of the Human Factors and Computing Systems CHI'01*, ACM Press, pp. 380 - 387
4. Cassell, J. (2002) *Genderizing*. The Handbook of Human-Computer Interaction
5. Cassell, J. and Ryokai, K. (2001) Making Space for Voice: Technologies for Supporting Children's Fantasy and Storytelling. *Personal and Ubiquitous Computing*, Springer-Verlag, vol. 5(3), pp. 169 - 190
6. Danesh, A., Inkpen, K. M., Lau, F., Shu, K., Booth, K. S. (2001) Geney: Designing a collaborative activity for the Palm handheld computer. In Proceedings of the Human Factors and Computing Systems CHI'01, ACM Press, pp. 388 - 395
7. Druin, A. and Solomon, C. (1996) Designing Multimedia Environments for Children. Wiley & Sons, New York
8. Druin, A. (1999) The Role of Children in the Design of New Technology. HCIL Technical Report No. 99-23, University of Maryland, USA
9. Druin, A. (1999) The Design of Children's Technology. Morgan Kaufmann Publishers, Inc., San Francisco, CA
10. Ellis, J. B. and Bruckman, A. S. (2001) Designing Palaver Tree Online: Supporting Social Roles in a Community of Oral History. In Proceedings of the Human Factors and Computing Systems CHI'01, ACM Press, pp. 474 - 481
11. Ericsson, K.A. and Simon, H.A. (1990) *Protocol Analysis. Verbal reports as data*, Cambridge Massachusetts
12. Gorriz, C. M. and Medina, C. (2000) Engaging Girls with Computers through Software Games. Communications of the ACM, vol. 43, No. 1, pp. 42 – 49
13. Hanna, L., Risden, K., and Alexander, K. J. (1997) Guidelines for Usability Testing with Children. In interactions, September + October, pp. 9 – 14

14. Inkpen, K. (1997) Three Important Research Agendas for Educational Multimedia: Learning, Children, and Gender. In Proceedings of Educational MultiMedia '97
15. Jensen, J. J. and Skov, M. B. (2005) A Review of Research Methods in Children's Technology Design. In Proceedings of the 4th International Conference on Interaction Design and Children (IDC'05), ACM Press
16. Kahler, H. (2000) Constructive Interaction and Collaborative Work. interactions, May + June, pp. 27 - 34
17. Karat, C.-M., Campbell, R., and Fiegel, T. (1992) Comparison of Empirical Testing and Walkthrough Methods in User Interface Evaluation. In Proceedings of the Human Factors and Computing Systems CHI'92, ACM Press, pp. 397-404
18. van Kesteren, I. E. H., Bekker, M. M., Vermeeren, A. P. O. S., and Lloyd, P. A. (2003) Assessing usability evaluation methods on their effectiveness to elicit verbal comments from children subjects. In Proceeding of the 2003 conference on Interaction design and children (IDC'03), ACM Press, pp. 41 - 49
19. Lester, J. C., Converse, S. A., Kahler, S. E., Barlow, S. T., Stone, B. A., and Bhogal, R. S. (1997) The Persona Effect: Affective Impact of Animated Pedagogical Agents. In Proceedings of the Human Factors and Computing Systems CHI'97, ACM Press, pp. 359 - 366
20. Lumbreras, M. and Sánchez, J. (1999) Interactive 3D Sound Hyperstories for Blind Children. In Proceedings of the Human Factors and Computing Systems CHI'99, ACM Press, pp. 318 - 325
21. Markopoulos, P. and Bekker, M. (2003) On the Assessment pf Usability Testing Methods for Children. Interacting with Computers, Elsevier, Vol. 15, pp. 227 – 243
22. Miller R. C. and Hart, S. G. (1984) Assessing the Subjective Workload of Directional Orientation Tasks. In Proceedings of 20th Annual Conference on Manual Control, NASA Conference Publication, pp. 85 – 95
23. Miyake, N. (1986) Constructive Interaction and the Iterative Process of Understanding. Cognitive Science, vol. 10(2), pp. 151 - 177
24. Moher, T., Johnson, A., Ohlsson, S., and Gillingham, M. (1999) Bridging Strategies for VR-based Learning. In Proceedings of the Human Factors and Computing Systems CHI'99, ACM, pp. 536 - 543
25. Montemayor, J., Druin, A., Farber, A., Simms, S., Churaman, W., and D'Amour, A. (2002) Physical Programming: Designing Tools for Children to Create Physical Interactive Environments. In Proceedings of the Human Factors and Computing Systems CHI'02, ACM Press, pp. 299 - 306
26. Nielsen, J. (1993) Usability Engineering. Academic Press
27. O'Malley, C. E., Draper, S. W., and Riley, M. S. (1984) Constructive Interaction: A Method for Studying Human-Computer-Human Interaction. In Proceedings of IFIP Interact '84, pp. 269 – 274
28. Rader, C., Brand, C., and Lewis, C. (1997) Degrees of Comprehension: Children's Understanding of a Visual Programming Environment. In Proceedings of the Human Factors and Computing Systems CHI'97, ACM Press, pp. 351 - 358
29. Resnick, M., Martin, F., Berg, R., Borovoy, R., Colella, V., Kramer, K., and Silverman, B. (1998) Digital Manipulatives: New Toys to Think With. In Proceedings of the Human Factors and Computing Systems CHI'98, ACM, pp. 281 – 287
30. Scaife, M., Rogers, Y., Aldrich, F., and Davies, M. (1997) Designing for or Designing with? Informant Design for Interactive Learning Environments. In Proceedings of the Human Factors and Computing Systems CHI'97, ACM Press, pp. 343 - 350

31. Skov, M. B., Andersen, B. L., Duhn, K., Garnæs, K. N., Grünberger, O., Kold, U., Mortensen, A. B., and Sørensen, J. A. L. (2004) Designing a Drawing Tool for Children: Supporting Social Interaction and Communication. In Proceedings of the Australian Computer-Human Interaction Conference 2004 (OzCHI'04)
32. Stewart, J., Bederson, B. B., and Druin, A. (1999) Single Display Groupware: A Model for Co-Present Collaboration. In Proceedings of the Human Factors and Computing Systems CHI'99, ACM, pp. 286 - 293
33. Strommen, E. (1998) When the Interface is a Talking Dinosaur: Learning across Media with ActiMates Barney. In Proceedings of the Human Factors and Computing Systems CHI'98, ACM Press, pp. 288 - 295