Usability is a key issue in the success of e-learning resources (Lahti, Siira, & Törmänen, 2012), so their design process should involve the users as early as possible. Gathering feedback from children, however, is challenging, as their language skills, confidence, and self-belief (Read & MacFarlane, 2006) and the presence of adult observers (Höysniemi, Hämäläinen, & Turkki, 2003) may affect their performance. Children are also eager to please (Hanna, Risden, & Alexander, 1997; Höysniemi et al., 2003) and tend to give overly positive feedback (Höysniemi et al., 2003; MacFarlane, Sim, & Horton, 2005).

In this case study, we evaluated a functional prototype of an artificial intelligence–based system called the Poetry Machine with 9- to 10-year-old children. The system is designed to facilitate poetry teaching and writing in Finnish primary schools. It aims to solve the problem of an empty paper by providing pupils with a computer-generated poem excerpt to start with, and rhyme suggestions and new lines as needed. The pupil may modify the excerpt through a user interface similar to poetry fridge magnets by dragging and dropping words, removing or editing them, or adding new words or lines (Figure 1).

We selected three evaluation criteria: the number of usability problems, usefulness, and fun. The applicability of the methods is discussed along with the resources required. Peer tutoring revealed numerous problems, whereas group testing and the Feedback Game were quick to conduct and analyze.

**EVALUATION METHODS**

Children in our target age group (9–10 years old) are accustomed to working with adults in a school environment and are able to describe what they see and do (Hanna et al., 1997). Even so, social support and given instructions affect their performance (Hourcade, 2008). For example, children are able to concentrate for a longer time if they are enjoying themselves (Markopoulos & Bekker, 2003). To create an enjoyable and relaxed testing environment and to promote natural interaction between the children, we selected methods involving multiple participants: Peer tutoring and paired interviews involve pairs of users, and group testing and the Feedback Game involve small groups. To provide the children a familiar test environment, we conducted the tests at their school, so we were also able to interview their teacher.

In peer tutoring, a person instructs another person with a similar status. When used as an evaluation method with children, the first child learns to use the evaluated system and then becomes a tutor instructing
the other child, the tutee (Edwards & Benedyk, 2007). As children easily emphasize positive features when instructing their friends, supplementary methods are recommended to elaborate negative comments and dislikes (Höysniemi et al., 2003). We used paired interviews, giving the benefit of peer support also in the interview. The questions were open-ended to reduce the children's tendency to please (Read & MacFarlane, 2006).

We selected group testing as another evaluation method to assess the Poetry Machine in a typical teaching context with a small group of children. Interviewing more than two children simultaneously is challenging, so we studied contemporary alternatives. We did not find a suitable method but found some potential ideas, such as the Fun Toolkit (MacFarlane et al., 2005; Read & MacFarlane, 2006; Read, MacFarlane, & Casey, 2002; Sim et al., 2006), which uses smileys for assessing single products and tactile scales for ranking several products with a single user. We combined these ideas with elements from focus groups, interviews, and board games, producing the Feedback Game. Analogous to board games, it has clear rules facilitating interaction, a physical game board, and game tokens. The Fun Toolkit inspired the game board: a table with eight questions and a 5-point Likert-type scale using smileys from a sad face to a laughing face (Figure 2). This test was the first experimentation with the game and an attempt to validate the method.

**TEST PROCEDURE**

The evaluations were conducted at a school that had participated in several pedagogical research programs, so the pupils were used to observers in their classes. The school environment may, however, discourage creative thinking (Obrist et al., 2011), so we used paired and group-based methods to facilitate discussions and creativity. We also recruited all the participants from the same class and asked the class teacher to select suitable pairs and groups who were comfortable in working together.

Altogether, 22 pupils participated in our tests: six pairs in peer tutoring and paired interviews and two groups of five pupils in group testing and the Feedback Game. Both settings included an equal number of girls and boys, and they were all 9 or 10 years old. We requested parental consent from all the participating pupils with a form sent through the school. We

Figure 1. The evaluated version of the Poetry Machine user interface with the fridge magnet metaphor.
asked the class teacher to form the peer-tutoring pairs from friends of the same sex to ease discussion and to avoid hostility (Höysniemi et al., 2003), whereas close friends were avoided in the group testing, following the instructions for focus groups (Patel & Paulsen, 2002).

We conducted six peer-tutoring and paired interview sessions in 2 days and two group-testing and Feedback Game sessions a week after. We organized the sessions during normal school hours in a small classroom and recorded them with one camera. The children were rewarded with a card, a sticker sheet, and candy.

**Peer tutoring and paired interviews.** Each peer-tutoring session involved five persons: the tutor and the tutee, one evaluator acting as a facilitator, and two observers. First, we introduced ourselves, presented the test setting, and asked the tutor to fill in a background questionnaire. Then we trained the tutor to write a poem with the prototype, after which the tutor moved aside to read a book, while we introduced the test setting and background questionnaire to the tutee. After that, the actual peer tutoring started, and the tutor instructed the tutee to write a poem with the system. The peer-tutoring sessions lasted on average 33 min, plus 7 min for the paired interviews.

We let the children work independently to promote their natural interactions, so the facilitator helped only when needed. Some pupils also needed encouragement and praise, as Höysniemi et al. (2003) had indicated. Following their instructions, we seated each pair so that the tutee controlled the mouse.

After the peer tutoring, the facilitator interviewed the pair at the same time. We started with warm-up questions asking how many stars the children would give to the Poetry Machine out of five and what was the best or worst thing about it. Then we moved on to specific problems faced during the test. Next, we asked if the children would recommend the tool to their friends and reasons for their choice. Finally, we asked if the children would like to participate in a similar test again.

**Group testing and Feedback Game.** Each group testing session involved nine persons: five pupils, their teacher, one facilitator, and two observers. First, we introduced ourselves, and the pupils filled in the background questionnaire. Then the teacher showed how to use the system and instructed the children as they wrote their own poems. The group-testing sessions lasted about 45 min, plus 15 min for the Feedback Game.

We played the Feedback Game after group testing to elicit more comments and improvement ideas. First, the players selected the color of their game tokens and thereby also their nickname during the game. The facilitator then explained the rules of the game and started being a game master. The game master started each round by asking a question on the game board, such as “Were the beginnings offered by the Poetry Machine good?” All the players answered at the same time by
placing their token on the most suitable smiley column. Next, the game master asked further questions, such as “What was good/poor about them?” Each player answered on his or her own turn, and everyone got to start a round in turn. As Figure 2 illustrates, children were allowed to move their tokens during the game. The used tokens were left on the board to visualize the progress of the game. An ongoing game session is depicted in Figure 3.

Analysis. Our evaluation criteria included fun, usefulness, and the number of usability problems. We were also interested in the number of user comments gathered with each method. Therefore, the analyses required several walk-throughs of the collected materials. The usability problems were classified in three levels:

3 = Problem prevents the use of a feature.
2 = Problem complicates the use of a feature.
1 = Problem slows down performance or bothers the user.

The peer-tutoring analysis focused on usability problems, but we also analyzed the behavioral signs, feature usage, and the pupils’ comments. To reduce evaluator effect in the problem detection, we had a pool of three evaluators, from which two evaluators independently analyzed each test recording, as recommended by Hertzum and Jacobsen (2001). One evaluator analyzed the other factors, classifying the behavioral signs as positive or negative based on literature (see, e.g., Hanna et al., 1997; Read et al., 2002). A neutral category was also needed for conflicting signs, such as frowning, which can indicate problems (Hanna et al., 1997) or positive concentration (Read et al., 2002) in the Nordic expressive behavior. This evaluator also counted the feature usage to see what intrigues the children and to assess the test coverage.

The group-testing analysis focused on usability problems and comments, but these recordings did not support the analysis of behavioral signs or feature usage. One evaluator analyzed the recordings supported by in situ notes from all the observers. She also transcribed the paired interviews and the Feedback Game to detect development ideas and comments on the concept and the evaluation criteria.

TEST RESULTS AND METHOD EXPERIENCES

In the test results, we emphasize our evaluation criteria: number of usability problems found, usefulness of the system, and fun. We also consider how useful the methods were in evaluating these criteria. Because usability evaluation methods with children are often assessed by their ability to detect usability problems and the resources they require (Markopoulou & Bekker, 2003), as well as their ability to elicit authentic comments from the children (van Kesteren, Bekker, Vermeeren, & Lloyd, 2003), we also investigate these factors.

Overall, we found 89 usability problems. Peer tutoring revealed most of the problems (76), giving also detailed information about their reasons. The most frequent problems were related to moving and targeting the words in the poem. The most frequently used features during peer tutoring were editing words (102 times), throwing words into trash bin (88), and moving words (81). The paired interviews revealed 16 problems, of which six were new. Group testing revealed only 18 problems, mainly because our focus had already switched to getting feedback on the system. Only seven problems were new, but they ranked high on the final problem listing by their severity. These problems were related to adding and moving whole lines in the poem. The Feedback Game, as the last debriefing method, revealed only three problems, without new problems.
The most severe problem in both effect and frequency was
the limited target area for dropping the words into their new
places. This problem was found with all methods, but a similar
problem with moving whole rows was found only through
group testing. As the group testing included a short introd-
uction to using the tool, it also revealed problems in memorabil-
ity. For example, the children had problems in recalling how to
add new words. The paired interviews, in turn, gave valuable
insight into missing features: One pair wanted to remove the
whole poem, and another wanted to write their own words
into the rhyming tool directly. The Feedback Game did not
give new information about the usability problems but
emphasized the problems in moving the words in general.

Both the pupils and the teacher found the system useful.
The peer-tutoring sessions gave little information about
usefulness, so additional information from group testing, the
Feedback Game, and teacher interviews was essential. For
example, one Feedback Game participant specifically men-
tioned that the poem excerpt helped in continuing the poem,
and the teacher noted that the pupils were faster to get to
work, worked more autonomously, and were more engaged
than in usual classes. Later on, the teacher also reported that
one pupil had started writing poems as a hobby after the test.

Evaluating the criterion fun was complex, as it included the
classification of gestures in peer tutoring. The children’s
self-reports and comments in the paired interviews and the
Feedback Game were an important supplement in this
evaluation. We categorized as many as 634 individual gestures
in the peer-tutoring sessions into 233 positive, 153 neutral, and
248 negative gestures. Although negative gestures
outnumbered the positive, only three children showed negative
gestures the most, two neutral gestures, and seven mainly
positive gestures (see Figure 4). Also the children’s self-reports
imply that the Poetry Machine was fun to use: In the paired
interviews, four pupils named fun as a reason to recommend
the tool to their friends, and the 10 Feedback Game
participants unanimously agreed that the Poetry Machine was
fun to use.

Peer-tutoring sessions were cumbersome and time-consuming
to conduct and analyze: Conducting the six sessions took about
3 hr 15 min, and the analyses for usability problems took almost
20 hr when done by two evaluators. Analyzing facial gestures,
feature usage, and comments took an additional 10 hr from one
evaluator. The six paired interviews took about 40 min, and
their analyses 4 hr 20 min, also by two evaluators. The two
group-testing sessions took about 1 hr, and their analyses 3 hr,
from one evaluator. Finally, the two Feedback Games took only
30 min, and their analyses only 1 hr 30 min, also by one
evaluator. In addition to being quicker to conduct, the group
testing let the evaluators concentrate on observation, as the
teacher instructed the pupils. The Feedback Game sessions, on
their part, were the easiest to analyze, as the children clearly
formulated their answers on top of the previous ones, and the
physical tokens made possible biases easy to spot and interpret.

Both in the peer tutoring and the group testing, the pupils’
questions were beneficial in revealing usability problems.
Especially the tutors’ answers were helpful, as they revealed also
the tutors’ mental models and hesitations. In peer-tutor-
ing sessions, the tutors made 76 comments during tutor
training and 124 comments when instructing the tutees,
whereas the tutees made only 65 comments. The most useful comments were the tutors’ instructions to the tutees (48) and descriptions of the poem writing process (7). In group testing, pupils made 85 comments: mostly questions for the teacher (34) or comments about the poems (10) but also hints to their friends (6), and some negative comments (4).

Altogether, the pupils gave 10 improvement ideas: six ideas in the paired interviews and four in the Feedback Games. The ideas in the Feedback Game underlined conceptual issues, whereas the ideas in the peer tutoring also dealt specific user interface issues.

If we compare peer tutoring and group testing, although applied with different aims, peer tutoring was very effective in detecting usability problems, revealing 6.3 problems per test user, whereas group testing revealed only 1.8 problems. However, if we take the usability problem analyses into account, it took about 18 min to find a problem in peer tutoring, whereas in group testing, it took only 13 min. On the part of the comments during the test sessions, peer tutoring prompted 265 comments in 3 hr 15 min (82 per hour) with 12 pupils, and group testing prompted 85 comments in 1 hr with 10 pupils. Table 1 summarizes the method comparisons.

### DISCUSSION

The number of test participants in our study (22) was in the same size range as in similar studies with children (e.g., Sim et al., 2006; van Kesteren et al., 2003), but it also shared the potential problem of having all the test participants from a limited sample of one class. This is a limitation that we shall address by conducting more tests at other schools.

In peer tutoring, we used only one general task, as open exploration is encouraged for this age group (Hanna et al., 1997). Even so, the test coverage was good, as each feature was tested by at least two pupils. However, the open task seemed to confuse the tutors, and a simple warm-up task might have been appropriate to break the ice. Selecting the more extrovert pupil in each pair as the tutor would also have given them more confidence in instructing the tutee.

Peer bias seemed to be stronger in the paired interviews than in the Feedback Game. In the Feedback Game, the player role and taking turns made it easier for the children to articulate both their agreements and disagreements, whereas in the paired interviews, either the tutor or the tutee usually dominated the discussion. The paired interviews also stressed positive answers, as all the pupils gave one of the two highest ratings for the tool, conforming to other studies with children giving overly positive feedback (e.g., Obrist et al., 2009; Read & MacFarlane, 2006). The answers in the Feedback Game, however, were more varied, including also poor ratings.

The tutors seemed to stress considerably as three adults observed them, whereas children participating in the group testing did not mind the large number of adults. In the analysis phase, having at least two evaluators is preferable to increase the reliability of the findings. For example, our three evaluators were very unanimous, as 86% of the usability problems discovered in peer tutoring were reported at least by two of them. Having fewer evaluators or less analysis time

| Method                  | Peer Tutoring | Group Testing | Paired Interview | Feedback Game |
|-------------------------|---------------|---------------|------------------|---------------|
| Participants            | 12            | 10            | 12               | 10            |
| Total duration          | 3 hr 15 min   | 1 hr          | 40 min           | 30 min        |
| Analysis time           | 20h (+10 hr gesture analysis) | 3 hr | 4 hr 20 min | 1 hr 30 min |
| Analyzers               | 2 (+1)        | 1             | 2                | 1             |
| Usability problems      | 76            | 18            | 16               | 3             |
| Problem severity (3/2/1) | 37/25/20     | 11/4/3        | 4/6/6            | 1/2/0         |
| Problems per user       | 6.3           | 1.8           | 1.3              | 0.3           |
| Problems per hour (sessions + analysis) | 3.3 | 4.6 | 3.1 | 1.5 |
| Comments                | 265           | 85            | Not available    | Not available |
| Comments per test hour  | 82            | 85            | Not available    | Not available |

Note. Group testing focused more on collecting feedback on the concept than on problem detection.
could leave some problems unnoticed, especially minor ones that still disturb the use.

**CONCLUSIONS**

We used a multimethod approach to evaluate an interactive prototype with children. Peer tutoring provided detailed information on the usability problems but was time-consuming to conduct and analyze, whereas group testing confirmed not only the most rudimentary usability problems but was more efficient. The new Feedback Game was quick and fun to play and easy to analyze. It was especially helpful in evaluating the criteria usefulness and fun, and in collecting ideas and feedback for the concept. However, the Feedback Game could focus only on a limited set of questions.

The evaluation results clearly show that multimethod approach is beneficial. No method alone could have produced equally useful results for such a broad scope. The various approaches also clearly supported each other’s results. Thereby, our first attempt to validate the Feedback Game seems successful. The results also show that quick and practical results can be obtained even with group-based methods. Therefore, we recommend using a set of group-based methods already in the early phases of development to test new ideas, detect missing features, and focus on specific questions puzzling the development team. Peer tutoring and other, more specific methods are relevant later on in the development process, as the details of the user interface are finalized.

**REFERENCES**

Edwards, H., & Benedyk, H. (2007). A comparison of usability evaluation methods for child participants in a school setting. In Proceedings of the 6th International Conference on Interaction Design and Children, IDC ’07 (pp. 9–16). New York, NY: ACM.

Hanna, L., Risden, K., & Alexander, K. (1997). Guidelines for usability testing with children. *Interactions*, 4(5), 9–14.

Hanna, L., Risden, K., Czerwinski, M., & Alexander, K. J. (1999). The role of usability research in designing children’s computer products. In A. Druin (Ed.), *The design of children’s technology* (pp. 3–26). San Francisco, CA: Morgan Kaufmann.

Hertzum, M., & Jacobsen, N. E. (2001) The evaluator effect: A chilling fact about usability evaluation methods. *International Journal of Human-Computer Interaction*, 13, 421–443.

Hourcade, J. P. (2008). Interaction design and children. *Foundations and Trends in Human–Computer Interaction*, 1, 277–392.

Höysniemi, J., Hämäläinen, P., & Turkk, L. (2003). Using peer tutoring in evaluating the usability of a physically interactive computer game with children. *Interacting With Computers*, 15, 203–225.

Lahti, J., Siira, E., & Törmänen, V. (2012). Development and evaluation of media enhanced learning application. In *Proceedings of the 11th International Conference on Mobile and Ubiquitous Multimedia*, MUM ’12 (pp. 5:1–5:10). New York, NY: ACM.

MacFarlane, S., Sim, G., & Horton, M. (2005). Assessing usability and fun in educational software. In *Proceedings of the 2005 Conference on Interaction Design and Children*, IDC ’05 (pp. 103–109). New York, NY: ACM.

Mack, R. L., & Nielsen, J. (1994). Executive summary. In J. Nielsen & R. L. Mack (Eds.), *Usability inspection methods* (pp. 1–23). New York, NY: Wiley.

Markopoulos, P., & Bekker, M. (2003). On the assessment of usability testing methods for children. *Interacting With Computers*, 15, 227–243.

Obrist, M., Igelsböck, J., Beck, E., Moser, C., Riegler, S., & Tscheligi, M. (2009). Now you need to laugh! Investigating fun in games with children. In *Proceedings of the International Conference on Advances in Computer Entertainment Technology*, ACE ’09 (pp. 81–88). New York, NY: ACM.

Obrist, M., Moser, C., Fuchsberger, V., Tscheligi, M., Markopoulos, P., & Höfstätter, J. (2011). Opportunities and challenges when designing and developing with kids @ school. In *Proceedings of the 10th International Conference on Interaction Design and Children*, IDC ’11 (pp. 264–267). New York, NY: ACM.

Patel, M., & Paulsen, C. A. (2002, June). *Strategies for recruiting children for usability tests*. Paper presented at the Meeting of the Usability Professionals Association, Orlando, FL. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.526.2950&rep=rep1&type=pdf.

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

Read, J., MacFarlane, S., & Casey, C. (2002). Endurability, engagement and expectations: Measuring children’s fun. In *Interaction design and children*, Vol. 2 (pp. 1–23). Eindhoven, Netherlands: Shaker.

Sim, G., MacFarlane, S., & Read, J. (2006). All work and no play: Measuring fun, usability, and learning in software for children. *Computer & Education*, 46, 235–248.

van Kesteren, I. E. H., Bekker, M. M., Vermeerem, A. P. O. S., & Lloyd, P. A. (2003). Assessing usability evaluation methods on their effectiveness to elicit verbal comments from children subjects. In *Proceedings of the 2003 Conference on Interaction Design and Children*, IDC ’03 (pp. 41–49). New York, NY: ACM.

Anna Kantosalo works as a PhD student at the Discovery Research group at the University of Helsinki. Her ongoing PhD work focuses on methods for describing and evaluating human–computer co-creativity in a meaningful way.

Sirpa Riihiaho has worked as a senior lecturer in Aalto University and a postdoctoral researcher at the University of Helsinki. Her research and teaching has focused on the methods for user-centered product development, especially on usability evaluation methods. Her doctoral thesis, “Experiences With Usability Testing: Effects of Thinking Aloud and Moderator Presence,” combined an extensive literature review with experiences of usability testing from 22 years, covering 143 usability studies.

We thank Karoliina Tiuraniemi and Mikko Hynninen for assisting with the test sessions and analysis, professor Hannu Toivonen for useful comments, and the pupils and the teacher who participated in the tests. This study was funded by the Academy of Finland (Decision 276897, CLIC) and the European Commission (FET Grant 617733, ConCreTe).

The authors prepared the work as employees of the University of Helsinki.