Enabling Unsupervised Eye Tracker Calibration by School Children through Games

Oleg Špakov
University of Tampere
Tampere, Finland

Howell Istance
University of Tampere
Tampere, Finland

Tiia Viitanen
Tampere University of Applied Sciences
Tampere, Finland

Harri Siirtola
University of Tampere
Tampere, Finland

Kari-Jouko Räihä
University of Tampere
Tampere, Finland

ABSTRACT

To use eye trackers in a school classroom, children need to be able to calibrate their own tracker unsupervised and on repeated occasions. A game designed specifically around the need to maintain their gaze in fixed locations was used to collect calibration and verification data. The data quality obtained was compared with a standard calibration procedure and another game, in two studies carried out in three elementary schools. One studied the effect on data quality over repeated occasions and the other studied the effect of age on data quality. The first showed that accuracy obtained from unsupervised calibration by children was twice as good after six occasions with the game requiring the fixed gaze location compared with the standard calibration, and as good as standard calibration by group of supervised adults. In the second study, age was found to have no effect on performance in the groups of children studied.

CCS CONCEPTS

• **Human-centered computing** → **Human computer interaction (HCI)**; • **Applied computing** → *Education*;

KEYWORDS

calibration, games, school children, low cost tracker, unsupervised

ACM Reference Format:
Oleg Špakov, Howell Istance, Tiia Viitanen, Harri Siirtola, and Kari-Jouko Räihä. 2018. Enabling Unsupervised Eye Tracker Calibration by School Children through Games. In *ETRA '18: 2018 Symposium on Eye Tracking Research and Applications, June 14–17, 2018, Warsaw, Poland.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3204493.3204534

1 INTRODUCTION

One area that offers many possibilities for exploiting the advent of low-cost eye tracking systems is education. Gaze measurement may be integrated into activities such as learning to read. Monitoring reading progress of a school class in real-time could enable the teacher to see which readers are having problems, and which parts of a reading text are causing problems to several readers. Collecting gaze-based reading performance data could enable the teacher to monitor the progress of individual students over time.

The extent to which this is feasible with low cost systems will depend very much on the data quality obtainable from the eye tracker over the course of a lesson lasting, say, 40 minutes. There are three necessary pre-requisites. First, the eye tracker is calibrated on each use occasion. With low sample trackers, modifying previously stored calibration data is unlikely to produce sufficiently accurate tracking performance. Second, the calibration needs to be done unsupervised by the individual student as there are not the teaching staff resources to oversee the calibration. This might be the case if the tracker is used in the laboratory or individual field test situation. Third, the calibration procedure may not take a disproportionate amount of time out of the lesson in comparison to the perceived benefits either to the teacher or to the student.

It is necessary to find a means of encouraging each child to calibrate their eye tracker quickly and carefully without supervision. As this will need to be done every time there is a reading lesson, the motivation to calibrate quickly and carefully needs to be maintained over multiple sessions.

This paper describes two studies carried out in three Finnish primary schools to compare data quality resulting from using two games to collect calibration data with a standard calibration procedure. The first study compares the data quality resulting when the procedures were tested by the same class over six repeated occasions, while the second study compares the same outcomes for children of different ages on a single occasion.

2 BACKGROUND

Games now constitute a fundamental part of digital entertainment culture and are a natural starting point for when considering suitable computer-based activities to motivate and engage children.

Broadly, a computer game consists of participation in an activity which has one or more players, rules, and a victory condition [Rogers 2014]. Characteristics of a game include: *Genre*, which relates to the gameplay interactions of the game external to the story or theme; *Theme*, which is the setting or scenario of the gameplay; *Actions*, which are the interactions the player makes with the game, such as opening chests, firing a weapon, or casting a spell; *Progression*, which marks how far along in the game a player is,

ETRA '18, June 14–17, 2018, Warsaw, Poland

ACM ISBN 978-1-4503-5706-7/18/06…$15.00
https://doi.org/10.1145/3204493.3204534

which could be an increase in level of the character, or progression through an over-arching story; *Rules* define the constraints a player must keep within; *Game mechanics*, which are the most basic building blocks of a game; a rule or description that covers a specific, single aspect of play, for example, pressing a sequence of buttons to make a character jump, or rotating a falling block of shapes; *rewards and achievements*, which are given to the player as a result of progression and performance.

Achievements are rewards collected alongside the main game and may be carried over between play occasions [Hamari and Eranti 2011]. Badges are a common type of achievement and may be accompanied by a reward, such as an amount of virtual currency.

The impact of an achievement system based on badges for class-based work in a large sample of university students was investigated by Denny [Denny 2013]. A highly significant positive effect was found on the quantity of contributions made by students, without any negative impact on quality. In general, they enjoyed being able to earn badges. The impact of gamifying two sets of tasks without much intrinsic interest carried out by 5 to 7 year old children was studied in a laboratory setting [Brewer et al. 2013]. Gamifying the task led to a significantly higher proportion of participants completing the tasks in both sets.

The application of games to a broad range of calibration activities has been discussed [Flatla et al. 2011]. They propose a framework that involves identifying calibration type, identifying the core calibration tasks, identifying one or more game mechanics for each task, and then adding additional games design elements. Three games for calibration were compared with their respective standard procedures, which were not related to eye tracking. The games produced a higher rating of enjoyment and in some cases better data quality.

Alternative means of eye tracker calibration have been studied with a view to making this activity, necessary on each use occasion, less tedious. Calibration using smooth pursuit eye movements instead of the usual fixed location approach has been studied [Pfeuffer et al. 2013]. It was observed by these, and other authors [Reingold 2014], that there is no guarantee that a user actually looks at the required location while calibration data is collected in the fixed location approach. In the smooth pursuit approach, the correlation between changes in gaze position in a time window with changes in the coordinates of moving objects on screen was studied. A person was encouraged to follow the path of a known object and then gaze data at known points on the object's trajectory was sampled. The advantages of this approach claimed by the authors are that the technique is tolerant to interruption and is able to calibrate users without them being aware of it. They did not claim advantages in relation to superior data quality over the fixed location approach.

The testing and reporting of data quality obtained from eye trackers rather than relying on data published by manufacturers has been advocated recently, particularly as the cost of eye tracking systems fall and the situations in which they are used increase. Standardized procedures for doing this have been proposed ([Holmqvist et al. 2012], [Akkil et al. 2014], [Niehorster et al. 2017], [Feit et al. 2017], [Špakov et al. 2017]). These share similar features and use accuracy and precision as the main quality metrics. After a system has been calibrated, then gaze position data is collected when a person is assumed to be looking at a known location. The difference between the known location and the average of the locations of gaze position samples collected when looking at that location is used as the measure of accuracy, and the dispersion of those samples is used as the measure of precision. Both measures may or may not be reported separately for the X and Y directions.

3 DESIGNING GAMES FOR CALIBRATION

Three games conditions were designed to collect calibration and verification data and compared with the standard fixed point calibration procedure. The game was layered over the standard fixed location calibration procedure. One reason for this is speed of completion in order to minimize the impact of this task on the classroom lesson in progress. Other approaches such as smooth pursuit are possible, as noted previously, and may be appropriate if these can be completed within an acceptable time. These were not investigated further in this study. In each condition, there were five fixed calibration locations, and between 4 and 10 verification locations. The eye tracker gave a value for the quality of the data collected at each calibration location. If the reported value was below a certain threshold, the point was re-calibrated. When all points had been calibrated, verification data was collected.

3.1 The Ball game (Standard procedure)

This was the standard calibration procedure where a ball moved between the calibration locations. This was called a game in order not to make this appear differently to the participants, although no feedback was provided. The player was instructed to follow the ball with their eyes and when it stopped, to wait for a second and then click the mouse button. The click caused the ball to move to the next location, where it stopped. After calibration was completed, the ball moved to the verification locations without any break.

3.2 The Mission game

The main game mechanic was designed to keep the player looking at a small area (40 x 40 px) on screen just before and just after eye position data was collected. An attribute (color or shape) of the small area cycled quickly through four options and the action required was that the player had to press a key (the space bar) when the target attribute was displayed. Each option appeared for a period between 0.43 to 0.8 seconds (1 in Figure 1). There was a 1 second delay before feedback was given at the same location showing whether or not the selection had been successful (2 and 3 in Figure 1). It was during this second that the calibration data was collected. During the calibration phase, five similar objects appeared at different locations on screen, one at a time. Prior to this, the player had been given the value to select on an instruction screen, for example select the 'yellow' object, or select the 'triangular' object. An important design consideration was that the mechanic is extensible, and can be embedded into different themes for games. The number of options cycled through could be increased or reduced, as could the display time.

The theme used in Level 1 was to open a door to let the player escape following a radioactive leak by releasing five locks, which were at the required calibration locations. There were four verification locations which followed on directly from the calibration. For additional motivation, a personal customizable player character,

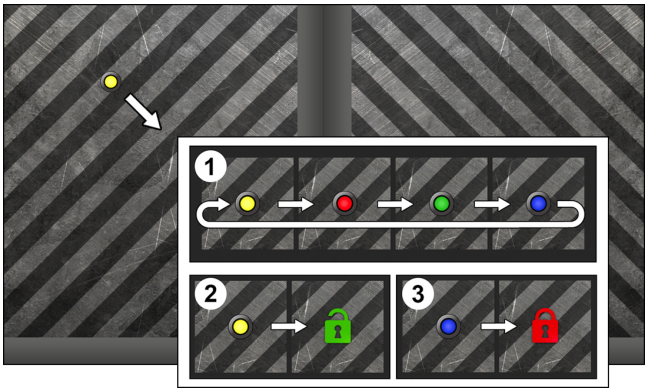


Figure 1: Main mechanic in Level 1, Mission game

levels and rewards were added. As a reward, points were awarded for speed and deducted for locks not opened in order to a) win the mission and b) gain achievements in the form of virtual currency and bonus accessories. In these trials, the game had two levels with different missions, each with their own background stories.

3.3 The Firefly/Troll game

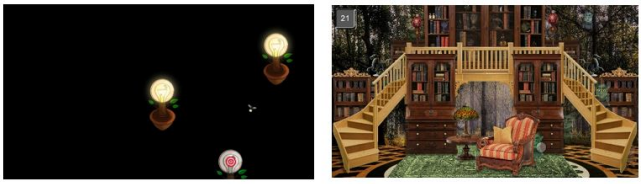
The Firefly game was similar to the standard calibration, but it was visualized as a firefly flying on the screen and turning on lamps. There was a bullseye on top of the light bulb to which the firefly would fly to and the user was instructed to look at. When the bullseye disappeared, the user would click the mouse button to turn the lamp on. Then a new lamp would appear and the firefly would fly to it (Figure 2a). Although instructed to do so, there was no penalty for not looking at the fly when the button was clicked.

When the calibration had been completed, a separate verification game immediately followed (the Troll game). A room appeared where the user played a hidden object game where they had to find and click on 10 troll characters displayed semi-transparently in the room (Figure 2b). The player was awarded a score based on the time and amount of trolls found within the time limit of 30 seconds, and their high scores were saved. The difference between the location of the mouse pointer and the location of the closest fixation to the pointer in the second prior to the click event was taken as the basis for validation (see Section 4.2.3). This is similar to the idea of Hornof and Halverson's *required fixation locations* [Hornof and Halverson 2002], where the user was required to click on a small target and was assumed to be looking at the pointer on the target when it was clicked.

There was no carry-over between game-playing occasions, other than the display of a 'Best' sign at the end of the Troll game if this was the player's highest score to date.

3.4 Encouraging the player to sit in a position acceptable for satisfactory tracking

To encourage the player to sit in a position and at a distance to provide suitable eye images for the tracker's cameras, an initial screen preceded the games (shown in Figure 3).



a) Firefly game b) Troll game

Figure 2: Firefly game (calibration) and Troll game (verification)

a) Ball Game and Firefly game b) Mission game

Figure 3: Means of encouraging the player to sit within the head box of the eye tracker

With the Ball game and Firefly game, the player was asked to sit so that the yellow circles were the center of the brown box (Figure 3a). When this happened the 'Aloita' ('Begin') button appeared, the player clicked on this and the game screen appeared. If the player moved out of position before clicking the button, the button disappeared. In the Mission game, the player was asked to move their position to move the eye images shown as red circles to match the location of the character's eyes (Figure 3b). When this happened, the initial game instruction screen appeared automatically. There was little difference between the acceptability thresholds in both methods. There was an automatic accept in the Mission game after a delay when an acceptable position was logged, and a manual accept in the other two conditions where a button appeared after the same delay. Positioning participants only took place in the beginning before each game and probably changed during the game for all games. They were not made to keep their head or eyes in that area during the actual period of game play.

4 EVALUATION OF THE EFFECTIVENESS OF GAMES FOR EYE TRACKER CALIBRATION

The purpose of using games for eye tracker calibration was to investigate whether unsupervised calibration by elementary school students could produce satisfactory levels of data quality over multiple use occasions at an acceptable cost in terms of time taken and other disruption. This can be resolved into two research questions:

RQ1: do games-based calibration procedures provide better quality of data than the standard procedure at an acceptable cost in terms of time taken to complete the procedure?

RQ2: is there a change in calibration data quality over repeated occasions?

The ages of children in elementary school in Finland range from 7 to 11, which prompts a further question.

RQ3: does the effect of games-based calibration procedures on data quality depend on the age of the students?

4.1 Equipment

Six Dell E7520 laptops with a screen resolution of 1366 x 768 and a 12.5 inch screen were used, each equipped with a Visual Interaction myGaze n eye tracker, and a mouse. The myGaze n was a remote binocular tracker with a sample rate of 30Hz, with manufacturer's estimates of accuracy and spatial resolution of 0.4° and 0.05° respectively. The three game solutions were built into a customized version of the general eye tracker driver software, ETU Driver,

4.2 Study 1: (RQ1 and RQ2) Investigation of the impact of games-based calibration on data quality

All of the students in a second-grade class (7-8 years old) of one of the schools completed the games in a single session, which was then repeated on five subsequent occasions. The session took about 10 minutes to complete. The class teacher had randomly allocated students to four groups of up to six students each, and the students remained in the same group for each session. The sessions were completed on separate days over a 10-day period, three in one week and three in the following week. The order of the games that each participant completed was counterbalanced over the six sessions. Sessions took place in a room adjacent to their normal classroom (see Figure 4). Prior to the data collection, a presentation of the games was made to the whole class. After the final session, there was a 10-minute discussion about each of the games with each group. School procedures for obtaining parental consent for each student's participation in the trials were followed. A pilot study to test the data collection procedures with a large group of participants in the school was not carried not as this would have meant that they would have already seen and played the games before the first session of the study. It would have been difficult (though not impossible) to have conducted a pilot test on a different group of participants, and this would have reduced the data loss in Session 1 of the study.

The dependent variables were:

- the data quality, assessed by the accuracy and precision of the data collected in the verification, and by the number of points requiring recalibration.
- the time taken to complete the games
- ratings of how engaging each game was (made after each session)

It was expected that the Mission game would produce the best data quality, initially and over the repeated sessions, due to the mechanic requiring attention to be maintained on the verification point, and the greater engagement expected with the game. It was expected that the Firefly/Troll game would result in greater engagement than the standard procedure, although not necessarily better data quality as the calibration mechanic was similar.

4.2.1 Observations from the data collection trials. It was intended that, as far as possible, each session would be conducted without any direct supervision or intervention. During the first session, however frequent interventions were required to adjust the sitting position of the participant in relation to the eye tracker. In general, the sitting position was too low in relation to the desk and the eye tracker and required the screen angle to be increased to an almost vertical position. For the second and subsequent sessions, cushions were provided (used in the session shown in Figure 4), very little intervention was required, and participants completed the games generally unsupervised.

Figure 4: A group of six students taking part in a data collection session (Study 1)

4.2.2 Definition of a fixation. The fixation detection algorithm accumulates gaze points to a single fixation as long as their location is not further from the current fixation center than a certain threshold [Špakov 2012]. The threshold was 50 pixels (~1.5°) as the myGaze tracker applied a two-state low-pass filter and temporally adjacent gaze points within a fixation lay relatively close to each other. The fixation center was calculated as a simple average of all gaze points assigned to this fixation. Two consecutive gaze points within 50 pixels of each other falling outside of the current fixation formed a new fixation.

4.2.3 Definition of the verification fixation. In the Ball game and the Firefly/Troll game the notion of a 'verification' fixation was used. Instead of using the fixation in progress when the selection event was made as the datum point for verification, the fixations made in the 1 second preceding the selection event were examined. The closest fixation to the verification reference point was selected, provided the duration of this fixation exceeded 350 ms. The reference point for the verification was taken to be the location of the pointer for the Firefly/Troll game. The size of an individual troll was relatively large, and it was not known where exactly a player would look when selecting one. The center of the target object in the Mission game and the Ball game was taken as its reference point.

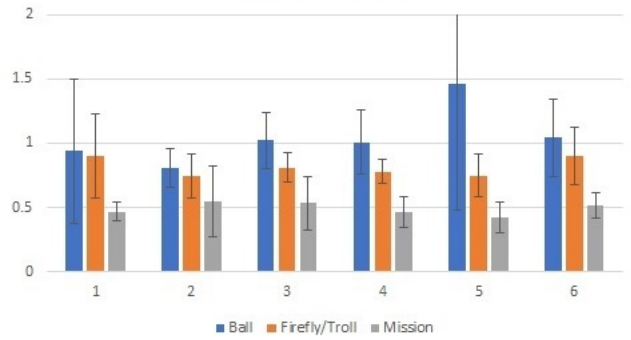


Figure 5: Average accuracy (cm) for the 3 games, sessions 1 to 6, error bar: 95% Confidence Interval

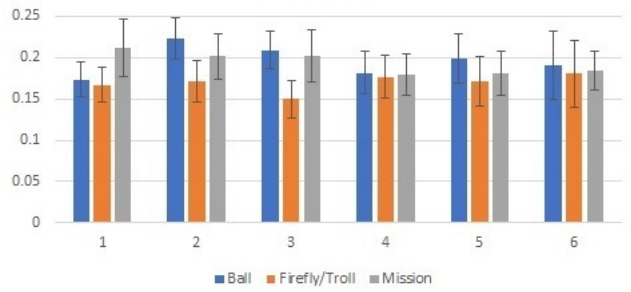


Figure 6: Average precision (cm) for the 3 games, sessions 1 to 6, error bar: 95% Confidence Interval

The distance from the center of the closest fixation to the reference point was used as the measure of accuracy and the standard deviation of the gaze points within the whole verification fixation was used as the measure of precision, in accordance with the TraQuMe formulae [Akkil et al. 2014].

4.2.4 Results - data quality. Of the 23 students in the class, one student was absent for five sessions, one was absent for four sessions and one for two sessions. Data from 368 games were collected in total from a possible maximum of 381, which represents an overall data loss of 3.4%. In Session 1, data from 55 games out of a possible 63 were collected (a 12.7% data loss) due mostly to problems with the data collection system rather than the eye trackers.

The accuracy and precision in a game were computed as averages in centimeters for all individual verification points across participants that had valid gaze data.

Figure 5 shows the average accuracy values in centimeters. The viewing distance was not fixed, nor recorded, so accuracy data is not given in degrees of visual angle but in centimeters. The smaller values indicate greater accuracy. The accuracy provided by the Mission game (~0.5 cm) is about twice as good as that provided by the standard calibration (~1.0 cm). The error bars show the 95% confidence intervals of where the population mean lies in relation to the sample mean. The differences in accuracy between the games accord with the expected outcomes for RQ1. Regarding RQ2, there

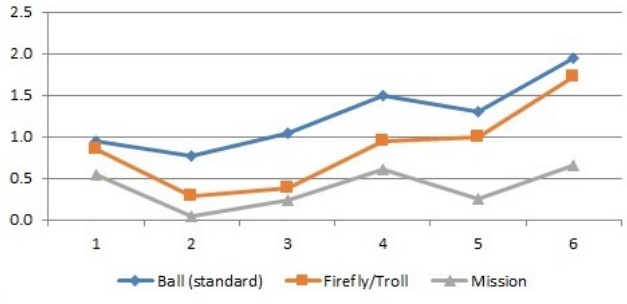


Figure 7: Average number of recalibrations for every 5 calibration points over all 6 sessions

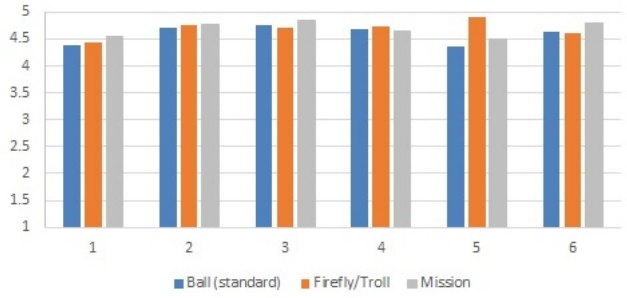


Figure 8: Average ratings made after each session, 5 = really good, 1 = really boring

is no evidence that accuracy gets worse over the six sessions for any of the calibration conditions.

Figure 6 shows the corresponding data for precision. The range of precision values is very small, probably as a result of the filtering applied to consecutive data samples by the eye tracker. The averages of most precision measures shown lie within a 0.1 centimeter range. There is no evidence of a systematic difference between games or over repeated sessions in the measures of precision.

Each game used five calibration locations on the screen, and Figure 7 shows the average number of recalibrations required for each game. Over the sessions, the number of recalibrations for the Ball game doubles, while that of the Mission game is nearly half this amount to start with, and remains at, or below, the initial level throughout. The performance of the Firefly calibration is initially similar to the Ball game, then improves in Sessions 2 and 3, but then deteriorates over sessions 4 to 6 and mirrors that of the Ball game. By the final session, the number of recalibrations required for the standard calibration method is three times that of the Mission game. There is then an effect in the predicted direction over time (RQ2) when the number of recalibrations is taken as the measure of data quality.

4.2.5 Subjective feedback. Contrary to expectations, the students were very positive about all the games over each of the sessions in the ratings made after each session. Figure 8 shows the average of rating values (5 being the most favourable and 1 the least). A 'Smileyometer' was used to obtain the ratings, which has

been found to be usable with students at elementary school level [Read et al. 2002]. There was no apparent effect on rating values either by game or by session. It was noted that on occasion, a student could make negative comments during a game, but afterwards still give it a very positive rating. As the students were not experienced games testers, in spite of being told what to rate, some of them may have been responding to how well they thought they did.

To gain more detailed feedback, group discussions were held and recorded with each group after the last session. The students were asked which game they liked the best and the least, and why. Ten out of the 20 students present in the final session considered the Ball game to be the most boring of the three. Most could not come up with any good qualities, but a few commented that they liked the movement of the ball. On the other hand, some also considered that as being negative. The Firefly game divided opinions more evenly, with nine votes for it being the best, and ten for it being the worst of the three. Positive things mentioned about it were the trolls and making a high score. However, the initial calibration (firefly and lamps) was still considered boring and too slow. The Mission game received 11 votes for the best game, and none to be the worst. Some of the students felt that the game was too difficult. Positive feedback included opening the locks (with preference to shapes instead of colors) and making a "high score" with the amount of locks opened.

4.2.6 Duration. The overall durations of the three games (Figure 9) are affected by the time for the verification process and the number of re-calibrations, as well as the time required to read the instruction screen. These are very consistent after the first day and the Ball game is completed about twice as quickly than the Mission game, the difference being about 40 seconds. A possible reason for the increase in the duration of the Mission game in Session 6 is that all students were moved to Level 2. The mechanic was similar in that a target shape had to be selected instead of a target color. However, moving to the new level necessitated reading new instructions, which would affect the overall duration.

The time required for the calibration only when the initial instructions and verification are discounted is consistent after the first session. This represents a minimum time overhead. In the Mission game, the player had to wait while the target cycled through each of the color options or shape options, whereas in the Firefly and Ball game, there were a greater number of recalibrations. The standard calibration was consistently faster by about 5 seconds.

4.2.7 Comparison of results obtained from supervised calibrations. It is informative to compare the data quality obtained from the school with that collected from adults during a supervised calibration using the standard procedure with the same eye tracker. These data provide a benchmark of what could be expected in laboratory testing conditions. Table 2 shows the averages for the numbers of recalibrations, accuracy and precision for Session 1 and Session 6 compared with a group of 12 adults undertaking the standard calibration under supervision. Here, the seating position, viewing distance and screen angle were checked, and adjusted if necessary, before the calibration using the Ball game only. For the standard calibration, the data quality from the unsupervised students is clearly much worse than from the supervised adults. However, the data quality obtained from the students using the

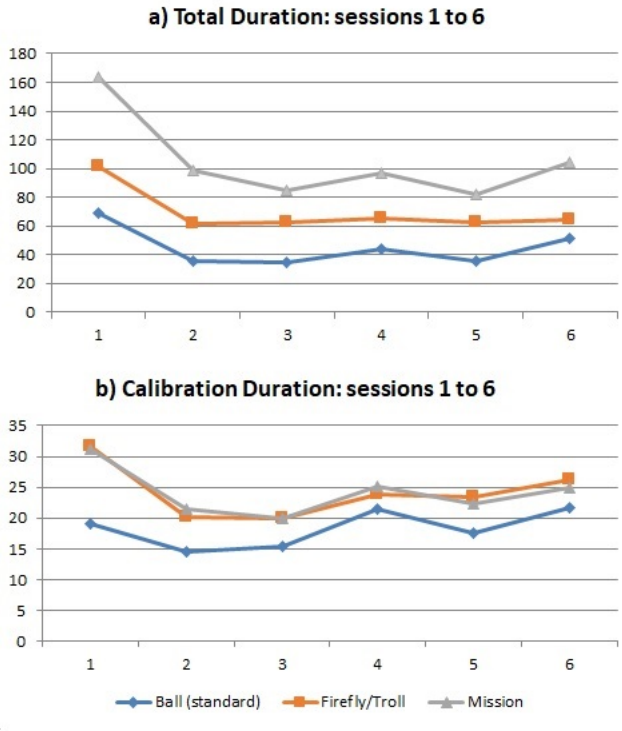


Figure 9: Durations of games (in seconds) over each session

Mission game is as good if not better than from the supervised adults using the standard procedure. Comparing data from Session 1, the difference between the accuracy of adults using the standard calibration (mean = 0.68cm) and students using the Mission game (mean = 0.47cm) approaches a significant difference ($p = 0.06$). Comparing the accuracy from the students in Session 6 using the Mission game (mean = 0.52) with the adult data, the difference is not significant.

We can compare our data with data quality values obtained from a study in which a large number of adults performed a task similar to the verification task in two different lighting conditions and with two different eye trackers [Feit et al. 2017]. The tracker used in our study is a low cost version of one of the trackers used in that study, the SMI REDn scientific. The mean values for accuracy and precision from unsupervised students using the Mission game in our study are similar to those reported for adults using a similar tracker with a higher sample rate (60Hz) in Feit et al.'s study.

4.3 Study 2: (RQ3) Investigation of the impact of age on the effect of games-based calibration

Study 2 repeated the first session of Study 1 with elementary school students in the third grade and fifth grade. There were:

- School A : 21 second grade students, age 7-8, (this data was from Session 1, Study 1)
- School B : 19 third grade students, age 8-9
- School C : 21 third grade students, age 8-9

Table 1: Average accuracy and precision (cm) for students in Study 1 and a control group of adults

			Ball game (standard calibration)			Mission Game		
		n	Recalibs *mean*	Accuracy *mean (sd)*	Precision *mean (sd)*	Recalibs *mean*	Accuracy *mean*	Precision *mean (sd)*
students	session 1	21	1	0.94 (1.31)	0.17 (0.05)	0.5	0.47 (0.13)	0.21 (0.06)
	session 6	18	1.9	1.04 (0.65)	0.19 (0.09)	0.7	0.52 (0.23)	0.18 (0.05)
adults		12	0.75	0.68 (0.34)	0.22 (0.05)			
Feit et al.		81	X	0.58 (0.75)	0.51 (0.91)			
			Y	0.66 (0.57)	0.51 (0.64)			

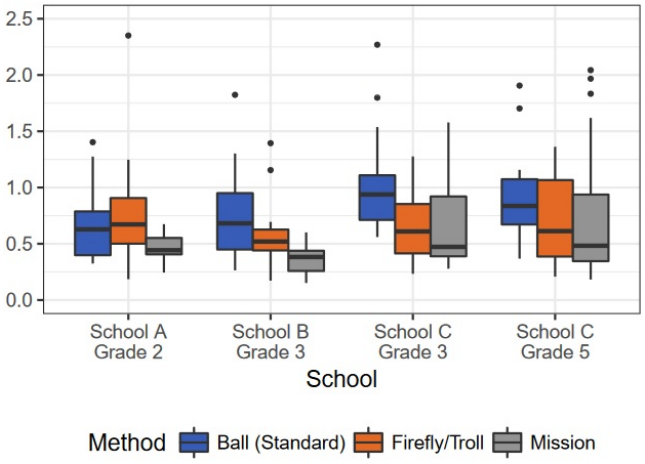


Figure 10: Median accuracy (cm) according to age groups and school (Study 2)

- School B : 20 fifth grade students, age 10-11

It is less easy to predict what the impact of age on the study outcomes will be. Older students will probably read instructions more quickly, and may find games less challenging and interesting, than younger students.

There were some differences between schools in how the data was collected. In School C, third and fifth grades, the data was collected in the classroom with other students present rather than in a separate adjacent area. In some cases other students looked over the shoulders of the test students, possibly distracting them.

Table 2 shows the accuracy and numbers of recalibrations from Study 2. The precision data is omitted as the ranges of average values between all conditions were very small as was the case in Study 1. The top row is the data from Session 1, Study 1 for Grade 2 students and the other rows contain the data collected from the other two schools. The differences in average accuracies between the standard calibration and the Mission game are less apparent here than in Study 1. However, the Mission game still requires fewer recalibrations than the standard procedure. The high number of recalibrations for the Ball game in School C, third grade, was influenced heavily by two participants, who together required 17 recalibrations, which may have been the result of the additional distraction.

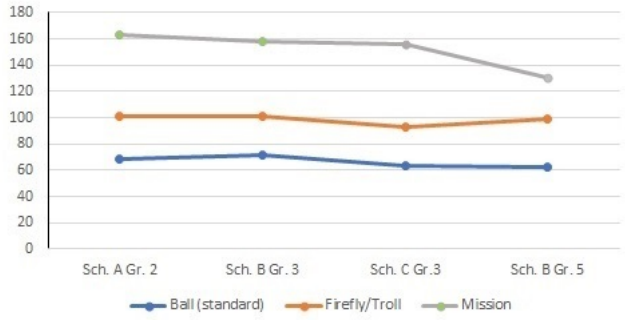


Figure 11: Study 2: Durations of the trials(seconds)

Looking instead at the medians of the accuracy to reduce the impact of individual students who may have had specific problems (Figure 10), the accuracy of the verification data is in the same order as obtained in Study 1 across the three new groups tested.

The overall durations of the game play (Figure 11) suggest an age effect possibly related to reading speed. The Mission game requires most reading, and the time to complete this game decreases with the age of the participants by 20 seconds.

5 DISCUSSION

Although game-based calibration is not new, the main contribution of this work to the field is presenting and quantifying the impact of using a specifically designed game mechanic for eye tracker calibration in a particularly demanding situation (unsupervised calibration by children). The outcomes of the two calibration games, one with and one without the specific mechanic show that simply making calibration into a game is not sufficient. Abstracting the core calibration task, building a mechanic around this, and then embedding this in the Mission game leads to better data quality than another game-based approach to calibration (the Firefly/Troll game), and better than the standard procedure (the Ball game).

Over repeated sessions, the Mission game provided consistently better accuracy than the Firefly/Troll game and the standard procedure. While there was no evidence that accuracy deteriorated over the sessions with any of the conditions, the numbers of calibration points requiring to be redone increased for the Firefly/Troll game and standard procedure but not for the Mission game.

Table 2: Average accuracy (cm) and Recalibrations - Study 2

		Ball game (standard calibration)		Firefly/Troll		Mission Game	
	n	Recalibs	Accuracy	Recalibs	Accuracy	Recalibs	Accuracy
students		*mean*	*mean (sd)*	*mean*	*mean (sd)*	*mean (sd)*	*mean (sd)*
School A, Grade 2	21	1.0	0.94 (1.31)	0.9	0.90 (0.77)	0.5	0.47 (0.13)
School B, Grade 3	21	1.1	0.73 (0.39)	0.5	0.71 (0.71)	0.9	0.70 (1.59)
School C, Grade 3	19	2.2	1.03 (0.44)	0.8	0.90 (1.15)	0.6	0.90 (1.20)
School C, Grade 5	20	1.1	0.89 (0.39)	0.8	0.79 (0.60)	1.1	0.75 (0.63)

The data quality provided by the standard calibration carried out by adults using the same eye tracking equipment under supervised laboratory conditions provides a useful benchmark. While the data quality of the unsupervised students was understandably much worse than the supervised adults using the standard procedure, the data quality provided by the Mission game was actually better than that obtained from the adults.

The framework advocated by Flatla et al. in Section 2 has been shown to be very effective in this application, even though eye tracker calibration was not specifically addressed in their work. The essential calibration task of maintaining the gaze point at a series of specific locations was abstracted and formed the basis for designing a game mechanic. Subsequently, a game that used this mechanic was designed, which included a reward and achievement system to promote motivation over several repeated sessions of use. Significantly, there was little or no emphasis on calibrating the eye tracker in describing the reason to the participants for playing the games.

Making the distinction between the mechanic and the game in which it is embedded provides a good approach to the issue of needing to repeat the calibration every time the eye tracker is used which could be several times a week in a school classroom. Different games and different levels within the same game can be designed around the same mechanic. In addition to selecting color and shape as attributes of the target object, for example, letters or numbers could be selected to make a password, provided that feedback about the selection is given at exactly the same location as the object, as was show in Figure 1. The number of options cycled through may be changed to make the task more or less difficult, and the time each is exposed for can also be varied.

The system of rewards and achievements added to motivate improving performance over repeated occasions can be transferred between games. In this study, a customizable character was provided together with the idea of current level and acquiring virtual currency to enable the individual to buy assets to further customize their character.

The second study where data quality outcomes were compared between age groups and across schools shows that this approach to unsupervised calibration is valid in the second grade and upwards in elementary schools. Further design work is needed to reduce the requirement for participants to read instructions. Then similar tests need to be carried out using first grade students. Carrying out the tests in the classroom as opposed to an adjacent room, as was the case in School C, showed that data quality may be affected by other students distracting the student while playing a game. This highlights the need for field testing of gaze-enabled educational aids under realistic conditions.

The design of the game or games needs to be such that the activity does not occupy too much time as a proportion of a lesson. The duration of the Mission game with verification in Session 6 was on average 100 seconds. Of this the actual calibration data collection took on average 23 seconds. Having established the benefits of using the game mechanic, further work is needed to investigate how much the game duration can be reduced without compromising engagement with the game. There is a case for making the verification phase a check on the data quality after calibration, in order to trigger a complete recalibration. However there is a danger that a student may deliberately score badly to extend the game play time.

Seating in the classroom was an issue that caused significant problems in terms of the height of the students' eyes above the desk and the eye tracker. In Study 1, the seated eye height was too low. This was corrected by using cushions, which were not normally used in the classroom. This was not a problem in Schools B and C in Study 2 however. In another school that took part in one of our previous studies, parents were encouraged to provide a 'stability ball' for their children to sit on as this afforded postural benefits. One effect was that the students' eye height above the desk was too great for an eye tracker. Changing the angle of the screen led to reflections from interior lighting.

6 CONCLUSIONS

The study has shown a game with the core mechanic designed around the need to keep the gaze point fixed for short periods at specific locations on the screen is an effective means of encouraging elementary school students to undertake unsupervised calibration of an eye tracker. The mechanic is essential to providing good data quality. A calibration game without this mechanic did not deliver the same data quality. The same mechanic can be embedded in different game themes meaning that it is not necessary to always use the same game for repeated calibrations. It was shown too that the game solution adopted is capable of motivating children of different ages with an elementary school. The issue of motivating adults to undertake unsupervised careful calibration of eye tracking equipment built into personal computing equipment will need to be addressed if the promise of ubiquitous gaze-based interaction is to be realized. Games may well offer an effective route to this objective.

ACKNOWLEDGMENTS

We would sincerely like to thank staff and students at Nuolialan, Lamminpää and Peltolammi elementary schools in the Tampere region who took part in the study so enthusiastically, and, in particular, Irja Kivikangas of Nuolialan School, and Sanna Salonen of Lamminpää School for their help. The work was supported by the Academy of Finland under grant number 2501287895.

REFERENCES

Deepak Akkil, Poika Isokoski, Jari Kangas, Jussi Rantala, and Roope Raisamo. 2014. TraQuMe: A Tool for Measuring the Gaze Tracking Quality. In *Proceedings of the Symposium on Eye Tracking Research and Applications (ETRA '14)*. ACM, New York, NY, USA, 327–330. https://doi.org/10.1145/2578153.2578192

Robin Brewer, Lisa Anthony, Quincy Brown, Germaine Irwin, Jaye Nias, and Berthel Tate. 2013. Using gamification to motivate children to complete empirical studies in lab environments. In *Proceedings of the 12th International Conference on Interaction Design and Children*. ACM, 388–391.

Paul Denny. 2013. The Effect of Virtual Achievements on Student Engagement. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13)*. ACM, New York, NY, USA, 763–772. https://doi.org/10.1145/2470654.2470763

Anna Maria Feit, Shane Williams, Arturo Toledo, Ann Paradiso, Harish Kulkarni, Shaun Kane, and Meredith Ringel Morris. 2017. Toward Everyday Gaze Input: Accuracy and Precision of Eye Tracking and Implications for Design. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. ACM, New York, NY, USA, 1118–1130. https://doi.org/10.1145/3025453.3025599

David R. Flatla, Carl Gutwin, Lennart E. Nacke, Scott Bateman, and Regan L. Mandryk. 2011. Calibration Games: Making Calibration Tasks Enjoyable by Adding Motivating Game Elements. In *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology (UIST '11)*. ACM, New York, NY, USA, 403–412. https://doi.org/10.1145/2047196.2047248

Juho Hamari and Veikko Eranti. 2011. Framework for Designing and Evaluating Game Achievements.. In *Digra Conference*.

Kenneth Holmqvist, Marcus Nyström, and Fiona Mulvey. 2012. Eye Tracker Data Quality: What It is and How to Measure It. In *Proceedings of the Symposium on Eye Tracking Research and Applications (ETRA '12)*. ACM, New York, NY, USA, 45–52. https://doi.org/10.1145/2168556.2168563

Anthony J. Hornof and Tim Halverson. 2002. Cleaning up systematic error in eye-tracking data by using required fixation locations. *Behavior Research Methods, Instruments, & Computers* 34, 4 (01 Nov 2002), 592–604. https://doi.org/10.3758/BF03195487

Diederick C. Niehorster, Tim H. W. Cornelissen, Kenneth Holmqvist, Ignace T. C. Hooge, and Roy S. Hessels. 2017. What to expect from your remote eye-tracker when participants are unrestrained. *Behavior Research Methods* (15 Feb 2017). https://doi.org/10.3758/s13428-017-0863-0

Ken Pfeuffer, Melodie Vidal, Jayson Turner, Andreas Bulling, and Hans Gellersen. 2013. Pursuit Calibration: Making Gaze Calibration Less Tedious and More Flexible. In *Proceedings of the 26th Annual ACM Symposium on User Interface Software and Technology (UIST '13)*. ACM, New York, NY, USA, 261–270. https://doi.org/10.1145/2501988.2501998

Janet C Read, SJ MacFarlane, and Chris Casey. 2002. Endurability, engagement and expectations: Measuring children's fun. In *Interaction design and children*, Vol. 2. Shaker Publishing Eindhoven, 1–23.

Eyal M. Reingold. 2014. Eye tracking research and technology: Towards objective measurement of data quality. *Visual Cognition* 22, 3-4 (2014), 635–652. https://doi.org/10.1080/13506285.2013.876481 arXiv:https://doi.org/10.1080/13506285.2013.876481 PMID: 24771998.

Scott Rogers. 2014. *Level Up! The Guide to Great Video Game Design* (2nd ed.). Wiley.

Oleg Špakov. 2012. Comparison of Eye Movement Filters Used in HCI. In *Proceedings of the Symposium on Eye Tracking Research and Applications (ETRA '12)*. ACM, New York, NY, USA, 281–284. https://doi.org/10.1145/2168556.2168616

Oleg Špakov, Harri Siirtola, Howell Istance, and Kari-Jouko Räihä. 2017. Visualizing the Reading Activity of People Learning to Read. *Journal of Eye Movement Research* 10, 5 (2017). https://bop.unibe.ch/index.php/JEMR/article/view/3735