Calibration Games: Making Calibration Tasks Enjoyable by Adding Motivating Game Elements

David R. Flatla¹, Carl Gutwin¹, Lennart E. Nacke², Scott Bateman¹, Regan L. Mandryk¹
¹Department of Computer Science
University of Saskatchewan
Saskatoon, Saskatchewan, Canada
david.flatla@usask.ca, gutwin@cs.usask.ca, lennart.nacke@acm.org, scott.bateman@usask.ca, regan@cs.usask.ca

²Faculty of Business and Information Technology
University of Ontario Institute of Technology
Oshawa, Ontario, Canada

ABSTRACT
Interactive systems often require calibration to ensure that input and output are optimally configured. Without calibration, user performance can degrade (e.g., if an input device is not adjusted for the user’s abilities), errors can increase (e.g., if color spaces are not matched), and some interactions may not be possible (e.g., use of an eye tracker). The value of calibration is often lost, however, because many calibration processes can be tedious and unenjoyable, and many users avoid them altogether. To address this problem, we propose calibration games that gather calibration data in an engaging and entertaining manner. To facilitate the creation of calibration games, we present design guidelines that map common types of calibration to core tasks, and then to well-known game mechanics. To evaluate the approach, we developed three calibration games and compared them to standard procedures. Results showed that users found the game versions significantly more enjoyable than regular calibration procedures, without compromising the quality of the data. Calibration games are a novel way to motivate users to carry out calibrations, thereby improving the performance and accuracy of many human-computer systems.

ACM Classification: H5.2. Information interfaces and presentation: User Interfaces.

Keywords: Calibration, computer games, modeling

INTRODUCTION
Many kinds of interaction with computers require some type of calibration. For example, a user’s performance with a new input device may need to be tested in order to choose an optimal control-to-display ratio; a user’s ability to see different colors must be established before choosing colors for a visualization; or two representations of a space must be brought into alignment (e.g., a tabletop display and the direct-input device used with it). Calibration, or the lack of it, can have substantial effects on the success of an interaction with a computer system: for example, people may perform poorly with an input device that is incorrectly configured, people may miss important signals if stimuli cannot be detected or differentiated, and selection errors may result from improper alignment of input and output space.

Unfortunately, calibration is often a tedious and dull process, and as a result, many people do not carry out the calibration. Compounding the problem, some situations require that a calibration be carried out multiple times (e.g., as environmental conditions change, or as sensors drift). Our solution to this problem is to make calibration more enjoyable and less tedious. If calibration is less painful, more people will do it, and more people can gain the benefits of correctly-configured devices and interfaces. To achieve this goal, we create calibration games that gather calibration data in an engaging and entertaining manner. To aid in the design of these games, we present a design framework for creating a calibration game from a standard calibration procedure. We identify nine major types of calibration, decompose these types to find the core perceptual and motor tasks that provide the calibration data, match each core task to common game mechanics that can be used as the basis for a calibration game, and specify additional game elements (challenge, theme, reward, and progress) that can be added to improve the quality of the game.

We have developed several calibration games using this framework including a space-invaders-style game to determine color perceptibility, a shooting-gallery game to set optimal C:D ratios, and a launching game that measures the input range for a physiological sensor.

We investigated two main questions with these games: first, whether calibration games are more enjoyable than their standard counterparts, and second, whether the game versions introduced any problems in the quality of the data that was gathered. Our studies showed that users found all three of the game versions to be significantly more enjoyable – suggesting that we can in fact improve the likelihood that users carry out calibrations. In addition, although there were differences in the actual data collected by the game and standard versions, these differences did not reduce the utility of the game data for calibration.

Our work makes four contributions. First, we identify the idea of gamifying calibration as a way to improve motivation to carry out a calibration procedure. Second, we present a framework that simplifies the design of calibration games and shows the broad applicability of the idea. Third, we demonstrate the feasibility of calibration games with three examples. Fourth, we show that calibration games are
significantly more enjoyable than standard versions, without compromising data quality.

RELATED WORK

Calibration

Calibration is a necessary task for many different input and output technologies (e.g., [1,5,6,8]). The need for calibration arises out of individual user differences [5,15,22], and environmental and situational changes [5,8,23]. Examples of systems that have addressed calibration include technologies for interactive touch [8,23], 3D sound [15], location-sensing technology [6], eye-tracking [20], heads-up displays [1], glove-based input [10], personalized color displays [5], and physiological sensing — for detecting emotional state [22] or muscle exertion [11].

There is little previous research on improving the enjoyability of calibration — we found only one project that proposed a ‘game’ for calibration. However, this work did not describe how their activity was a game or why a game for calibration would be desirable (instead, the authors describe the game as a simple task that met the requirements of the calibration) [6]. One other similar project considered the idea of using first-person shooter games for Fitts’s Law tasks [9], and found that FPS targeting is well modeled by Fitts’s Law. This work suggests that adding game elements to calibration may not unduly alter the core task.

The majority of work on improving calibration has focused on reducing costs: through new mathematical models that reduce sampling requirements [5,15,20], improved calibration protocols that address current inefficiencies [1,10], or dynamic adaptation during the procedure [11]. However, these strategies are not applicable to all types of calibration, and do not address motivation issues.

Gamification

Gamification refers to the use of gameplay mechanics in non-gaming applications to encourage a desired type of behavior [4]. By using techniques such as scoreboards and personalized fast feedback, people feel more ownership and purpose when engaging with tasks [16].

HCI research has previously investigated the use of game elements to engage people in work (e.g., the ESP Game [21] or PSDoom [3]). This research suggests that incorporating game elements into work activities improves motivation [19], but also cautions that it is important to carefully design the integration of game elements with work tasks [21]. In our work, we extend this notion by providing a framework that clearly maps elements of the task to be accomplished (core calibration tasks) to elements of games.

There are several organizations of game elements that have been considered for gamification. First, Malone’s early work identified three key concepts of games: challenge (i.e., providing a goal with uncertain attainment), fantasy (i.e., evoking images of non-present and vicarious objects or situations), and curiosity (i.e., motivating users to learn independently) [13]. Reeves et al. [17] list ten ingredients of great games that could be used to improve services and non-game applications: self-representation with avatars, three-dimensional environments, narrative context, feedback and behavior reinforcement, reputations (ranks and levels), marketplace and economies, competition within rules, teams, parallel communication systems, and time pressure. Another set of design elements are Brathwaite and Schreiber’s ‘game design atoms’ [2]: the game state, game view, player representation, game mechanics (i.e., the game rule system), game dynamics (i.e., game mechanics set in motion by players), goals, and game theme. Most of these atoms are built on a formal game framework called MDA (mechanics, dynamics, aesthetics) [7]. This framework distinguishes mechanics as game components on the algorithmic level, dynamics as run-time interaction with the player, and aesthetics as the desired player emotion triggered by mechanics and dynamics.

A DESIGN FRAMEWORK FOR CALIBRATION GAMES

Our framework has four parts, which can be used as a basic design process for building many types of calibration games. The sections below discuss the parts of the framework and cover nine main types of calibration, core tasks that obtain the calibration data, game mechanics that can be built from the core tasks, and additional game elements that improve the subjective aspects of the game.

To develop this framework, we used an iterative process of finding calibration types through literature search, isolating the core task for each calibration type, and identifying the associated game mechanic. Previous literature helped identify the four game design elements [7,13].

Framework Part 1: Calibration Types

There are two main types of calibration in HCI — one addressing human abilities and limitations, and another addressing systems and technologies. First, calibrating for human capabilities determines performance characteristics for perceptual input and motor output.

- **Perceptual thresholds.** These calibrations determine baselines of human perceptual ability, such as the quietest tone that can be heard, or the minimum light level that can be seen in a dark room. Maximum values are used to determine upper limits (e.g., the loudest sound that can be listened to comfortably).

- **Motor thresholds.** These calibrations determine baselines and ranges in human motor ability, including motor skill (e.g., fine motor control or ambidexterity), muscle actuation (e.g., maximum bicep flex), and body measurements (e.g., maximum lung inflation).

- **Just-noticeable differences (JNDs).** These calibrations determine the minimum difference in the value of some attribute (e.g., color, frequency, pressure) such that two stimuli can be differentiated - for example, the minimum difference between two colors necessary for a user to perceive them as different.

- **Perceptual performance.** These calibrations determine a person’s abilities to carry out perceptual tasks such as reacting to a stimulus, visual search, or pattern match-
ing. The performance measures in these tasks are usually response time (e.g., time to find a visual target) and errors (e.g., patterns misidentified in a sequence).

- **Motor performance.** These calibrations determine the details of a person’s sensorimotor performance in several areas of HCI: aiming, pointing, steering, and pursuit tracking. The calibration is often used to establish coefficients for a performance model (e.g., Fitts’s Law), in order to accurately represent individual differences.

- **Memory performance.** These calibrations determine individual performance in memory tasks including short-term memory (e.g., recall of sets of items), spatial memory (e.g., remembering object locations), or the ability to associate items from one set to another. As with motor performance, calibration can be used to provide coefficients for existing models (e.g., Hick’s Law).

A second main class of calibration involves measurement of technology alone, and does not consider human abilities. A common calibration process in human-computer systems is **registration.** Registration is used to bring two spaces into alignment, where the congruence of the spaces is needed to enable subsystems to work together. Three examples of registration involve physical, color, and sound space:

- **Physical space registration.** These calibrations are used to bring two representations of physical space into alignment. A common example is that of registering input and output spaces for direct touch on a digital table.

- **Color space registration.** These calibrations are used to adjust the color output of a device so that it matches some representation of ‘ground truth.’ Monitor calibration by measuring the maximum output of each RGB channel is an example of color space registration.

- **Auditory registration.** These calibrations are used to match the audio output of different systems (e.g., an orchestra tuning each instrument to the oboe).

**Framework Part 2: Core Tasks of Calibration**

In moving from calibration to games, the main object of the calibration must be maintained – that is, the game must calibrate the same thing as the standard procedure. However, it is not required that the game use exactly the same presentation as the traditional system.

For example, a standard color JND calibration might present pairs of color swatches and ask the user whether they are the same or different; however, it is not essential that a game version use this same approach. The core task underlying the ‘same or different’ question is signal discrimination, and so any presentation that implements this core task should be possible for a game version of the calibration. In the color game described below, the signal discrimination task is carried out implicitly, in that the player can only shoot targets that they can see against the background.

Gamification thus requires that we identify tasks that are at the core of calibration procedures, and ensure that these tasks are the basis of game versions. We identify several of these core tasks below (a non-exhaustive list, but one that covers a wide range of possible calibration types):

- **Signal detection:** the minimum level of intensity that a person can perceive for a given stimulus (e.g., light).

- **Signal discrimination:** the minimum amount of difference needed for a person to determine that two stimuli (e.g., two colors, two tones) are different.

- **Maximum / minimum muscle activation:** the maximum and (non-zero) minimum amounts that a person can activate a muscle (e.g., flexing the bicep muscle).

- **Minimum movement:** the minimum amount that a body part can be moved, as a measure of fine motor control.

- **Ambidexterity:** the ability to perform coordinated simultaneous movements with left and right limbs.

- **Reaction time:** the time needed to initiate a response to a perceptual stimulus.

- **Visual search:** the ability to find a visual target in a field of distractors.

- **Pattern matching:** the ability to determine that a given perceptual pattern matches a target pattern.

- **Aiming:** the ability to accurately point a device (e.g., a rifle) at a target, without feedback about the direction.

- **Pointing:** the speed at which a person can move a pointer to a target (with feedback).

- **Steering:** the speed at which a person can move a pointer along a path without going outside the path’s borders.

- **Pursuit tracking:** the ability to move a pointer so as to accurately match the location of a moving object.

- **Short-term memory:** a person’s ability to memorize and retrieve sets of items, sequences, and mappings.

- **Spatial location memory:** ability to remember locations of items without persistent visual cues.

**Framework Part 3: Game Mechanics for Core Tasks**

Making games that are based on the core tasks described above can be accomplished because many games are already designed around exactly these building blocks. Figure 1 sets out the ways that core tasks of calibration can be translated into the kinds of game mechanics that are used in many existing games. For example, signal detection, which is a core task of perceptual threshold calibration, can be translated to a ‘presence of enemies’ game mechanic. Similarly, a core task like visual search, which underlies some perceptual performance calibrations, can be translated to a ‘find treasure’ game mechanic.

Although some calibration types use tasks that are currently rare in games, new game mechanics can be designed to fulfill these needs. For example, while some work has looked at using physiological sensors in games [14], the core tasks of muscle activation and minimum movement are not typically used as game mechanics. We explore this area further with one of our example games (BabyLaunch) that demonstrates how non-typical tasks can be gamified.

**Framework Part 4: Game Design Elements**

The final stage in our process is to add elements from game design that make the calibration procedure more game-like.
We include four basic elements of games, derived from previous taxonomies of game elements [7,13].

- **Challenge:** providing challenging goal elements tied to rewards. The system needs to provide clear task goals (e.g., collecting items, shooting objects) that trigger a challenging activity for the player to engage in. The challenge may increase within system boundaries as the game progresses, to keep player interest.

- **Theme:** Vicarious aesthetic representation and theme. The elements of the calibration system can be cloaked by putting them in a fantasy context (e.g., targets are spaceships, different parts of the game have descriptive names evoking some form of mental imagery), allowing the player to experience a vicarious setting. This applies to the player representation (e.g., the avatar) as well.

- **Reward:** Rewards and behavior reinforcement. Challenge alone does not provide sufficient motivation to stay engaged in a task, so activities need to be rewarded periodically to provide feedback about player progress and to maintain motivation. Easily-implemented rewards include simple visual animations or pleasant sounds. To trigger curiosity, rewards at random locations can be used to keep the player cognitively engaged.

- **Progress:** Progress units and markers. Feedback is a central aspect of game design. Providing progress units (e.g., levels, worlds, quests) and achievement markers (e.g., badges, score information) allows players to know how far they have progressed into the game and how well they are doing. Progress units can also serve as reinforcement (e.g., using time pressure to change the speed of an action). In a social context, progress markers such as high scores also enable reputation scores, fostering competition and increasing replay value.

Including these basic elements allows us to design a game scenario around any basic calibration mechanic and thereby increase the engagement and fun in an otherwise dull task. It is important to note that all of these elements need not be present to make a game, and that a single game feature may cover one or more of the elements above.

**THREE CALIBRATION GAMES**

We developed three games as examples of how calibration can be gamified, using three different types of calibration and three different core tasks.

**Calibrating Color Just-Noticeable-Differences (JNDs)**

The ability to differentiate color varies between individuals because of many factors, such as environmental lighting, age, fatigue, or the presence of color vision deficiency. To make the best color choices for information visualization, a system can use a model of the color differentiation abilities of the user. In order to generate this model, an in-situ calibration is performed that measures the individual’s ability to differentiate between colors [5].

The standard method for calibrating the model asks the user to perform a sequence of signal discrimination tasks. A rectangular field of circles is shown to the user (see Figure 2), with half of the circles one color and half another. The user indicates whether the colors are the same or different. About 10 trials (at one second each) are needed to measure a single differentiation limit, and the model requires 192 of these limits, for a total of 1920 seconds (32 minutes).

![Figure 2. The traditional calibration task. Users respond to each presentation of bicolored circles with 'same' or 'different'. Approximately 1920 of these presentations are used to calibrate the model.](image-url)

The standard calibration is both lengthy and dull; in addition, it may need to be repeated whenever the environment changes (e.g., lighting levels change). The cost of this procedure quickly outweighs the benefits for users, who are thus hesitant to perform the calibration as often as needed.

To produce a game-based calibration, we built a ‘space invaders’ style game, which is a 2D side-view shoot-em-up style game. In the original, the player controls a vehicle at the bottom of the screen from which missiles can be launched. The numerous computer opponents progress in formation from the top of the screen to the bottom. The goal of the game is to shoot every computer opponent with a missile before they reach the bottom of the screen.
The original game relies on the targets being differentiable from the background, and we modified this mechanic to use the ‘signal discrimination’ core task identified above. By altering the background and target colors, we replicated the ‘same or different’ task in the game. Instead of the user explicitly indicating whether the colors are the same or different, the game version relied on implicit indication. If the user can see a target (i.e., they see it as a different color than the background), then they shoot it with a missile. The change in presentation from the original calibration procedure meant that we also had to prevent people from differentiating colors based on luminance or movement; therefore, we used a random luminance mask (RLM) over the entire screen, which removes luminance contrast and reduces movement effects, thereby forcing users to differentiate solely based on color [18] (see Figure 3).

In addition, the implicit nature of the signal-discrimination task means that the game version must deal with errors differently. In the original system, there is no way to know whether the user is making an error in judgment, and so the procedure must use repeated trials to obtain a probability measure. In the game version, people may make errors that are based only on the game mechanics (i.e., shooting at a target and missing), and these errors are difficult to interpret in terms of the signal discrimination task. To reduce errors based on simple aiming (which is not our core task of interest), we quantized the horizontal dimension of the game – both the player’s vehicle and the invading enemies move discretely across the 20 fixed columns on the screen.

This helps prevent blanket shooting (which would result in shooting targets that the user cannot see).

- **High score (reward, progress, challenge).** By showing the current high score, the player is encouraged to achieve a higher score by shooting more targets and maintaining a higher accuracy.

- **New items for high accuracy (reward, challenge, theme).** When the player achieves a high accuracy on a level, they are rewarded with a new missile representation and an additional target explosion sound effect.

### Calibrating C:D Parameters for Targeting

Many activities in desktop applications involve targeting tasks, and many optimization processes seek to improve targeting by identifying a user's performance characteristics in selection and targeting activities - for example, determining the correct parameters for a model-based interaction technique. In this game we calibrate the control-to-display (C:D) ratio (the relationship between control device movement and on-screen cursor movement [12]).

User performance characteristics are generally determined through repeated-trials calibration sessions, in which the user carries out a number of test tasks. For selection and targeting, calibration typically involves repeated selection of two-dimensional targets on a plain background. In order to stabilize performance, users must often complete a large number of these trials (e.g., 50-100 trials per condition [12]). These repetitions can become tedious and tiring, and it is difficult for users to maintain performance at their peak level - which is an important requirement for some models.

To gamify this calibration, we identified targeting as the core task underlying the calibration, and translated that core task into a pointing-based game mechanic that is common in many computer-based shooting-gallery games (see Figure 1). The shooting game thus used the same presentation as the traditional calibration system; but to this basic platform we added several game-design elements:

- **Plot (theme).** We added a minimal story to the game; the universe is being attacked by ‘evil blorgs,’ and users must help to defend against the attack. In addition, the different C:D ratios were characterized as different ‘worlds’ where the cursor would behave differently (e.g., in ‘sludge world’ a lower C:D ratio was employed).

- **Visible timer (challenge).** The calibration session runs for ten seconds; in the game version of the system, the timer is made visible using a progress bar, encouraging users to shoot as many targets as possible before the bar fills up.

- **Laser gun and explosion effects (theme and reward).** In the game system, the cursor is represented by cross-hairs, and clicking the mouse causes a ‘laser beam’ effect (see Figure 4). Hitting a target causes it to break into pieces that fall to the bottom of the screen.

- **Sound effects (theme and reward).** The game plays different sound effects for firing, hitting and missing targets, bonuses, and the start and end of a ten-second trial.
• **Bonuses (theme, reward, and progress).** Two kinds of bonus were used (see Figure 4): first, if the user hit five targets in a row, they received a gold star; second, if the user hit more targets than they had in the previous round, the explosions changed (from broken pieces to bubbles). The changes made for this game were thus all cosmetic, and there was minimal change in the presentation of the core task. For testing, we designed both a standard calibration version as well as the shooting game. In both systems the goal was to find which of four C:D ratios was optimal for the user’s mouse-based input. Both systems tested four C:D ratios, with four 10-second sessions at each ratio.

The standard calibration procedure for chest inhalation sensors is fast (it only requires breathing in deeply a few times), so there is no problem with tedium in this case. We developed BabyLaunch to explore the effects that gamification can have on calibration data for very simple procedures. In particular, we were interested in the effects of rewards (the laughing baby, colorful visuals), challenge (to see how far you can launch the baby), theme (baby playing in the ball pit), and progress (current maximum distance) on people’s performance. In addition, we were interested in whether the game elements could detract from performance by distracting the user from the calibration task.

**EVALUATION**

Our testing of the calibration games involved two main questions: first, whether gamifying a calibration solves the problems identified previously (lack of enjoyment and lack of motivation); and second, whether the addition of game elements reduces the accuracy of the calibration data.

To answer these questions, we compared the gamified (CG) and traditional calibration procedures (SC) in three separate studies. For each study, we recruited 12 people from the local university community, and had them test the two calibration systems in counterbalanced order. We compared the calibration data using paired-samples t-tests to determine whether the game versions produced results that were different from the traditional versions, and we compared people’s answers to questions about their enjoyment and effort during the sessions using Wilcoxon Signed Ranks Tests.

**Results 1: Subjective Experiences**

We measured perceptions of difficulty, frustration, tedium, and enjoyment through a survey given after each session.

**Color perception calibration**

Survey responses showed three main results. First, people found the game version of the system significantly more enjoyable than the standard system (Table 1, rows 1 and 2). Second, people did not indicate much difference in terms of difficulty or frustration – neither of these survey questions resulted in significant differences (rows 3 and 4, Table 1).

![Figure 4. The C:D calibration game. The blue bar on the left marks time passed in the level, and the interface shows last and current hits as well as score. The star in the upper-right corner is a badge for successfully hitting five blorgs in a row.](image)

![Figure 5. The BabyLaunch game allows a user to launch a laughing baby out of a picnic basket into a pile of colorful balls, by breathing in deeply.](image)

Third, the gamification clearly had the effect of **increasing** the perception of effort: people felt that the game version forced them to work harder and rushed them more than the standard version (rows 5-8 of Table 1). This additional effort is not necessarily a problem, however: it could mean that people were trying harder (see discussion below).
**C:D-ratio calibration**

Survey responses for the C:D calibration showed similar trends in terms of fun and enjoyment – responses were significantly higher for the game versions (rows 1 and 2, Table 2). Given that the only difference between the two versions of the C:D system was the game elements, it seems that even simple decoration can have an effect on enjoyment (one person commented that “the lack of feedback in the non-game version made it less fun;” another said “it’s just more fun when you have a purpose behind it”).

Unlike the color calibration, there were no other subjective differences: people ranked the two versions similarly in terms of effort, difficulty, and frustration. The lack of differences here may have to do with the fact that the presentation of the core task was the same in both versions of the C:D game, unlike in the color calibration game.

Table 2. Survey responses for C:D-ratio calibration.

<table>
<thead>
<tr>
<th>Likert statement (5-point scale)</th>
<th>CG</th>
<th>SC.</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 I enjoyed completing the task</td>
<td>3.8</td>
<td>2.9</td>
<td>.040</td>
</tr>
<tr>
<td>2 Completing the task was fun</td>
<td>3.9</td>
<td>2.8</td>
<td>.004</td>
</tr>
<tr>
<td>3 Completing the task was difficult</td>
<td>2.3</td>
<td>1.8</td>
<td>.119</td>
</tr>
<tr>
<td>4 Completing the task was frustrating</td>
<td>2.0</td>
<td>1.7</td>
<td>.219</td>
</tr>
<tr>
<td>5 The task was physically demanding</td>
<td>2.0</td>
<td>2.3</td>
<td>.390</td>
</tr>
<tr>
<td>6 I felt rushed when completing the task</td>
<td>3.3</td>
<td>3.2</td>
<td>.527</td>
</tr>
<tr>
<td>7 I had to work hard to complete the task</td>
<td>2.1</td>
<td>2.0</td>
<td>.739</td>
</tr>
<tr>
<td>8 The task was mentally demanding</td>
<td>2.7</td>
<td>2.1</td>
<td>.161</td>
</tr>
</tbody>
</table>

For the C:D calibration, we also asked participants two additional questions – which system they would prefer if they had to do the calibration once per day, and which they would prefer if they had to do it ten times per day. For the ‘once per day’ preference, people were evenly divided between the game and the standard calibration. For the ‘ten times per day’ preference, however, only two people chose the standard system, and ten chose the game version.

**Respiration sensor calibration**

For the respiration sensor calibration, users again reported significantly higher enjoyment with the game-based version over the standard calibration (row 1, Table 3). However, unlike the other two calibrations, there was no significant difference in fun between the two calibration versions. This suggests that the visual and audio effects provided an entertainment value to the users (thereby increasing their enjoyment), but the input technique (breathing deeply) was too mundane to result in the participants having fun.

Table 3. Survey responses for respiration sensor.

<table>
<thead>
<tr>
<th>Likert statement (5-point scale)</th>
<th>CG</th>
<th>SC.</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 I enjoyed completing the task</td>
<td>4.2</td>
<td>3.3</td>
<td>.007</td>
</tr>
<tr>
<td>2 Completing the task was fun</td>
<td>4.2</td>
<td>3.6</td>
<td>.107</td>
</tr>
<tr>
<td>3 Completing the task was difficult</td>
<td>1.7</td>
<td>1.7</td>
<td>.000</td>
</tr>
<tr>
<td>4 Completing the task was frustrating</td>
<td>2.0</td>
<td>1.8</td>
<td>.776</td>
</tr>
<tr>
<td>5 The task was physically demanding</td>
<td>3.6</td>
<td>2.9</td>
<td>.131</td>
</tr>
<tr>
<td>6 I had to work hard to complete the task</td>
<td>2.3</td>
<td>1.8</td>
<td>.167</td>
</tr>
<tr>
<td>7 I felt rushed when completing the task</td>
<td>2.5</td>
<td>1.9</td>
<td>.230</td>
</tr>
<tr>
<td>8 The task was mentally demanding</td>
<td>2.3</td>
<td>1.9</td>
<td>.157</td>
</tr>
</tbody>
</table>

All other results showed no significant differences between the two versions of calibration. Similar to the C:D calibration, it may be that the core game mechanic of each version (breathing deeply) was so similar that there were no differences in effort, difficulty, or frustration.

**Results 2: Calibration Accuracy**

**Color perception calibration**

JNDS can be described mathematically by sigmoid functions. Normalized sigmoid functions transition from a minimum of 0 to a maximum of 1 as the x-axis value increases. For JNDS, this translates into moving from not differentiable to completely differentiable (y-axis) as you increase the difference between two signals (x-axis). In psychophysics, this is called a psychometric function. In Figure 6 six JND psychometric functions (3 RGB channels x 2 calibration techniques) are shown. Each of these functions is the average JND for all participants for a particular RGB channel and calibration type.

It is evident from Figure 6 that there are differences between the JND functions for the two types of calibration. We found a significant difference for all (p<0.05) but one (p=0.06) for the horizontal shift of each function. This suggests that the functions are relatively translated. We also examined the horizontal span of each function as it transitions from 0 to 1 and found no significant differences. This suggests that the functions are not relatively skewed. We will come back to these results in the discussion.

**C:D-ratio calibration**

The C:D calibration system is designed to determine people’s performance at different control-display ratios. Figure 7 shows the mean results from the two systems. Although the relative performance was similar, completion times per target were lower at all ratios for the game system, and this difference was significant (p<0.05).

Our observations suggest that the improved performance may have been caused by people trying harder in the game: that is, it is possible that the game version of the calibration actually produces more accurate data, since it did a better job of encouraging people to try their hardest at the tasks.
We were also interested in whether the game elements would change people’s behaviour in ways that would lead to a less accurate calibration. We looked at three issues: selection errors, the gold stars (given after a run of five error-free trials), and the bubble bonus (given if the player beat their previous score). We analysed the actual data, and also asked people if they changed their strategy for any of these elements. There were no significant differences in the number of errors, the number of error-free runs, or the frequency of beating previous scores. In addition, people’s responses indicated that only one person made a conscious effort to work toward the stars and the bonus.

**Respiration sensor calibration**

In contrast to the color and C:D calibrations, no significant difference was found between the maximum respiration reading between the two versions of respiration sensor calibration (p>0.05). This may suggest that the lack of a specific reward for higher achievement (the game response was identical no matter how far the baby was launched) caused the participant to not try harder during the game version. It is also possible that the brevity of each calibration version (under one minute each) prevented the user from becoming immersed in the game environment.

**DISCUSSION**

Our examples and comparisons show five main findings:

• calibration procedures are good candidates for gamification, because they often share core tasks with games;
• calibration games can be successfully built either using the same presentation as the original procedure (the C:D game), or a very different presentation (the color game);
• calibration games were clearly seen as more enjoyable than standard versions, and were strongly preferred;
• some of the game versions did produce significantly different data than the standard calibrations;
• gamification can also change other qualities of the user’s experience in addition to enjoyment; the pacing and challenge elements of game versions may lead to experiences that are seen as more effortful and more demanding.

In the following sections we explore several questions including issues of generalizability, feasibility, and cost.

**Do game versions solve calibration’s motivation problems?**

Our evidence suggests that game-based calibrations can solve both motivation problems of standard calibration procedures. The first problem was that people are unlikely to carry out standard calibrations at all, because the procedures are often tedious and unenjoyable. Our studies clearly showed that gamification makes these procedures significantly more fun and more enjoyable.

It is not surprising that people find games more fun; however, it is important to remember that our ‘games’ were really just dressed-up versions of standard calibration tasks, and so the significant increase in enjoyability represents a confirmation that we can in fact introduce some of the pleasurable experiences of games to tasks like calibration, even without dramatically changing the nature of the task.

The second motivation problem is that people are unmotivated to perform at the level that is needed for many kinds of calibrations. Our evaluations also suggest that the addition of encouragement and reward structures to the calibration procedure can motivate people to try harder.

This conclusion is not unequivocal, since our evidence only shows that people performed differently in the game versions, not why those differences occurred. The changes could also have arisen because of differences in the presentation of the stimuli, or interaction effects with other game elements. However, our observations and comments from participants suggest that there is a motivating effect of the game’s rewards and encouragements.

**Is game data still usable for calibration?**

The risk in changing a calibration procedure is that the data gathered from the game will be inaccurate or biased. Our evaluations show that the data from the game versions is different from that of the standard versions – significantly so – but this difference itself does not mean that the game data is less valuable for calibration.

A problem with all human-based calibrations is that there is no true ‘ground truth’, and so any calibration procedure is an approximation of the real value – and it may be that the game versions actually provide a more accurate calibration (e.g., by motivating people to try harder). More study is needed to determine why our game-based calibrations differed from the standard versions, but here we consider the potential ramifications of the differences we saw.

If we assume that the standard calibration is correct, we can consider how the game-data differences would affect real-world decisions that might be made from the calibration. In the case of using the C:D calibration to choose the best ratio, the game data would have led us to choose a ratio of 1.0 more often than we should (i.e., since 1.0 was the best ratio for the game, whereas 2.0 was best in the standard system). This difference, however, represents a performance difference of only 25 milliseconds per selection (i.e., the performance difference between a C:D ratio of 1 and 2 for either system); therefore, the results of calibrating with the game version are nearly indistinguishable from that of the standard version. In addition, if we are comparing the game-based calibration to situations where we have no calibration at all (due to a lack of motivation for doing the process), then the game version is a clearly better approach.

**Figure 7. Mean selection time by C:D ratio.**
This example does raise a related issue, however, that could become a problem for game-based calibration in some settings. If games motivate people to try harder, it is possible that their better performance is dependent on a higher level of effort. In some cases this is not a problem (e.g., when determining relative coefficients); but for other purposes (e.g., choosing the best value to use in a real application), a decision based on game data could lead to a choice that does not accurately reflect the user’s mental state or effort level in the situation of use. The generalizability of a calibration task to the real-world situation of interest is an ongoing problem for all human-based calibration, but gamification adds additional issues for designers to consider.

In the case of color calibration, the game version appeared to be more sensitive than the standard calibration. This difference may have arisen through improved motivation, but may also be caused by the difference in presentation (e.g., the addition of random noise in the game may not have removed all perceptual traces of luminance or movement). If the game-based calibration is in fact over-sensitive, it would lead to more false positives in the differentiation predictor (i.e., the predictor would incorrectly state that the user will be able to differentiate some pairs of colors). Again, however, the real-world difference is extremely small and easily solved by using a small fixed-offset value, as proposed in earlier work [5]. This fixed-offset value translates the psychometric functions horizontally, thereby aligning the color-based calibration with the standard calibration.

The color-calibration game shows that changing the presentation of the core calibration task requires an additional step of the system designer – that is, the designer must ensure that any changes in the game version do not make the task different (e.g., easier or harder). This can be accomplished by comparing the standard and new presentations in a non-game context, to control for motivational effects.

It is also worth noting that players of calibration games are likely to be aware of the fact that a calibration process is going on behind the game. The knowledge that there is a larger purpose that benefits the user themselves may help to control any behaviour that would reduce the value of the calibration data. For example, it is unlikely that a user will cheat in order to win the game, when they know that the outcomes of the game are decisions that will affect their performance or satisfaction with the system.

What other calibrations could be gamified?

Figure 1 shows that many calibration types have connections to common game mechanics. The only types listed that do not have strong connections to games are some motor-control thresholds (i.e., minimum movement). As further examples of the range of systems that are possible, we briefly introduce three additional places where calibration could be carried out using gamified versions.

Determining spatial memory abilities. Some interaction techniques work best for users who have good spatial location memory, and calibration can be carried out to assess each person’s individual capability. Playing an interactive game based on Concentration (where the user must remember locations in order to match pairs) could provide this assessment; in addition, the presentation could be tuned to the stimuli that are to be used in the interaction technique.

Registering input and output spaces on a table. Projector-based table systems involve representations of space for the input device (e.g., vision-based tracking) and the output (e.g., a projector). These spaces must be aligned, but with some setups a large number of registration points are required (e.g., when nearby metal warps the input space of a magnetic tracker). A targeting or pursuit-tracking game could provide a large number of registration points, even with users such as schoolchildren who are unlikely to carry out traditional calibration. The main design issue in this game would be in determining when the user’s input actions were actually on the target (similar to the error issue discussed for the JND game); this could be addressed by making use of varying difficulty levels (e.g., a slower target in a pursuit task is more likely to result in accurate data).

Calibrating a color monitor. Monitor calibration is a task that is rarely if ever carried out by most users, but a properly-calibrated monitor can be extremely useful (e.g., for producing accurate printouts). A game based on visual search and pattern matching (e.g., similar to popular games like Bejeweled) could be used to make monitor calibration less tedious. Printed color swatches could be used as physical tokens in the game, and the player’s goal would be to find colors on the screen that match the printed swatches. A staircase approach could be used to deal with errors and home-in on an accurate calibration.

What are the limits to calibration games?

Although calibration games can be widely applied, there are limits to the idea:

- The quality of the calibration data must be the first priority. Although we concluded that the game elements did not compromise the quality of the calibration in our three examples, there are clearly limits to the number and type of game elements that can be added to the system. For example, adding narrative elements or reward structures that dramatically change people’s strategy in the game could alter the way they behave, compromising the calibration. Similarly, adding challenge elements to the game could add noise to the data, making it more difficult to interpret the user’s behaviour when errors occur.
- If an existing standard calibration procedure is not tedious (dull and time-consuming), then there is little need to convert the procedure to a game. For example, ergonomic measurements (e.g., needed to set up a vision-based gestural input system) would only need to be taken once, and would be easy to carry out.
- Games require that the user switch mental contexts, even minimally, to think about the goals and tasks of the game (e.g., the ‘invading blorgs’ story of the C:D game). In some task situations, this context shift could disrupt a us-
er’s work focus (e.g., it could cause them to forget important information for a post-calibration task step).

- Some work settings (e.g., formal offices) may not be appropriate places to play games, and the sound and visual effects of games may disrupt other workers.

**What are the added costs of building a game version?**

Development costs vary, but in our experience, the added development costs of gamifying calibration are relatively low. All three game implementations required approximately three additional days of design and programming compared to the traditional versions. We estimate that games using the same core task as the original calibration could be developed in as little as two days, although some amount of additional playtesting time will be required to ensure that the game elements work and that the game version actually is more fun than the standard version.

For systems that change the presentation of the core task, development time will necessarily be longer, particularly if testing must be carried out to ensure that the new presentation is providing accurate data. As the idea of game-based calibration sees wider adoption, however, some of these development costs will be reduced through re-use.

**CONCLUSIONS AND FUTURE WORK**

Calibration is an important part of many human-computer systems, but is often neglected because calibration procedures are tedious and uninteresting. To address this problem, we introduced the idea of calibration games. We presented a design framework that can aid in the development of these games, and demonstrated the process through three example games. Our evaluations showed that the game versions of calibration were much more enjoyable than the standard versions, and were strongly preferred. In addition, although there were differences in the data gathered from the two versions, these differences do not compromise the utility of calibration data obtained from games.

There are several avenues for continued work in this area. First, we plan to develop more calibration games to further test and refine our design framework. Second, we will undertake a more in-depth exploration of the way that data differences actually affect different kinds of decisions, and further investigate the issue of errors raised in the discussion. Third, we plan to package and deploy a set of calibration games, in order to gather additional information about the use of these games in real usage scenarios.

**REFERENCES**