Behav Res (2013) 45:301–318
DOI 10.3758/s13428-012-0264-3

# Gamelike features might not improve data

Guy E. Hawkins · Babette Rae · Keith V. Nesbitt · Scott D. Brown

Published online: 9 October 2012

**Abstract** Many psychological experiments require participants to complete lots of trials in a monotonous task, which often induces boredom. An increasingly popular approach to alleviate such boredom is to incorporate gamelike features into standard experimental tasks. Games are assumed to be interesting and, hence, motivating, and better motivated participants might produce better data (with fewer lapses in attention and greater accuracy). Despite its apparent prevalence, the assumption that gamelike features improve data is almost completely untested. We test this assumption by presenting a choice task and a change detection task in both gamelike and standard forms. Response latency, accuracy, and overall task performance were unchanged by gamelike features in both experiments. We present a novel cognitive model for the choice task, based on particle filtering, to decorrelate the dependent variables and measure performance in a more psychologically meaningful manner. The model-based analyses are consistent with the hypothesis that gamelike features did not alter cognition. A postexperimental questionnaire indicated that the gamelike version provided a more positive and enjoyable experience for participants than the standard task, even though this subjective experience did not translate into data effects. Although our results hold only for the two experiments examined, the gamelike features we incorporated into both tasks were typical of—and at least as salient and interesting as those usually used by—experimental psychologists. Our results suggest that modifying an experiment to include gamelike features, while leaving the basic task unchanged, may not improve the quality of the data collected, but it may provide participants with a better experimental experience.

**Keywords** Computer game · Gamelike features · Response time · Multialternative choice · Change detection · Speed-accuracy trade-off · Particle filter

## Introduction

The potential for computer games in various streams of psychological research is substantial. In applied contexts, computer games and simulations have been used to promote learning in such varied areas as the military, business, health care, and aviation (for reviews, see Salas & Cannon-Bowers, 2001; Wilson et al., 2009). Computer games have also been increasingly utilized to explore human performance in the laboratory. In particular, games and tasks incorporating the visual, auditory, and/or game play aspects of computer games (i.e., gamelike features) present ways of complementing standard methodological approaches in experimental psychology.

Psychology experiments often require participants to complete many trials in a monotonous task. Such boredom might reduce the engagement of participants, leading to poorer task performance and, perhaps, also more variable data (assuming that the effects of boredom are different for different participants or at different times throughout the experiment). Since many people find computer games engaging and motivating (Wood, Griffiths, Chappell, & Davies, 2004), it might be possible to improve participants' engagement in an experiment by incorporating some gamelike features. Recent decades have seen this approach taken in many areas of psychology, and using many different gamelike features. In some cases, the gamelike features are essential elements of the central task of the participant. For example, testing children requires special consideration, and experimental tasks must be specially designed to suit younger

G. E. Hawkins (✉) · B. Rae · S. D. Brown
School of Psychology, University of Newcastle,
Callaghan NSW 2308, Australia
e-mail: guy.e.hawkins@gmail.com

K. V. Nesbitt
School of Design, Communication & Information Technology,
University of Newcastle,
Callaghan NSW 2308, Australia

minds, most often by having them play games. There are many quite sophisticated examples of gamelike experiments for children, including "Dragon Master" (Metcalfe, Kornell, & Finn, 2009), "Frog Game" (Dunbar, Hill, & Lewis, 2001), and spaceships (Spencer & Hund, 2002, 2003; for more examples, see Berger, Jones, Rothbart, & Posner, 2000; Droit-Volet, Tourret, & Wearden, 2004; Kujala, Richardson, & Lyytinen, 2010; Ploog, Banerjee, & Brooks, 2009; Stevenson, Sundqvist, & Mahmut, 2007; Yildirim, Narayanan, & Potamianos, 2011). Other researchers working with children have simply presented an unaltered experimental task to the child as if it were a game. For example, "We are going to play a game now. Would you like to play a game?" (see, e.g., Andrews & Halford, 2002; Carneiro, Fernandez, & Dias, 2009; Hanauer & Brooks, 2003; Price & Connolly, 2006; Raijmakers, Dolan, & Molenaar, 2001; Thibaut, French, & Vezneva, 2010; Toppino, Fearnow-Kenney, Kiepert, & Teremula, 2009; Yuzawa, 2001).

Another situation in which the game is fundamental is when existing video games (such as Tetris and Madden) are used to explore cognition (Hansberger, Schunn, & Holt, 2006; Kirsh & Maglio, 1994; Maglio, Wenger, & Copeland, 2008). Some researchers extend this approach by using the development engines of various games to create gamelike environments for the exploration of spatial cognition and social behavior (Alloway, Corley, & Ramscar, 2006; Drury et al., 2009; Frey, Hartig, Ketzel, Zinkernagel, & Moosbrugger, 2007; Gunzelmann & Anderson, 2006; Hutcheson & Wedell, 2009; Radvansky & Copeland, 2006). This type of experiment is less common in cognitive psychology than other gamelike experiments. A much more popular way of using gamelike features in cognitive psychology experiments is to modify an experiment's appearance without changing the fundamental properties of the stimuli or the experimental design and procedure. Colors, animations, and sound effects may be added to the standard display. Gamelike features can also be introduced through the creation of a back story and a performance-based point system.

We refer to the introduction of cosmetic modifications, which do not change the fundamental stimuli, design, or procedure, as "gaming up" an experiment. One assumption made when gaming up an experiment is that the gamelike additions improve the participants' experience. This assumption is uncontroversial, because gamelike features are almost certainly more interesting and fun than the plain features of standard experiments. A less certain, and often implicit, assumption about participants' underlying cognitions is that the improved experience of participants will manifest as greater engagement with the task and better motivation. In turn, it is assumed that this greater engagement and motivation will manifest in the data as improved performance or, perhaps, reduced between-subject variability that may arise due to variable boredom or motivation between subjects. While this assumption appears to be endorsed (at least implicitly) among experimental psychologists, conventional wisdom in computer game design suggests the opposite. One maxim of computer game design is that the fundamental repeated task determines enjoyment level, not the superficial gamelike features (e.g., Schonfeld, 2010); if the underlying task is not interesting and engaging, as in the majority of psychological experiments, then no amount of gamelike features can improve performance. We investigate these two competing notions by examining whether gamelike features increase the engagement and motivation of participants and, therefore, improve performance on experimental psychology tasks.

## Gaming-up experiments

Gamed-up experiments have become popular in almost all areas of experimental psychology. We briefly review some examples in order to give an idea of how prevalent this practice is and also to illustrate the typical kinds of gamelike features that are introduced. A particularly common application of gamelike features has been in learning experiments. For instance, participants learn to navigate through a three-dimensional space while interacting with "characters" later used for identification tests (Wade & Holt, 2005). Arcediano, Ortega, and Matute (1996) developed the "Martians" game to explore classical conditioning using Martians and explosions as stimuli (see also Baeyens et al., 2005; Blanco, Matute, & Vadillo, 2010; Franssen, Clarysse, Beckers, van Vooren, & Baeyens, 2010). Gamelike tasks have been used to study instrumental learning with stimuli presented as balloons that must be shot from the sky (Krageloh, Zapanta, Shepherd, & Landon, 2010), minefields to be navigated (Baker, Mercier, Vallee-Tourangeau, Frank, & Pan, 1993), or a host of similar examples (Lie, Harper, & Hunt, 2009; Molet, Jozefowiez, & Miller, 2010; Paredes-Olay, Abad, Gamez, & Rosas, 2002; Stokes & Balsam, 2001; Stokes & Harrison, 2002). Discrimination and generalization learning have been presented as melodies that participants must classify as belonging to different composers (Artigas, Chamizio, & Peris, 2001) or as torpedoes to be launched at certain flying objects but not others (Nelson & Sanjuan, 2008; Nelson, Sanjuan, Vadillo-Ruiz, & Perez, 2011). The game approach has also been extended to spatial learning, such as remembering the location of a previously displayed spaceship (Spencer & Hund, 2002), and the popular approach of creating a three-dimensional town where participants play the role of a taxi driver and must learn landmark, spatial, and temporal relations that are later tested for recall (Newman et al., 2007), sometimes with neurophysiological recordings (Ekstrom & Bookheimer, 2007; Weidemann, Mollison, & Kahana, 2009).

Causal reasoning and categorization experiments have also been gamed up. For instance, explicit categorization tasks can be presented as "diagnoses" (Castro & Wasserman, 2007;

Wasserman & Castro, 2005), and implicit categorization tasks can be presented as the detection of "secret code words" embedded in artificial grammars (Sallas, Mathews, Lane, & Sun, 2007). Causal reasoning has been presented as a scientist uncovering the workings of a "black box" with light rays and atoms (Johnson & Krems, 2001), or using electrical circuits (Johnson & Mayer, 2010), or many other back stories (Dixon & Banghert, 2004; Dixon & Dohn, 2003; Ozubko & Joordens, 2008; Stephen, Boncoddo, Magnuson, & Dixon, 2009). The detection and prediction of change has been investigated in a "tomato processing factory" (Brown & Steyvers, 2009). Experiments investigating meta-cognition and executive function have been variously presented to participants as driving simulations or spaceship wars (Finke, Lenhardt, & Ritter, 2009; Lien, Ruthruff, Remington, & Johnston, 2005; Mather, Gorlick, & Lighthall, 2009). Minimally gamed-up experiments in the same vein have used performance based points systems (e.g., Buchner, Mehl, Rothermund, & Wentura, 2006; van der Linden & Eling, 2006), and simple gamelike reward structures have been utilized in studies of comparative psychology (e.g., Artigas et al., 2001; Washburn & Gulledge, 1995). Social psychologists also have gamed up their experiments—for example, to study ostracism (Williams & Jarvis, 2006), conflict and cooperation strategies (Aidman & Shmelyov, 2002), attitude generalization (Fazio, Eiser, & Shook, 2004), and even alcohol consumption during drinking games (such as "beer pong"; Correia & Cameron, 2010).

The longest running approach to using games in psychological experimentation has been the application of arcade style games to assess skill acquisition, usually by having the participant pilot a spaceship (McPherson & Burns, 2007, 2008), shoot alien spaceships (Williams, Nesbitt, Eidels, & Elliott, 2011), or control space weapons (Jackson, Vernon, & Jackson, 1993; Salthouse & Prill, 1983; Talvitie & Singh, 2009). The longest running game specifically designed for experimental psychology research has been "Space Fortress" (Mane & Donchin, 1989). Space Fortress includes many gamelike qualities, including sound effects, visual explosions, and a performance-contingent points system. Many studies have utilized Space Fortress—for example, to study IQ (Rabbitt, Banerji, & Szymanski, 1989), skill acquisition (Arthur et al., 1995; Logie, Baddeley, Mane, Donchin, & Sheptak, 1989), and different training schedules and methods (Day, Arthur, & Shebilske, 1997; Fabiani, Buckley, Gratton, Coles, & Donchin, 1989; Mane, Adams, & Donchin, 1989; Shebilske, Goettl, Corrington, & Day, 1999).

### Does it help?

Although it is not usually stated explicitly, one goal of including gamelike features in an otherwise standard experiment is to improve data quality by increasing participant motivation. We could find very little research testing this assumption, and the one study we did find provided equivocal results: Washburn (2003) demonstrated that providing a back story to an otherwise standard task resulted in poorer accuracy but faster response times.

## Experiment 1

We aimed to directly test the assumption that gamelike features can improve data quality by randomly assigning participants to either a gamelike or a standard version of a simple cognitive experiment. Our task required participants to make judgments about a number of choice alternatives (displayed as squares) that dynamically collected dots over the course of a trial. On each trial, one square would accumulate dots slightly faster than the remaining squares, and the participants' goal was to select this target square as quickly and accurately as possible (a demonstration version of the standard task can be viewed online at http://psych.newcastle.edu.au/~sdb231/buckets/vanillaR.html). To ensure that our gamelike condition had the best chance of improving data quality relative to the standard condition, we included all of the gamelike elements that have been standardly incorporated in the experiments reviewed above: a detailed back story, animations related to stimulus presentation and also response feedback, audio feedback on responses, and a points-based scoring system.

## Method

### Participants

Two hundred first-year psychology students from the University of Newcastle participated online for course credit. Participants were randomly allocated to the gamelike or nongame version of the experiment. The gamelike and nongame versions of the task were statistically identical, with the only difference being the "gamed-up" appearance of the gamelike task.

### Properties common to gamelike and nongame versions of the task

In both tasks, decision latency and accuracy were measured as functions of the number of choice alternatives present in a display, which we denote with *K*. Each decision trial began with *K* empty squares randomly placed into 20 locations on a 4 × 5 grid, with each square measuring 100 pixels × 100 pixels (plus a 2-pixel border). A difficulty factor was introduced to the design where participants were randomly allocated to one of three levels defined by the number of squares displayed on any trial: an *easy* condition, where the number of squares displayed on any trial was randomly chosen from $K \in \{2, 4, 6, 8, 10\}$; a *medium* condition, where the number of squares displayed on any trial was randomly chosen from $K \in \{2, 4, 6, 8, 10, 12, 14, 16, 18\}$; and a *hard* condition, where the number of squares displayed on any trial was randomly chosen from $K \in \{10, 12, 14, 16, 18\}$. The difficulty factor was designed to examine behavior across a greater range of experimental conditions and, hence, give the gamelike features the best chance of improving data, without interfering with the critical gamelike versus nongame comparison. The three difficulty levels were crossed with the game version factor to create a 2 (game version: gamelike vs. nongame) × 3 (difficulty level: easy vs. medium vs. hard) between-subjects factorial design.

Over the course of a trial, each square dynamically accumulated small "dots." One square (chosen at random on each trial) accumulated dots slightly more rapidly than all the others, and it was the participant's task to identify this *target* square from *distractor* squares. During each trial, time proceeded in discrete steps of 15 events per second. On each step, each square either accumulated a new dot or not. The chance of each square accumulating a new dot was independent of all the other squares at .4, except the target square, which had a higher probability than all the others, at .5. This means that, on average, the target square accumulated 7.5 dots per second, while distractor squares each collected an average of 6 dots every second. The left panel of Fig. 1 shows an example of a trial with six choice alternatives in the nongame condition.

Squares began with a completely white background (unfilled), and each time a new dot was accumulated, a 2 × 2 pixel area within the square changed to a dark blue color. The position of the new dot was chosen randomly from the remaining unfilled area of the square. Participants were free to sample information from the environment until they felt confident with their decision. Nongame participants were simply informed that they should aim to identify the target as quickly and accurately as possible, but if they responded too early, they might incorrectly select a distractor square that had by chance accumulated the most dots thus far in the trial.

After the participant chose a target square, a very fast animation illustrated many more time steps very quickly. This provided feedback on whether the participant's choice was the true target (which always ended up accumulating more dots than did the other squares) or not. When the target square was correctly identified, its border was turned green. If an incorrect selection was made, the incorrect square's border was turned red, and the true target square's border was turned green.

## Properties specific to the gamelike version of the task

The gamelike task operated on the same principles as the nongame task, except with a more interesting "gamed-up" facade. We tried to make the gamelike features of this task at least as salient and interesting as those included in the experiments reviewed above, to maximize any game-driven effects in the data. The gamelike features did not influence the physical properties of the stimulus display, with stimuli appearing on screen in the same color, size, shape, and location as in the nongame task.

The nongame task began with a single, plain text instruction screen. In contrast, although the gamelike task instructions conveyed the same message as the nongame ones, this information was displayed through a detailed series of screens

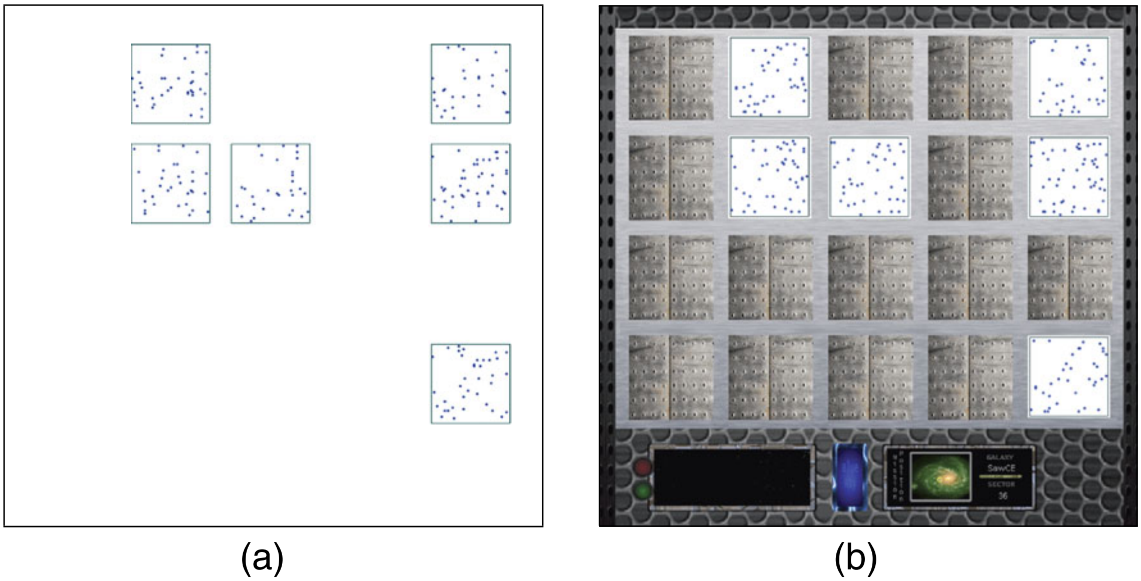


**Fig. 1** Illustrative example of the display in Experiment 1 during a typical trial for the nongame task (**a**) and the gamelike task (**b**)

to create a gamelike environment (a demonstration version of the gamelike task can be viewed online at http://psych.newcastle.edu.au/~sdb231/buckets/emf.html).

At the beginning of the gamelike task, participants were introduced to "EMFants: Last Light," a simple game set in a space environment. Participants were provided with a back story describing the EMFants species, which eat all forms of electromagnetic radiation (ElectroMagnetic Feeding ants—i.e., EMFants), that have escaped from a twin universe. EMFants have been detected in numerous galaxies and must be stopped before they rapidly spread to all known galaxies.

Participants viewed a mission brief instructing them that they were commander of "Dark-Stealth-6," a spaceship with a "shadow-scope" to detect alien EMFants, "blue-ray" armament to destroy EMFant colonies, and "time-hop" propulsion. Various aspects of standard experimental tasks were augmented to be more akin to a game. For example, when beginning a new block of trials, participants were required to manually engage Dark-Stealth-6's time-hop capabilities to navigate from one galaxy to another, initiating a short animation and sound effect. Observing the standard evidence accumulation of the nongame task was also made more interesting in the gamelike task. Participants were told that they used their shadow-scope to detect EMFant colonies (i.e., squares, the choice alternatives). The EMFant colony growing at the fastest rate indicated the home of the EMFant queen (the target square; see Fig. 1 for comparison of the nongame and gamelike displays during a trial). By clicking a target, participants fired their blue-ray, described as an intense pulse of long-wavelength radiation, to destroy the EMFant colony. When a square was selected, the entire display quickly flashed blue as the blue-ray fired, followed by an outline of green (for a correct answer) or red (for an incorrect answer) on the selected EMFant colony. A correct answer was accompanied by the sound of a cheering crowd. An incorrect answer produced a disappointed "urrrgghh" sound. Importantly, the statistical properties and the physical appearance of the stimuli were identical in the gamelike and nongame conditions.

The goal of the game was identified as locating and destroying EMFants. Participants were informed that speed was essential to prevent EMFants spreading to other galaxies. Participants were also instructed that accuracy was essential, since they had only one chance in each mission to fire the blue-ray, and if they did not destroy the colony of the queen, the EMFants would multiply and invade other galaxies.

Trials in the gamelike task took longer than nongame trials due to untimed events such as charging the blue-ray, postshot animations, sound effects, and so on. Accordingly, to match total experiment time, participants in the gamelike conditions completed fewer trials: 140 (game) versus 180 (nongame) in the easy condition; 126 versus 162 in the medium condition; and 100 versus 140 in the hard condition. Each $K$ appeared equally often in each block for all conditions.

## Results

Due to the different number of total trials across the six conditions, each $K$ was completed a different number of times by each group. To balance trials per $K$ between conditions, only the first 14 trials per $K$ for each participant were analyzed.[1] Data from participants with fewer than 33 % correct responses or whose computers displayed fewer than an average of 13 time steps per second were excluded, leaving data from 80 gamelike and 86 nongame participants for analysis. The proportion of participants excluded from analysis due to low accuracy rates did not differ between gamelike and nongame conditions, $\chi^2_{(1)} = 0.08$, $p = .77$. Remaining data were screened for outlying trials with responses faster than 1 s or slower than 150 s or individual trials where the host computer displayed fewer than 13 time steps per second, which were removed from analysis, resulting in the exclusion of data from 369 trials (2.46 % of total trials).

The upper panel of Fig. 2 shows mean response time for the nongame and gamelike conditions represented as a function of difficulty level. As was expected, a 2 (game version: gamelike vs. nongame) × 3 (difficulty level: easy vs. medium vs. hard) between-subjects analysis of variance indicated a highly significant main effect of difficulty level on response latency, $F(2, 160) = 17.04$, $p < .001$, where the fastest responses were provided by the easy groups, followed by the medium and then hard conditions. Of key interest, however, there was no main effect of game version or interaction between difficulty level and game version on response time, $p = .98$ and $p = .28$, respectively.

The lower panel of Fig. 2 displays mean accuracy for nongame and gamelike conditions as a function of difficulty level. Unlike the response time data, there was no significant main effect of difficulty level on accuracy, $p = .16$, and neither the game version or the interaction between difficulty level and game version effect was statistically reliable, $p = .56$ and $p = .49$, respectively.

### Do gamelike features have any effect on data?

Our analyses identified no significant differences between gamelike and standard versions of the experiment, but this does not imply that data from the two conditions were identically distributed. That is, the tests reported above do

[1] Fourteen trials per $K$ was chosen, since this was the lowest number of trials per $K$ completed by one of the conditions (gamelike medium difficulty). A similar pattern of results was obtained when analyses were conducted using all data.

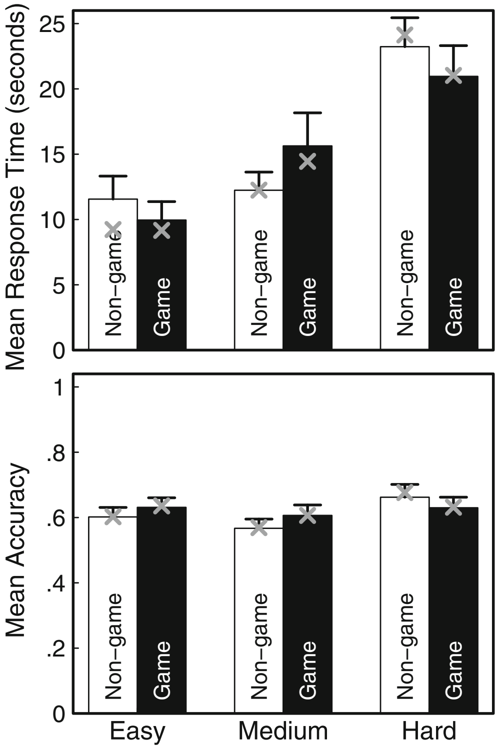


**Fig. 2** Experiment 1 mean response times (upper panel) and accuracy (lower panel) for nongame (white bars) and gamelike (black bars) conditions represented as a function of difficulty level. The error bars represent ±1 standard errors of the means. Gray crosses show predictions of the particle filter model (described in the main text and Appendix)

not confirm the null hypothesis, because they do not differentiate between null effects and low statistical power. We addressed this problem using the Bayesian *t*-test developed by Rouder, Speckman, Sun, Morey, and Iverson (2009). Rouder et al.'s method estimates a Bayes factor for the comparison of the null hypothesis and a well-specified alternative. We conducted two-sample Bayesian *t*-tests, assuming the uninformative JZS prior,[2] to compare the gamelike and nongame conditions within each difficulty level, for both mean response time and accuracy data. Table 2 shows the Bayes factors from each test, where values greater than one indicate support for the null hypothesis, and following the conventions described in Table 1. In all comparisons, the direction of support was in favor of the null hypothesis. Although the evidence did not always strongly favor the null, there is no evidence in favor of the alternative hypothesis—that is, that gamelike features alter measurable aspects of performance.

**Table 1** Interpretation of Bayes factors (*BF*; adapted from Raftery, 1995)

| Null hypothesis ($H_0$) | Strength of evidence | Alternative hypothesis ($H_A$) |
|---|---|---|
| $1 \le BF \le 3$ | Slight | $.33 \le BF \le 1$ |
| $3 \le BF \le 10$ | Positive | $.1 \le BF \le .33$ |
| $10 \le BF \le 100$ | Strong | $.01 \le BF \le .1$ |
| $BF > 100$ | Decisive | $BF < .01$ |

Stronger evidence for the null is obtained when a combined Bayes factor was considered. There are numerous possible approaches to aggregate Bayes factors across conditions or experiments. For example, one could take the product of independent Bayes factors, but this method does not respect the resolution of the data (i.e., the fact that there is increasing sample size with increasing number of conditions; Rouder & Morey, 2011). Instead, we use the Bayesian meta-analysis approach of Rouder and Morey, conducted separately on response time and accuracy data. This analysis assumes that each comparison (e.g., gamelike easy condition vs. nongame easy condition) and the samples of participants on which they are based are independent, which was true in our design. These meta-analytic Bayes factors indicated strong evidence in favor of the null hypothesis (no effect of gamelike features) for both response time and accuracy, as indicated by the Overall row in Table 2.

## Decorrelating response time and accuracy data to assess effort and riskiness

Our analyses of the Experiment 1 data indicate that adding gamelike features to an experimental task does not alter performance to a measurably noticeable level, even across a range of difficulty settings. However, the previous analyses assumed that response time and accuracy are independent, which is almost never true in data. For instance, the absence of a main effect of difficulty level on accuracy illustrates a common pitfall in analyzing accuracy and response time data separately, since one would expect harder difficulty levels to produce higher error rates. To maintain an acceptable accuracy rate, participants instead made increasingly slower responses as difficulty level increased

**Table 2** Bayes factors for response time and accuracy data in Experiment 1

| Difficulty level | Response time | Accuracy |
|---|---|---|
| Easy | 3.94 | 3.94 |
| Medium | 2.75 | 3.48 |
| Hard | 3.83 | 3.99 |
| Overall | 14.95 | 24.80 |

[2] We report only JZS Bayes factors, but note that the direction of all reported Bayes factors did not change when a unit-information prior was assumed.

(which manifested as a main effect of difficulty on choice latency). Indeed, the correlation between response time and accuracy was very strong for all difficulty levels, all *r*s > .8. The dependence of response time and accuracy is clear in scatterplots shown in the left column of Fig. 3. The three scatterplots show individual participant mean response time and accuracy in each of the easy, medium, and hard groups. In each scatterplot, the shape of the response-time–accuracy relationship is typical of the speed–accuracy trade-off common in choice experiments. Some people make fast and error-prone responses, seen in the lower left regions of the plots in Fig. 3, and others make slow and careful decisions, as in the upper right of each panel.

The strong correlation between accuracy and response time weakens the results described above, because our finding of a null result in one measure (say, response time) makes it almost unnecessary to even check the other measure (accuracy), because a corresponding finding is almost certain to be observed. To decorrelate response time and accuracy, we apply a process model based on particle filtering to the data. Particle filters are sequential Monte Carlo methods that approximate Bayesian posterior distributions without some of the computational difficulties that burden other integration methods (see, e.g., Doucet, de Freitas, & Gordon, 2001). Particle filters provide useful frameworks for describing cognition, particularly in tasks where evidence accumulates over time, as is assumed in almost all decision-making paradigms. The distributions of particles can approximate posterior distributions arbitrarily closely if there is a sufficient number of particles, but also quite poorly if there are few particles; it is this property that allows the models to smoothly move between statistically optimal behavior and flawed, human-like behavior. Models based on particle filtering have been applied to problems, including categorization (Sanborn, Griffiths, & Navarro, 2006), language comprehension (Levy, Reali, & Griffiths, 2009), change detection (Brown & Steyvers, 2009), multiple object tracking (Vul, Frank, Alvarez, & Tenenbaum, 2010),

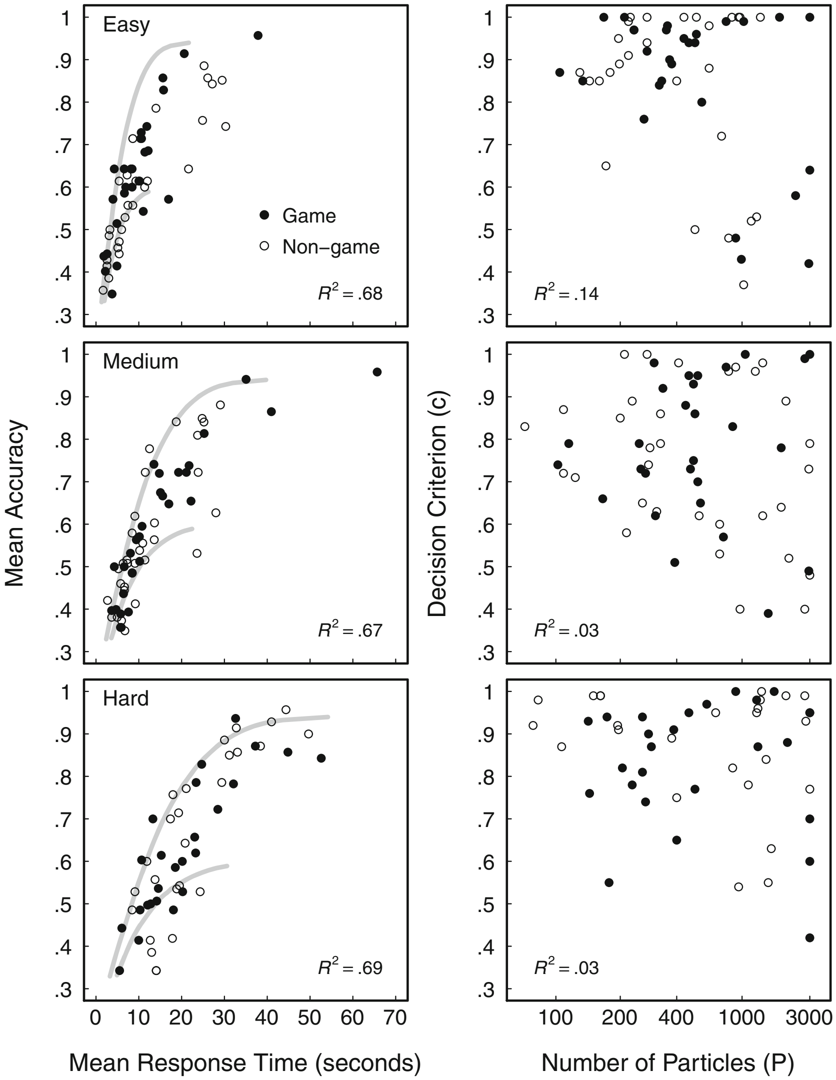


**Fig. 3** Scatterplots of individual participant data from Experiment 1 (left column) and particle filter parameter estimates (right column) for the easy, medium, and hard conditions (upper, middle, and lower rows, respectively). The left column shows that faster responses were generally less accurate in all conditions for game and nongame participants (filled and unfilled symbols, respectively). A particle filter modeled the data by varying decision criterion and the number of particles, $P = 200$ and $P = 2{,}500$, provided to the model (lower and upper gray lines, respectively, in all plots). The right column shows individual participant parameter estimates for the number of particles and decision criterion, which did not reliably vary across game version or difficulty level

determining reward rate payoffs (Yi, Steyvers, & Lee, 2009), and animal learning (Daw & Courville, 2008; Gershman, Blei, & Niv, 2010). We focus here on the output of the particle filter and its psychological implications. For a full explanation of the model and general goodness of fit to data, see the Appendix.

The particle filter we implemented has two simple and psychologically plausible parameters: number of particles (*P*) and decision criterion (*c*). A low or high decision criterion captures the way some people were risky in their choices, while others were more cautious. The number of particles can be interpreted as a limit in cognitive resources, where some individuals have a greater task capacity than others. We interpret the number of particles here as an indicator of a participant's ability (and perhaps their effort). To determine whether task effort or decision riskiness was differentially influenced by game version, we used individual participant mean response time and accuracy data to estimate the parameters of the particle filter model separately for each person. This was accomplished by a grid search that effectively transformed each participant's mean accuracy and response time into estimates of the parameters *P* and *c*. This transformation was one-to-one: Any given accuracy and mean response time pair could be predicted by only one parameter pair. These parameter estimates provided a good fit to mean response time and accuracy data, shown as gray crosses in Fig. 2, and to the full distribution of response times, shown in Fig. 7 of the Appendix.

The right column of Fig. 3 redraws the scatterplots in terms of parameter estimates (*P* and *c*) rather than raw data means (response time and accuracy). The parameter estimates confirm that the addition of gamelike features did not alter the latent variables—the amount of effort applied to the task or the level of riskiness adopted in making the noisy judgments. Comparison of the left and right columns in Fig. 3 also confirms that the model-based analysis was successful in decorrelating response latency and accuracy measures. For the medium and hard difficulty levels, there was no longer any statistically reliable correlation between the number of particles and decision criterion, both *p*s > .05. There was a significant correlation between particle filter parameter estimates in the easy condition, $p < .01$; however, this was much weaker than the correlation coefficient observed between response time and accuracy for this group (see Fig. 3).

With the decorrelated parameter estimates, we conducted a 2 (game version: gamelike vs. nongame) × 3 (difficulty level: easy vs. medium vs. hard) between-subjects analysis of variance separately on the number of particles and decision criterion (log- and inverse-normal-transformed, respectively, to correct for nonnormality). There were no reliable effects of game version or difficulty level on the number of particles, all *p*s > .05. The only significant effect was that of difficulty level on decision criterion, $F(2, 160) = 5.31$, $p < .01$, where the medium difficulty condition demonstrated a lower mean decision criterion than did the easy and hard groups. Bayes factor comparisons confirmed that the gamelike features had no effect on the latent variables (shown in Table 3). When combined across difficulty levels (Rouder & Morey, 2011), there was once again strong evidence in favor of the null hypotheses, that both the number of particles and the decision criterion were unaffected by the addition of gamelike features to the task.

## Experiment 2

In Experiment 1, we found no evidence that gamelike features had any effect on the raw dependent measures or the latent variables predicted by the particle filter. We aimed to replicate the findings of Experiment 1 in a separate cognitive task in Experiment 2. We required participants to detect changes in an underlying data-generating process based on noisy outputs. In each block, there were two choice alternatives, and on each trial, an object would independently appear for these alternatives according to a prespecified underlying probability (unknown to participant). The probabilities were specified such that no alternative "paid out" objects more often, on average, across the experiment but, during most points throughout the task, one alternative paid out objects more often on average than did the other. The prespecified probabilities switched at a random point once in each block. The participants task was to accurately track changes in the underlying probability, on the basis of the observed object payouts, to succeed in the task.

As in Experiment 1, we aimed to ensure that the gamelike condition had the best chance of improving data quality, relative to the nongame condition, by including all of the gamelike elements traditionally invoked in gamelike tasks: colorful displays, animations related to stimulus presentation and response feedback, a detailed back story, and a points-based scoring system shown to participants at all times. In contrast to Experiment 1, where we failed to query subjective experiences, in Experiment 2 we surveyed participant experiences of the gamelike and nongame versions of the task. It could be that participants completing the gamelike version

**Table 3** Bayes factors for particle filter parameter estimates in Experiment 1

| Difficulty level | Number of particles (*P*) | Decision criterion (*c*) |
|---|---|---|
| Easy | 1.48 | 4.90 |
| Medium | 4.40 | 4.05 |
| Hard | 4.21 | 2.34 |
| Overall | 19.98 | 13.95 |

have a more enjoyable experience of the experiment than do those in the nongame task, even if gamelike features produce no measurable change to response data.

## Method

### Participants

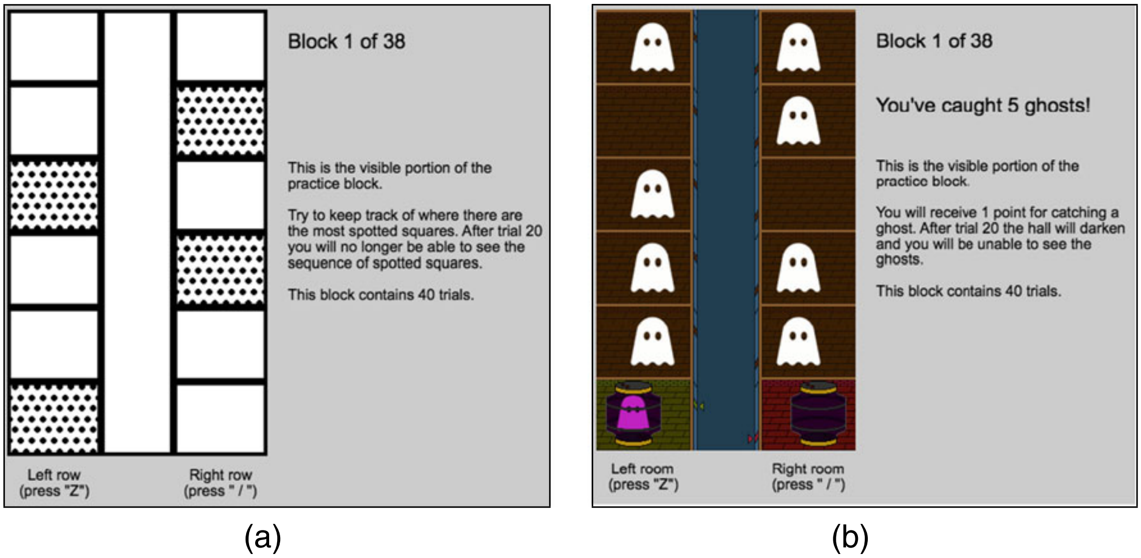
One hundred twenty-seven first-year psychology students from the University of Newcastle participated either in the laboratory ($N$ = 31) or online ($N$ = 96) for course credit. Participants were randomly allocated to the gamelike or nongame version of the task. As in Experiment 1, the gamelike and nongame versions of the task were statistically identical, with the only difference being the "gamed-up" appearance of the gamelike task. At the conclusion of Experiment 2, participants completed a brief questionnaire regarding their experiences in the experiment.

### Properties common to gamelike and nongame versions of the task

In this section, we describe the change detection task as it appeared to participants in the nongame condition, and in the following section, we describe the gamelike features we implemented to "game up" the task. The task involved making repeated binary decisions about noisy outputs based on unknown (to the participants) payoff distributions. The payoff distributions across the two alternatives changed throughout the experiment. To perform well, the participants had to detect the changes in the underlying payoff distributions on the basis of changes in the observed payouts and adjust their behavior accordingly. By appropriately adjusting their behavior to observed changes in payoffs, the participants could optimize their performance in the task.

At the beginning of the task, participants were introduced to the "Spots Experiment," where they were given the objective of finding spotted squares. A demonstration version of the standard task can be viewed online at http://psych.newcastle.edu.au/~sdb231/changeDet/nongame/nongame.html. A brief set of plain text instructions explained that a sequence of spotted and plain squares would move down two rows (described as the left and right rows), from the top of the screen to the bottom. Spotted squares could be collected only when they reached a collection square at the bottom of each of the left and right rows. On each trial, the sequence of spotted and plain squares would move down one position in its respective row. When a spotted square hit the collection square, it remained there until the participant collected it. Each collection square could hold only one spotted square at a time. Therefore, if a second spotted square passed onto an already occupied collection square, the participant was unable to collect the second spotted square. Participants were instructed that they could choose only one collection square on each trial: the left or right. A response for the left collection square was made by pressing the "Z" key, and for the right collection square by pressing the "/" (i.e., question mark) key. If the response collected a spotted square, a green tick was displayed. If the response missed a spotted square (i.e., was a plain square), a red cross was shown. The green tick and red cross were each shown for 250 ms.

An example screenshot of the nongame task is shown in the left panel of Fig. 4. All squares were 100 pixels (wide) × 80 pixels (tall). The plain squares were completely white, except for a 4-pixel black border. The spotted squares contained



Block 1 of 38

This is the visible portion of the practice block.

Try to keep track of where there are the most spotted squares. After trial 20 you will no longer be able to see the sequence of spotted squares.

This block contains 40 trials.

Block 1 of 38

You've caught 5 ghosts!

This is the visible portion of the practice block.

You will receive 1 point for catching a ghost. After trial 20 the hall will darken and you will be unable to see the ghosts.

This block contains 40 trials.

**Fig. 4** Illustrative example of the display in Experiment 2 during practice trials for the nongame task (**a**), where participants were asked to keep track of spotted squares; and the gamelike task (**b**), where participants were asked to capture ghosts in an abandoned building

approximately 48 circular black spots, each 37 pixels in size, evenly spread out on a white background with a 4-pixel black border. The left and right rows were made up of a collection square at the base and five squares placed on top (vertically) of the collection square. The left and right rows were separated by a vertically oriented plain colored middle section (78 × 480 pixels). Therefore, with the left and right rows and middle section, the test display was 278 × 480 pixels. The overall experiment window was 620 × 560 pixels. Below each of the left and right collection squares were text boxes that stated the response keys for each row at all times: "Left row (press 'Z')"; "Right row (press '/')."

Progress information was displayed to the right of the test window at all times. This information indicated progress through the experiment (current block of total blocks) and instructions for the current stage of the experiment. There were three experiment stages: visible and hidden portions of the practice blocks and test blocks. Participants completed 36 test blocks and 2 practice blocks. The first 20 (of 40) trials in the practice blocks were described as the "visible portion" of the block. In this visible phase the participant could observe the sequence of spotted and plain squares moving down the left and right rows, demonstrating to the participant that whether a spotted square appeared or not on a particular trial was relatively unpredictable but that, at any given time, one row tended to have more spotted squares than did the other. On each practice trial in the visible phase. the participant was shown the outcome of the collection square (spotted or plain square) for both the chosen and foregone alternatives. The left panel of Fig. 4 shows an example of the visible portion of the practice blocks.

In the second half of each practice block, the sequence of spotted and plain squares was "hidden" from the participant. All squares, including the collection square (and the middle section), were instead displayed in gray (but still with a black border), and on each trial, the participant was shown only the outcome of the chosen collection square (i.e., they were not shown the outcome of the foregone alternative). All data from practice trials were discarded from analyses. The test phase of the experiment appeared the same as the "hidden phase" of the practice blocks, but without an initial "visible phase," for the 36 test blocks of 60 trials each.

There was one change point in each block of the experiment. The trial of the change point was randomly sampled from an exponential distribution ($M = 10$), constrained not to fall within the first or last 10 trials of the block. That is, the change point would occur between the 11th and 50th trials in the test blocks (between the 11th and 30th trials in the practice blocks). When a change point occurred, the underlying probabilities for the left and right rows were randomly resampled, subject to the constraint that the probability ratios must change (i.e., not resample the current probability ratio, which would produce no apparent change to the participant). We used three probability ratios throughout the experiment, given as the probability that a spotted square would appear in the left versus right row on any trial: .5 versus .5 (both rows provide equal average payoff), .75 versus .25 (left row pays off approximately three times as often as the right row), and .25 versus .75. (right row pays off approximately three times as often as the left row). There was continuity across blocks such that the probability ratio at the end of one block was the same ratio that began the following block.

Participants were instructed that spotted squares would come down the left and right rows in batches and that, sometimes, more spotted squares would appear in the left row and, at other times, more spotted squares would appear in the right row. They were told that within each block, one row might have more spotted squares appear at the beginning of the block and fewer spotted squares by the end of the block. In order to collect the most spotted squares, participants were reminded that the spotted squares might change which row they traveled down (but participants were never explicitly instructed that there would be a change point in each block).

Our primary dependent measures on each trial were whether a spotted square was collected and the latency to make a response (in milliseconds). We recorded a cumulative count of the number of spotted squares collected throughout the task, but at no point was the participant made aware of their cumulative total. Participants who achieved the best scores (greatest number of collected spotted squares) were those who accurately tracked changes in the underlying data-generating process. That is, people who were aware of changes in the observed payoffs of spotted squares changed their response patterns accordingly so as to collect the maximal number of spotted squares.

### Properties specific to the gamelike version of the task

As in Experiment 1, the gamelike task operated on the same principles as the nongame task, except for the "gamed-up" facade, and did not alter any of the physical properties of the stimulus display (e.g., size, shape, location). We again attempted to include many "standard" gamelike features typically incorporated in gamelike tasks to give the gamelike version the best chance of affecting participant engagement and motivation and, consequently, observable performance. A demonstration version of the gamelike task can be viewed online at http://psych.newcastle.edu.au/~sdb231/changeDet/game/ghostGame.html.

The gamelike version of the task required participants to trap ghosts that were on the loose in an abandoned warehouse as efficiently as possible. At the beginning of the task, participants were introduced to the "Ghost Trap Experiment" (the name "Ghostbusters" seemed apt, but we refrained for fear of copyright infringement). The task instructions described the same underlying task as the nongame version, except that there was a detailed back story and animations that explained how the task worked.

Ghosts traveled through two rows of rooms set across a hallway (i.e., the left and right rows, the middle section). Ghost traps were set up in the two rooms at the end of the hallway (i.e., collection squares). Participants were told that when a ghost entered a room with an empty trap, the ghost would be caught. Each ghost trap could hold only one ghost at a time, so if a second ghost entered a room with a full trap, the ghost would pass through without capture. The capture (and evading capture) processes were illustrated with a short animation of the capture process that we expected would aid understanding of the task. Participants were instructed that they could check a room containing a full trap to collect the ghost (as in the nongame task: left trap, "Z" key; right trap, "/" key). Unlike the nongame task, participants were given explicit instruction about the points-based scoring system and were told that they would receive one point for every ghost caught. The current score was displayed at all times on the right side of the experiment display in the form: "You've caught # ghosts!," where # was the cumulative score across the experiment up to the current trial (see right panel of Fig. 4). At the end of each block, a reminder of the number of ghosts caught was also displayed. Participants were instructed that they could check only one ghost trap at a time and, thus, that they should change which side of the hallway they observed wisely.

The gamelike version also contained three experiment phases of the same length and type as the nongame task: *visible* and *hidden* practice trials and test blocks. In the visible practice trials, the hallway and rooms were lit so ghosts could be seen moving down the hallway. During the visible phase, there were animations on each trial showing the process by which ghosts could become stuck in empty traps, how ghosts could be retrieved and captured from a full trap, and how ghosts that approached a currently full trap would freely pass through the capture room (animations on each trial took approximately half a second). In this phase, the animations showed the outcome for both chosen and foregone alternatives, demonstrating how the task functioned. The hidden practice trials and test blocks no longer showed the hallways or lit the trap rooms, due to a "power outage." On each trial, participants observed only the outcome of the chosen trap room (left or right; i.e., they were not shown the status of the unchosen trap room).

As with the nongame task, participants were instructed that ghosts would move down the left and right hallways in batches—sometimes, more ghosts would be in the left hallway, and at other times, more ghosts would appear in the right hallway—and that within each block, one row might have more ghosts appear at the beginning than at the end of the block. Importantly, the gamelike version of the task had a far more consistent "story" than the nongame version: There were ghosts wandering an abandoned building (why objects moved down the left and right rows); traps could hold only one ghost at a time, so a second ghost could not be caught if a trap was currently occupied (why a collection square could hold only one object at a time); and there was a power outage (why there were visible and hidden phases). We believed that these back-story elements would aid understanding of the task and increase participant engagement.

### Questionnaire

We administered an 11-item questionnaire about subjective experiences in the experiment at the conclusion of the task. The questions generally related to understanding of the task, effort applied, task interest, motivation, enjoyment, and boredom (shown in Table 4). Each question was rated on a 7-point Likert scale, with 8 of the 11 items worded positively and 3 negatively. We calculated a Bayes factor for each questionnaire item individually and then created a combined score that was calculated separately for each participant by adding the scores of the positively worded questions (Qs. 1–6, 10, 11) and subtracting the scores of the negatively worded questions (Qs. 7–9; giving a possible range from 13 to 53). We provide a separate Bayes factor for this combined score (i.e., not a meta-analytic Bayes factor).

## Results

### Experimental task

Our primary focus in Experiment 2 was on overall performance as measured by the total number of objects collected (ghosts in the gamelike version; spotted squares in the nongame version) separately for game version (gamelike, nongame) and testing location (laboratory, online). As with Experiment 1, we supplement ANOVA analyses with Bayes factors to provide direct evidence for the null hypothesis (denoted in text with $BF_{H_0}$).

Consistent with the results of Experiment 1, gamelike features had no reliable effects on the data (see Table 5). There were no reliable differences in the number of objects collected between the gamelike and nongame versions, $F(1, 123) = 0.06, p = .80 (BF_{H_0} = 7.00)$, or between laboratory and online participants, $F(1, 123) = 1.27$, $p = .26 (BF_{H_0} = 3.52)$. The interaction between the two was also not reliable, $F(1, 123) = 1.15$, $p = .29$. We also examined whether there were any differences in mean choice latency on each trial. Again, we found no differences between game versions on mean response time, $F(1, 123) = 0.53, p = .47 (BF_{H_0} = 5.57)$, or an interaction between game version and testing location, $F(1, 123) = 0.39$, $p = .53$. There was, however, a marginally significant

**Table 4** The 11 items in the questionnaire given to participants after completing Experiment 2. Answers were provided on a 7-point Likert scale (1 = *lowest*; 7 = *highest*). The bottom entry shows a total score obtained by combining across questionnaire items. For all questionnaire items, and separately for gamelike versus nongame and laboratory versus online, the two left columns show means (*M*s) and standard deviations (*SD*s) followed by the Bayes factor indicating support for the null hypothesis $BF_{H_0}$. Bayes factors greater than one support the null hypothesis, and those less than one support the alternative hypothesis

| | Question | Gamelike vs. Nongame | | | Laboratory vs. Online | | |
|---|---|---|---|---|---|---|---|
| | | Game – *M*(*SD*) | Nongame – *M*(*SD*) | $BF_{H_0}$ | Lab – *M*(*SD*) | Online – *M*(*SD*) | $BF_{H_0}$ |
| 1 | How easily did you understand the instructions given at the beginning of the task? | 4.14 (1.95) | 3.72 (1.70) | 3.24 | 4.63 (1.68) | 3.73 (1.85) | 0.32 |
| 2 | How much effort did you put in throughout the task? | 4.53 (1.76) | 4.41 (1.60) | 6.77 | 5.34 (1.49) | 4.20 (1.65) | 0.01 |
| 3 | How interesting would you rate the graphics? | 4.20 (1.75) | 2.75 (1.51) | $9.1 \times 10^{-5}$ | 4.09 (1.94) | 3.35 (1.71) | 1.12 |
| 4 | How mentally stimulating did you find this task? | 3.18 (1.73) | 2.75 (1.57) | 2.57 | 3.56 (1.81) | 2.80 (1.58) | 0.78 |
| 5 | How worthwhile did you find the task to be? | 3.01 (1.63) | 2.75 (1.58) | 4.90 | 3.84 (1.69) | 2.59 (1.46) | 0.01 |
| 6 | How much did you enjoy participating in this task? | 3.59 (1.96) | 2.82 (1.65) | 0.44 | 4.16 (1.89) | 2.94 (1.76) | 0.06 |
| 7[a] | How boring did you find the task to be? | 4.10 (1.97) | 4.52 (1.95) | 3.54 | 3.84 (1.72) | 4.44 (2.02) | 1.86 |
| 8[a] | How repetitive did you find this task to be? | 5.87 (1.64) | 5.75 (1.89) | 6.88 | 5.97 (1.33) | 5.77 (1.87) | 5.22 |
| 9[a] | How much were you looking forward to the task finishing? | 5.55 (1.62) | 5.85 (1.28) | 3.72 | 4.97 (1.62) | 5.92 (1.35) | 0.10 |
| 10 | How willing would you be to complete this task again? | 3.73 (1.90) | 3.23 (1.75) | 2.26 | 4.59 (2.00) | 3.15 (1.66) | 0.01 |
| 11 | How well do you think you performed, compared to other people who have participated in this experiment? | 3.94 (1.38) | 3.63 (1.39) | 3.35 | 4.14 (1.17) | 3.69 (1.44) | 1.46 |
| | Total score | 14.82 (12.82) | 9.94 (11.04) | 0.56 | 19.58 (12.09) | 10.32 (11.45) | $9.3 \times 10^{-3}$ |

[a] Negatively worded questions

main effect of testing location on response latency, with online participants exhibiting faster responses (*M* = 209.42 ms, *SD* = 138.34) than laboratory participants (*M* = 263.52 ms, *SD* = 135.20) according to an ANOVA, *F*(1, 123) = 3.59, *p* = .06, but this effect would not be considered reliable according to a Bayes factor ($BF_{H_0} = 1.13$).

There were also no differences in task engagement as indirectly measured by the proportion of outliers in the number of total objects collected. We examined outliers in each condition, defined as participants who scored below two standard deviations (pooled across the four conditions) in the number of objects collected. We found that the proportion of participants judged as outliers varied very little between conditions—from a high of 6.8 % (in the nongame and online condition) to a low of 5.8 % (nongame in lab).

**Table 5** Experiment 2 mean items collected (ghosts in the gamelike version; spots in the nongame version) and total questionnaire scores, separately for game version and testing location

| Version | Location | Mean items collected (*SD*) | Mean questionnaire score (*SD*) |
|---|---|---|---|
| Gamelike | Laboratory | 1,413.19 (106.70) | 22.94 (12.63) |
| | Online | 1,411.17 (98.79) | 12.26 (11.87) |
| Nongame | Laboratory | 1,452.73 (102.92) | 15.77 (10.59) |
| | Online | 1,404.55 (109.43) | 8.04 (10.61) |

## Questionnaire

In addition to analyzing performance in the experimental task, our secondary focus was to analyze participants' self-reported experiences. This was to investigate whether gamelike features (or testing location) influenced participants' subjective experience in the experiment, even though the dependent measures from the experimental task were unchanged. Table 5 shows the total mean scores for the questionnaire in each of the conditions, while Table 4 shows Bayes factors for separate comparisons between the gamelike versus nongame and laboratory versus online conditions. Bayes factors greater than 1 indicate support for the null hypothesis (no difference between conditions).

Unlike the raw data measures from the experiment, there were some reliable differences in questionnaire scores between the gamelike and nongame conditions. The gamelike version had a greater total score than the nongame condition, indicating a more positive experience. In particular, participants in the gamelike version rated the graphics of the experimental task more interesting and had a more enjoyable experience than did those in the nongame version. Interestingly, even though the experience in the gamelike version was more positive overall, the gamelike features did not appear to be sufficiently motivating or interesting enough to reduce the repetitiveness and boredom experienced during the task, as evidenced by scores on the negative items.

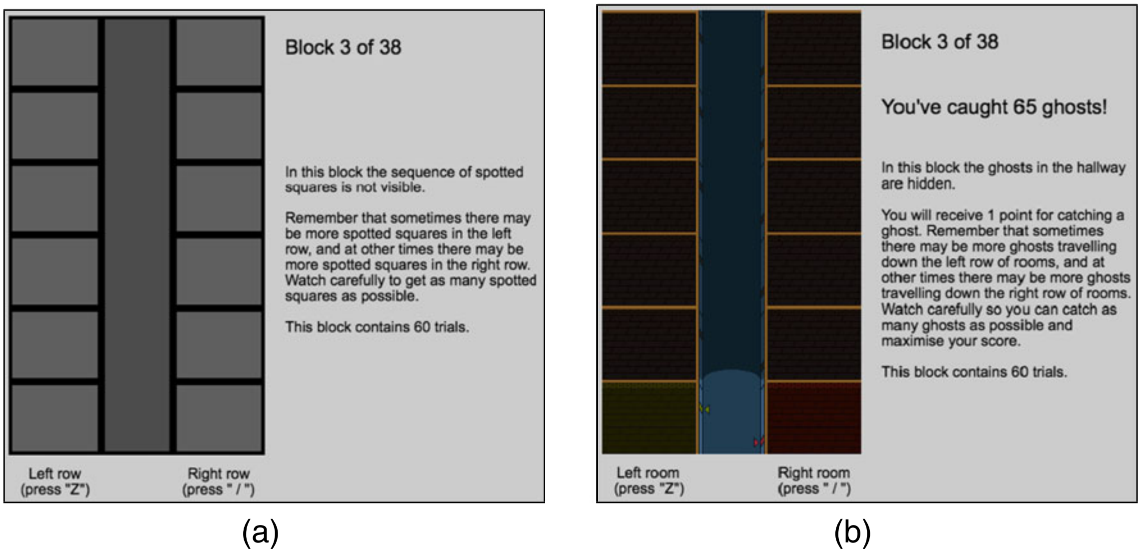
Our questionnaire item #2—self-reported effort applied to the task—provided an alternative approach to examining the criterion validity of our argument that the data do not differ across game versions. Any reasonable hypothesis about performance would state that task effort and task performance are positively associated, which was the case in Experiment 2 (using number of objects collected as the measure of performance), $r = .33$, $t(120) = 3.80$, $p < .001$. If the nongame task produced data of similar quality to the gamelike version, both the amount of effort applied and the strength of the effort–performance relationship should not differ between these two groups. As can be seen above (Table 4), the amount of effort applied did not differ between conditions. Furthermore, we found significant correlations between effort and performance in both the gamelike task, $r = .34$, $t(62) = 2.86$, $p < .01$, and the nongame task, $r = .33$, $t(56) = 2.61$, $p = .01$, and these two correlations were not significantly different: $Z = .07$, $p = .94$. This suggests that the data from both game versions were of equal quality and is consistent with the particle filter modeling of Experiment 1 data. In our modeling, we found no difference in the effort applied to the task between gamelike and nongame conditions (i.e., number of particles), which provides convergent validity for the parameter estimates of the particle filter, as well as our interpretation of those parameter estimates.

As compared with the game version, there were some reliable differences across testing locations. Again, there was a strong effect on the total score, with laboratory participants reporting a more positive experience of the experiment. There were also many strong effects on individual items. As compared with online participants, those who participated in the laboratory understood task instructions more easily, exerted more effort throughout the task, found the task to be more worthwhile and enjoyable, and were more willing to complete the task again.

## General discussion

A common practice in psychological research is to incorporate features from computer games into standard psychological experiments. The effect of gamelike features on data quality is assumed to be positive, because of plausible hypotheses about their effects on underlying cognitions—particularly, motivation and attention. However, the effect that gamelike features have on data is almost never tested empirically. We found that, on average, there were no differences in outcome measures (response latency, decision accuracy, points scored) between gamelike and nongame versions of statistically identical tasks. We also demonstrated the data from both versions of the choice task were consistent with a single quantitative model whose parameters did not vary between conditions. Furthermore, despite the lack of an effect of gamelike features on observed data, the gamelike version of a change detection task was subjectively perceived as more interesting and enjoyable than the nongame version.

Gamelike features may alter task performance, as Washburn (2003) observed, but evidence from Experiment 1 suggests this is likely the result of a speed–accuracy trade-off. In keeping with this hypothesis, Green and Bavelier (2006) found that computer game players set a much faster speed–accuracy trade-off than did other participants in a perceptual attention experiment, making much faster responses with very small changes in accuracy. In our Experiment 1 data (left column of Fig. 3), both gamelike and nongame tasks naturally produced large within-group variability in the preference for speed



In this block the sequence of spotted squares is not visible.

Remember that sometimes there may be more spotted squares in the left row, and at other times there may be more spotted squares in the right row. Watch carefully to get as many spotted squares as possible.

This block contains 60 trials.

In this block the ghosts in the hallway are hidden.

You will receive 1 point for catching a ghost. Remember that sometimes there may be more ghosts travelling down the left row of rooms, and at other times there may be more ghosts travelling down the right row of rooms. Watch carefully so you can catch as many ghosts as possible and maximise your score.

This block contains 60 trials.

**Fig. 5** Illustrative example of the display in Experiment 2 during test trials for the nongame task (**a**) and the gamelike task (**b**). In the test phase, the display for the nongame and gamelike tasks are almost identical

or accuracy, but there was no systematic effect of gamelike features. It seems that, if an experimenter really wanted to manipulate the speed–accuracy trade-off, there are simpler and more controlled ways than by utilizing gamelike features. For instance, simply asking participants to respond quickly or to respond carefully usually works well (e.g., Brown & Heathcote, 2005, 2008; Ratcliff & Rouder, 1998; Usher & McClelland, 2001).

There are several ways to interpret the null results from our two experiments. It is possible that our gamelike features were not sufficiently salient to have any effect. However, the gamelike features we introduced in each experiment were typical of those used in psychology experiments, and the implementation details (e.g. graphical, auditory, and animation quality) were at least as good as most other examples. Thus, if our gamelike features produced no effects on data, it is likely that other standard gamelike features will also have no effect. Also, in Experiment 2, we found that the very strong contextual difference between taking part in the experiment online versus in the laboratory made a strong impact on participants' experience (as reported in the questionnaire). Even this very strong effect—which is surely a larger context effect than could ever be induced by changing computer animations and other gamelike features—had no reliable effects on data. This speaks against the notion that gaming-up experiments, no matter how well executed, will ever improve data.

Another possibility is that a drawback of our tasks curtailed the potential benefits from gamelike features. That is, while the gamelike features may have worked to enhance performance, the more complicated stimulus display of the gamed-up task may have simultaneously worsened performance. If these effects were in balance, it could explain our observation of no difference in performance between the versions. This interpretation is possible for Experiment 1, but it seems very unlikely for Experiment 2. In Experiment 2, although the introductory phases of the experiment were very different between the nongame and gamelike tasks, the display during the data collection phase was very similar between versions (see Fig. 5). This makes it difficult to see how any putative improvement due to gamelike features may have been suppressed by irrelevant task elements. Nevertheless, future research could address this more carefully by using identical visual displays in the nongame and gamelike versions.

Nevertheless, and at a minimum, our results, combined with those of Washburn (2003), suggest that researchers should not simply assume that adding gamelike features to a task confers any benefits on performance. It may be that improved data will result from computer games only when the underlying game play has been altered, as computer programmers suggest, since the fundamental task is often rated as one of the most important components of a game (Wood et al., 2004). In the present study and most existing research using gamelike tasks, the game play (i.e., the experimental task) is deliberately unaltered, with gamelike features added only to superficial aspects.

Alternatively, our null results from experimental data, combined with the results of our Experiment 2 questionnaire, could be interpreted as positive news, implying that experiments can be "gamed up" without worrying that the data will be negatively influenced. Gaming up an experiment might be viewed as intrinsically worthwhile to the experimenter, or beneficial because it makes the participants' perception of the experiment more positive and enjoyable.

**Acknowledgements** We thank Mike Yi and Mark Steyvers for assistance in the development of the materials for Experiment 2. This research was supported by the Keats Endowment Research Fund and by an Australian Research Council Discovery Project (0878858).

## Appendix

### Particle filter details

The general idea of a particle filter is to take a set of $P$ particles, each of which is treated as a sample from the posterior distribution of interest (e.g., they may be "hypotheses" about which of the $K$ choice alternatives is the target). Before any data are observed, these particles are samples from a prior distribution. Each time a new datum is observed, the entire set of $P$ particles is "evolved." This evolution step can take many forms but usually involves resampling the particles according to their likelihood. Conditionally likely particles (those consistent with the datum) have a higher probability of being resampled than conditionally unlikely particles (those inconsistent with the datum), which become rarer. If the resampling algorithm meets certain conditions, the distribution across particles approximates the full posterior distribution—that is, based on the entire history of observations—even though this history is not explicitly stored.

Different particle filters can be developed to approximate different posterior distributions, by employing likelihood calculations based on particular assumptions about which environmental quantities are known and which are estimated. We set up a particle filter to model individual participant data from both game versions of Experiment 1, illustrated conceptually in Fig. 6. Each particle is a number from 1, …, $K$ corresponding to a belief about which square is the target. At the beginning of a decision, particles are randomly sampled from a uniform prior. An initial set of $P = 10$ particles for a decision between $K = 4$ choice alternatives is shown in the top row of the right-hand side of Fig. 6. In this example, three particles hypothesize that square 1 is the target (which it actually is), four particles that the target is 2, and so on. In our task, on each time step, a dot either appeared or did not appear in each square, and these are represented by the

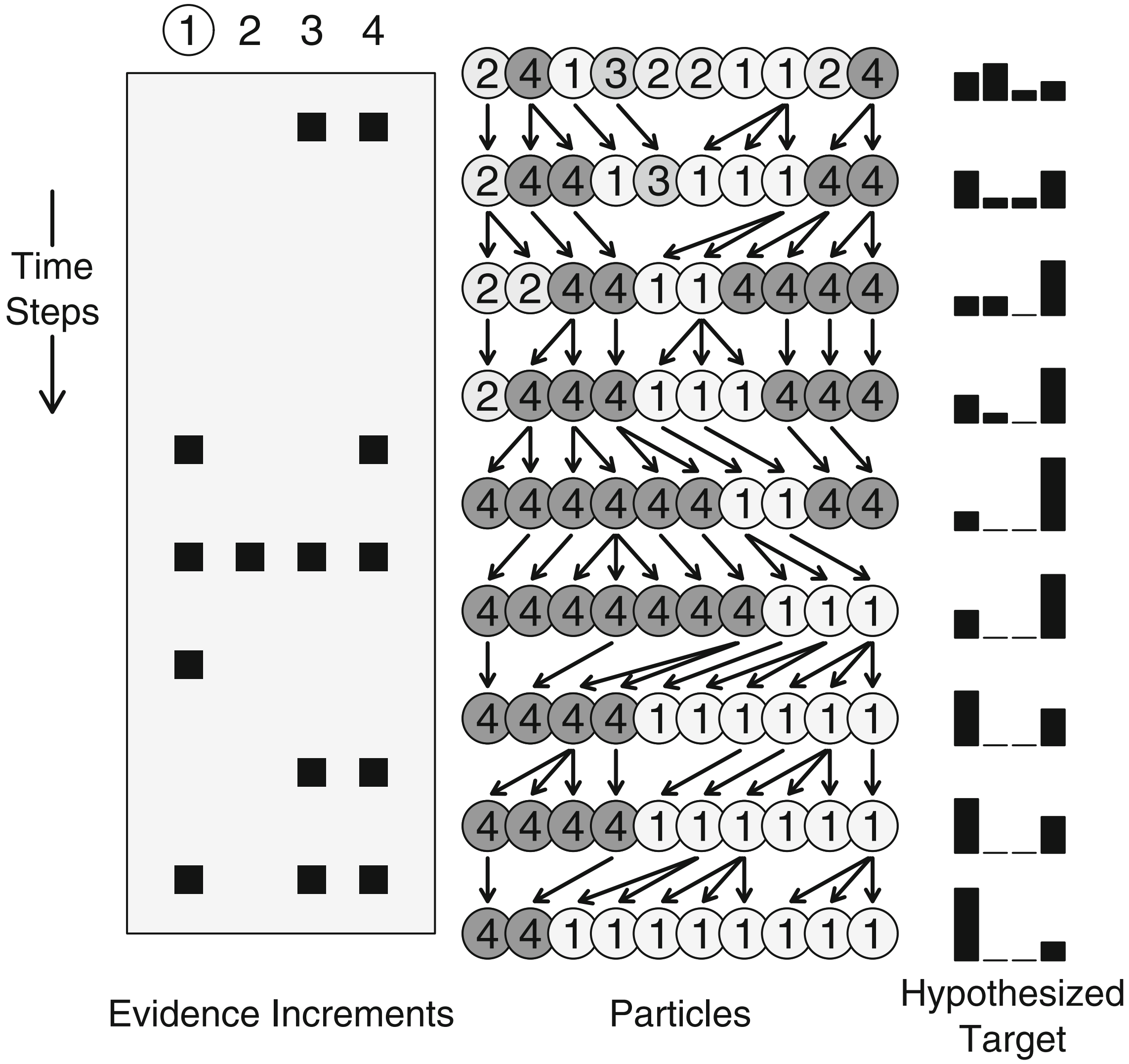


**Fig. 6** Illustrative example of the particle filter algorithm (see the Appendix text for details)

square "evidence increments" in the shaded rectangle on the left of Fig. 6. The uppermost row of evidence tokens illustrates the first time step: A dot appeared in both of squares 3 and 4, but not in squares 1 or 2. The probability of this pattern of dots across the squares can easily be calculated under each particle's hypothesis, if we assume that target and distractor probability rates are perfectly known. For example, the probability of observing the first time step's outcomes (new dots for squares 3 and 4, no new dots for 1 and 2) is 4.8 % if square 1 is really the target, 7.2 % if square 4 is the target, and so on. These probabilities are used to resample a new set of *P* particles for the next time step, with replacement. The outcome of this resampling is shown by the second row of particles.

After each step, the posterior probability that each square is the target can be estimated by calculating the proportion of particles representing that square, illustrated by the histograms on the far right side of Fig. 6. We assumed that a decision was triggered whenever the largest posterior probability exceeded a criterion threshold *c*. In our example, if *c* = .8, the particle filter would incorrectly respond (with square 4) after the fourth time step, since eight out of ten particles represented square 4 at that time. This demonstrates the difficult speed–accuracy trade-off in the present experiment: Responding too early may result in incorrectly selecting a distractor that has by chance accumulated the most evidence thus far in the trial. In the time steps that follow, the true target (square 1) gathers more evidence tokens and, hence, is conditionally more likely so particles begin to resample toward square 1.

By varying both the number of particles and decision criterion, the particle filter predicts the speed–accuracy trade-off seen in our data. For a large number of particles, *P* = 2,500, sweeping the decision criterion from *c* = .3 − 1 naturally produces the negatively accelerated upper gray line in each of the scatterplots in the left column of Fig. 3.

This $P = 2{,}500$ line represents an approximate upper bound on performance. That is, on average, people did not respond more quickly or accurately than this model.

Simply reducing the number of particles qualitatively changes model predictions. With fewer particles, the model predicts faster responses with lower accuracy. For a $P = 200$ model and decision criterion swept from $c = .5 - 1$, the particle filter predicts an approximate lower bound on performance, whereby most people perform as accurately or as quickly, or better, than this model, depicted by the lower gray lines in the left column of Fig. 3.

To make the model respond as slowly as the human participants, we had to scale response time predictions (this is why the upper and lower "bounds" are only approximate). We assume that it took participants just under 15 ms to inspect each square in the display, so this means that the response time predictions of the particle filter must be multiplied by 1.2 in the easy conditions, 2.3 in the medium groups, and 3.1 in the hard conditions. In other words, the entire set of particles was updated, on average, every 80, 153, and 207 ms in the easy, medium, and hard difficulty levels, respectively. It is this scaling factor that makes the same $P = 200$ and $P = 2{,}500$ model predictions differ in the left column of Fig. 3. The majority of data fall between the $P = 200$ and $P = 2{,}500$ model predictions shown in Fig. 3, indicating that the particle filter provides a good account of the data across both game versions. The particle filter also provides a good fit to the distribution of response times, even though the model was not fit to these data, as is shown in Fig. 7.

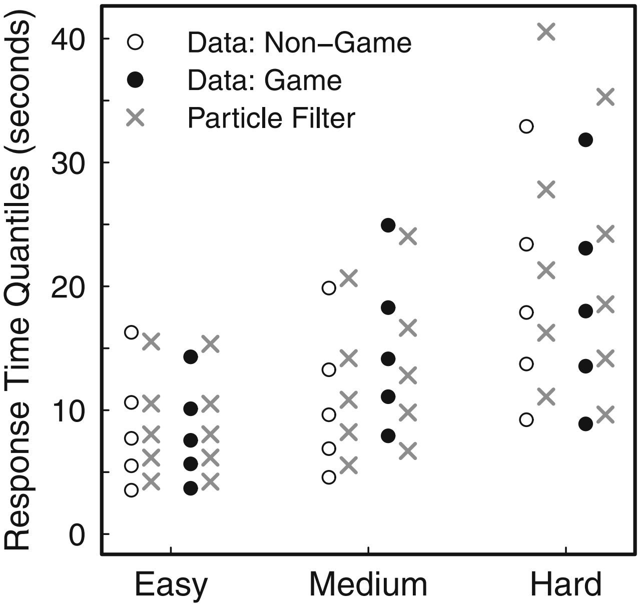


**Fig. 7** Fit of the particle filter model to response time distributions. Data from the nongame and gamelike conditions are shown as open and filled symbols, respectively. Predictions of the particle filter are shown with gray crosses. Symbols represent, from bottom to top, the 10 %, 30 %, 50 % (i.e., median), 70 %, and 90 % quantiles of the response time distribution

## References

Aidman, E. V., & Shmelyov, A. G. (2002). Mimics: A symbolic conflict/cooperation simulation program, with embedded protocol recording and automatic psychometric assessment. *Behavior Research Methods, Instruments, & Computers, 34,* 83–89.

Alloway, T. P., Corley, M., & Ramscar, M. (2006). Seeing ahead: Experience and language in spatial perspective. *Memory & Cognition, 34,* 380–386.

Andrews, G., & Halford, G. S. (2002). A cognitive complexity metric applied to cognitive development. *Cognitive Psychology, 45,* 153–219.

Arcediano, F., Ortega, N., & Matute, H. (1996). A behavioural preparation for the study of human Pavlovian conditioning. *Quarterly Journal of Experimental Psychology, 49,* 270–283.

Arthur, W., Jr., Strong, M. H., Jordan, J. A., Williamson, J. E., Shebilske, W. L., & Regian, J. W. (1995). Visual attention: Individual differences in training and predicting complex task performance. *Acta Psychologica, 88,* 3–23.

Artigas, A. A., Chamizo, V. D., & Peris, J. M. (2001). Inhibitory associations between neutral stimuli: A comparative approach. *Animal Learning & Behavior, 29,* 46–65.

Baeyens, F., Vansteenwegen, D., Beckers, T., Hermans, D., Kerkhof, I., & de Ceulaer, A. (2005). Extinction and renewal of Pavlovian modulation in human sequential feature positive discrimination learning. *Learning & Memory, 12,* 178–192.

Baker, A. G., Mercier, P., Vallee-Tourangeau, V., Frank, R., & Pan, M. (1993). Selective associations and causality judgments: Presence of a strong causal factor may reduce judgments of a weaker one. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19,* 414–432.

Berger, A., Jones, L., Rothbart, M. K., & Posner, M. I. (2000). Computerized games to study the development of attention in childhood. *Behavior Research Methods, Instruments, & Computers, 32,* 297–303.

Blanco, F., Matute, H., & Vadillo, M. A. (2010). Contingency is used to prepare for outcomes: Implications for a functional analysis of learning. *Psychonomic Bulletin & Review, 17,* 117–121.

Brown, S., & Heathcote, A. (2005). A ballistic model of choice response time. *Psychological Review, 112,* 117–128.

Brown, S., & Heathcote, A. J. (2008). The simplest complete model of choice reaction time: Linear ballistic accumulation. *Cognitive Psychology, 57,* 153–178.

Brown, S. D., & Steyvers, M. (2009). Detecting and predicting changes. *Cognitive Psychology, 58,* 49–67.

Buchner, A., Mehl, B., Rothermund, K., & Wentura, D. (2006). Artificially induced valence of distractor words increases the effects of irrelevant speech on serial recall. *Memory & Cognition, 34,* 1055–1062.

Carneiro, P., Fernandez, A., & Dias, A. R. (2009). The influence of theme identifiability on false memories: Evidence for age–dependent opposite effects. *Memory & Cognition, 37,* 115–129.

Castro, L., & Wasserman, E. A. (2007). Discrimination blocking: Acquisition versus performance deficits in human contingency learning. *Learning & Behavior, 35,* 149–162.

Correia, C. J., & Cameron, J. M. (2010). Development of a simulated drinking game procedure to study risky alcohol use. *Experimental and Clinical Psychopharmacology, 18,* 322–328.

Daw, N. D., & Courville, A. C. (2008). The rat as particle filter. In J. Platt, D. Koller, Y. Singer, & S. Roweis (Eds.), *Advances in Neural*

*Information Processing Systems 20* (pp. 369–376). Cambridge, MA: MIT Press.

Day, E. A., Arthur, W., Jr., & Shebilske, W. L. (1997). Ability determinants of complex skill acquisition: Effects of training protocol. *Acta Psychologica, 97,* 145–165.

Dixon, J. A., & Banghert, A. S. (2004). On the spontaneous discovery of a mathematical relation during problem solving. *Cognitive Science, 28,* 433–449.

Dixon, J. A., & Dohn, M. C. (2003). Redescription disembeds relations: Evidence from relational transfer and use in problem solving. *Memory & Cognition, 31,* 1082–1093.

Doucet, A., de Freitas, N., & Gordon, N. (2001). *Sequential Monte Carlo methods in practice*. New York: Springer.

Droit-Volet, S., Tourret, S., & Wearden, J. (2004). Perception of the duration of auditory and visual stimuli in children and adults. *Quarterly Journal of Experimental Psychology, 57,* 797–818.

Drury, J., Cocking, C., Reicher, S., Burton, A., Schofield, D., Hardwick, A., …, Langston, P. (2009). Cooperation versus competition in a mass emergency evacuation: A new laboratory simulation and a new theoretical model. *Behavior Research Methods*, *41*, 957–970.

Dunbar, G., Hill, R., & Lewis, V. (2001). Children's attentional skills and road behavior. *Journal of Experimental Psychology: Applied, 7,* 227–234.

Ekstrom, A. D., & Bookheimer, S. Y. (2007). Spatial and temporal episodic memory retrieval recruit dissociable functional networks in the human brain. *Learning & Memory, 14,* 645–654.

Fabiani, M., Buckley, J., Gratton, G., Coles, M. G. H., & Donchin, E. (1989). The training of complex task performance. *Acta Psychologica, 71,* 259–299.

Fazio, R. H., Eiser, J. R., & Shook, N. J. (2004). Attitude formation through exploration: Valence asymmetries. *Journal of Personality and Social Psychology, 87,* 293–311.

Finke, A., Lenhardt, A., & Ritter, H. (2009). The MindGame: A P300–based brain–computer interface game. *Neural Networks, 22,* 1329–1333.

Franssen, M., Clarysse, J., Beckers, T., van Vooren, P. R., & Baeyens, F. (2010). A free software package for a human online–conditioned suppression preparation. *Behaviour Research Methods, 42,* 311–317.

Frey, A., Hartig, J., Ketzel, A., Zinkernagel, A., & Moosbrugger, H. (2007). The use of virtual environments based on a modification of the computer game Quake III Arena in psychological experimenting. *Computers in Human Behavior, 23,* 2026–2039.

Gershman, S. J., Blei, D. M., & Niv, Y. (2010). Context, learning, and extinction. *Psychological Review, 117,* 197–209.

Green, C. S., & Bavelier, D. (2006). Effect of action video games on the spatial distribution of visuospatial attention. *Journal of Experimental Psychology: Human Perception and Performance, 32,* 1465–1478.

Gunzelmann, G., & Anderson, J. R. (2006). Location matters: Why target location impacts performance in orientation tasks. *Memory & Cognition, 34,* 41–59.

Hanauer, J. B., & Brooks, P. J. (2003). Developmental change in the cross–modal Stroop effect. *Perception & Psychophysics, 65,* 359–366.

Hansberger, J. T., Schunn, C. D., & Holt, R. W. (2006). Strategy variability: How too much of a good thing can hurt performance. *Memory & Cognition, 34,* 1652–1666.

Hutcheson, A. T., & Wedell, D. H. (2009). Moderating the route angularity effect in a virtual environment: Support for a dual memory representation. *Memory & Cognition, 37,* 514–521.

Jackson, D. N., III, Vernon, P. A., & Jackson, D. N. (1993). Dynamic spatial performance and general intelligence. *Intelligence, 17,* 451–460.

Johnson, T. R., & Krems, J. F. (2001). Use of current explanations in multicausal abductive reasoning. *Cognitive Science, 25,* 903–939.

Johnson, C. I., & Mayer, R. E. (2010). Applying the self–explanation principle to multimedia learning in a computer–based game–like environment. *Computers in Human Behavior, 26,* 1246–1252.

Kirsh, D., & Maglio, P. (1994). On distinguishing epistemic from pragmatic action. *Cognitive Science, 18,* 513–549.

Krageloh, C. U., Zapanta, A. E., Shepherd, D., & Landon, J. (2010). Human choice behaviour in a frequently changing environment. *Behavioural Processes, 83,* 119–126.

Kujala, J. V., Richardson, U., & Lyytinen, H. (2010). Estimation and visualizaion of confusability matrices from adaptive measurement data. *Journal of Mathematical Psychology, 54,* 196–207.

Levy, R., Reali, F. Griffiths, T. L. (2009). Modeling the effects of memory on human online sentence processing with particle filters. *Advances in Neural Information Processing Systems*, *22*.

Lie, C., Harper, D. N., & Hunt, M. (2009). Human performance on a two–alternative rapid–acquisition choice task. *Behavioural Processes, 81,* 244–249.

Lien, M.-C., Ruthruff, E., Remington, R. W., & Johnston, J. C. (2005). On the limits of advance preparation for a task switch: Do people prepare all the task some of the time or some of the task all the time? *Journal of Experimental Psychology: Human Perception and Performance, 31,* 299–315.

Logie, R., Baddeley, A., Mane, A., Donchin, E., & Sheptak, R. (1989). Working memory in the acquisition of complex cognitive skills. *Acta Psychologica, 71,* 53–87.

Maglio, P. P., Wenger, M. J., & Copeland, A. M. (2008). Evidence for the role of self–priming in epistemic action: Expertise and the effective use of memory. *Acta Psychologica, 127,* 72–88.

Mane, A. M., Adams, J. A., & Donchin, E. (1989). Adaptive and part–whole training in the acquisition of a complex perceptual–motor skill. *Acta Psychologica, 71,* 179–196.

Mane, A., & Donchin, E. (1989). The space fortress game. *Acta Psychologica, 71,* 17–22.

Mather, M., Gorlick, M. A., & Lighthall, N. R. (2009). To brake or accelerate when the light turns yellow? Stress reduces older adults' risk taking in a driving game. *Psychological Science, 20,* 174–176.

McPherson, J., & Burns, N. R. (2007). Gs invaders: Assessing a computer game-like test of processing speed. *Behavior Research Methods, 39,* 876–883.

McPherson, J., & Burns, N. R. (2008). Assessing the validity of computer game-like tests of processing speed and working memory. *Behavior Research Methods, 40,* 969–981.

Metcalfe, J., Kornell, N., & Finn, B. (2009). Delayed versus immediate feedback in children's and adults' vocabulary learning. *Memory & Cognition, 37,* 1077–1087.

Molet, M., Jozefowiez, J., & Miller, R. R. (2010). Integration of spatial relationships and temporal relationships in humans. *Learning & Behavior, 38,* 27–34.

Nelson, J. B., & Sanjuan, M. C. (2008). Flattening generalization gradients, context, and perceptual learning. *Learning & Behavior, 36,* 279–289.

Nelson, J. B., Sanjuan, M. C., Vadillo-Ruiz, S., & Perez, J. (2011). Experimental renewal in human participants. *Journal of Experimental Psychology: Animal Behavior Processes, 37,* 58–70.

Newman, E. L., Caplan, J. B., Kirschen, M. P., Korolev, I. O., Sekuler, R., & Kahana, M. J. (2007). Learning your way around town: How virtual taxicab drivers learn to use both layout and landmark information. *Cognition, 104,* 231–253.

Ozubko, J. D., & Joordens, S. (2008). Super memory bros.: Going from mirror patterns to concordant patterns via similarity enhancements. *Memory & Cognition, 36,* 1391–1402.

Paredes-Olay, C., Abad, M. J., Gamez, M., & Rosas, J. M. (2002). Transfer of control between causal predictive judgments and instrumental responding. *Animal Learning & Behavior, 30,* 239–248.

Ploog, B. O., Banerjee, S., & Brooks, P. J. (2009). Attention to prosody (intonation) and content in children with autism and in typical children using spoken sentences in a computer game. *Research in Autism Spectrum Disorders, 3,* 743–758.

Price, H. L., & Connolly, D. A. (2006). BatMon II: Children's category norms for 33 categories. *Behavior Research Methods, 38,* 529–531.

Rabbitt, P., Banerji, N., & Szymanski, A. (1989). Space Fortress as an IQ test? Predictions of learning and of practised performance in a complex interactive video–game. *Acta Psychologica, 71,* 243–257.

Radvansky, G. A., & Copeland, D. E. (2006). Walking through doorways causes forgetting: Situation models and experienced space. *Memory & Cognition, 34,* 1150–1156.

Raftery, A. E. (1995). Bayesian model selection in social research. In P. V. Marsden (Ed.), *Sociological methodology* (pp. 111–196). Cambridge: Blackwells.

Raijmakers, M. E. J., Dolan, C. V., & Molenaar, P. C. M. (2001). Finite mixture distribution models of simple discrimination learning. *Memory & Cognition, 29,* 659–677.

Ratcliff, R., & Rouder, J. N. (1998). Modeling response times for two–choice decisions. *Psychological Science, 9,* 347–356.

Rouder, J. N., & Morey, R. D. (2011). A Bayes factor meta–analysis of Bem's ESP claim. *Psychonomic Bulletin & Review, 18,* 682–689.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian *t*–tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review, 16,* 225–237.

Salas, E., & Cannon-Bowers, J. A. (2001). The science of training: A decade of progress. *Annual Review of Psychology, 52,* 471–499.

Sallas, B., Mathews, R. C., Lane, S. M., & Sun, R. (2007). Developing rich and quickly accessed knowledge of an artificial grammar. *Memory & Cognition, 35,* 2118–2133.

Salthouse, T. A., & Prill, K. (1983). Analysis of a perceptual skill. *Journal of Experimental Psychology: Human Perception and Performance, 9,* 607–621.

Sanborn, A. N., Griffiths, T. L., Navarro, D. J. (2006). A more rational model of categorisation. In R. Sun, N. Miyake (eds), *Proceedings of the 28th Annual Conference of the Cognitive Science Society.*

Schonfeld, E. (2010). *SCVNGR's secret game mechanics playdeck.* Tech Crunch, Retrieved August 14, 2012, from http://techcrunch.com/2010/08/25/scvngr-game-mechanics/

Shebilske, W. L., Goettl, B. P., Corrington, K., & Day, E. A. (1999). Interlesson spacing and task-related processing during complex skill acquisition. *Journal of Experimental Psychology: Applied, 5,* 413–437.

Spencer, J. P., & Hund, A. M. (2002). Prototypes and particulars: Geometric and experience–dependent spatial categories. *Journal of Experimental Psychology: General, 131,* 16–37.

Spencer, J. P., & Hund, A. M. (2003). Developmental continuity in the processes that underlie spatial recall. *Cognitive Psychology, 47,* 432–480.

Stephen, D. G., Boncoddo, R. A., Magnuson, J. S., & Dixon, J. A. (2009). The dynamics of insight: Mathematical discovery as a phase transition. *Memory & Cognition, 37,* 1132–1149.

Stevenson, R. J., Sundqvist, N., & Mahmut, M. (2007). Age–related changes in discrimination of unfamiliar odors. *Perception & Psychophysics, 69,* 185–192.

Stokes, P. D., & Balsam, P. (2001). An optimal period for setting sustained variability levels. *Psychonomic Bulletin & Review, 8,* 177–184.

Stokes, P. D., & Harrison, H. M. (2002). Constraints have different concurrent effects and aftereffects on variability. *Journal of Experimental Psychology: General, 131,* 552–566.

Talvitie, E., Singh, S. (2009). Simple local models for complex dynamical systems. In D. Koller, D. Schuurmans, Y. Bengio, L. Bottou (eds), *Advances in Neural Information Processing Systems* 21 (1617–1624).

Thibaut, J.-P., French, R., & Vezneva, M. (2010). Cognitive load and semantic analogies: Searching semantic space. *Psychonomic Bulletin & Review, 17,* 569–574.

Toppino, T. C., Fearnow-Kenney, M. D., Kiepert, M. H., & Teremula, A. C. (2009). The spacing effect in intentional and incidental free recall by children and adults: Limits on the automaticity hypothesis. *Memory & Cognition, 37,* 316–325.

Usher, M., & McClelland, J. L. (2001). On the time course of perceptual choice: The leaky competing accumulator model. *Psychological Review, 108,* 550–592.

van der Linden, D., & Eling, P. (2006). Mental fatigue disturbs local processing more than global processing. *Psychological Research, 70,* 395–402.

Vul, E., Frank, M. C., Alvarez, G. A., Tenenbaum, J. B. (2010). Explaining human multiple object tracking as resource-constrained approximate inference in a dynamic probabilistic model. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, A. Culotta (eds), *Advances in Neural Information Processing Systems* 22 (1955–1963).

Wade, T., & Holt, L. L. (2005). Incidental categorization of spectrally complex non–invariant auditory stimuli in a computer game task. *Journal of the Acoustical Society of America, 118,* 2618–2633.

Washburn, D. A. (2003). The games psychologists play (and the data they provide). *Behavior Research Methods, Instruments, & Computers, 35,* 185–193.

Washburn, D. A., & Gulledge, J. P. (1995). Game–like tasks for comparative research: Leveling the playing field. *Behavior Research Methods, Instruments, & Computers, 27,* 235–238.

Wasserman, E. A., & Castro, L. (2005). Surprise and change: Variations in the strength of present and absent cues in causal learning. *Learning & Behavior, 33,* 131–146.

Weidemann, C. T., Mollison, M. V., & Kahana, M. J. (2009). Electrophysiological correlates of high–level perception during spatial navigation. *Psychonomic Bulletin & Review, 16,* 313–319.

Williams, K. D., & Jarvis, B. (2006). Cyberball: A program for use in research on interpersonal ostracism and acceptance. *Behavior Research Methods, 38,* 174–180.

Williams, P., Nesbitt, K., Eidels, A., Elliott, D. (2011). *Balancing risk and reward to develop an optimal hot hand game*. Game Studies, 11, online.

Wilson, K. A., Bedwell, W. L., Lazzara, E. H., Salas, E., Burke, C. S., Estock, J. L., …, Conkey, C. (2009). Relationships between game attributes and learning outcomes. *Simulation & Gaming*, *40*, 217–266.

Wood, R. T. A., Griffiths, M. D., Chappell, D., & Davies, M. N. O. (2004). The structural characteristics of video games: A psycho-structural analysis. *Cyberpsychology & Behavior, 7,* 1–10.

Yi, S. K. M., Steyvers, M., & Lee, M. (2009). Modeling human performance in restless bandits with particle filters. *Journal of Problem Solving, 2,* 81–101.

Yildirim, S., Narayanan, S., & Potamianos, A. (2011). Detecting emotional state of a child in a conversational computer game. *Computer Speech and Language, 25,* 29–44.

Yuzawa, M. (2001). Effects of word length on young children's memory performance. *Memory & Cognition, 29,* 557–564.