Computers in Human Behavior 29 (2013) 2388–2393

Contents lists available at SciVerse ScienceDirect

Computers in Human Behavior

journal homepage: www.elsevier.com/locate/comphumbeh

Exploring the effectiveness of commercial and custom-built games for cognitive training

Shamus P. Smith [a,c,*], Marina Stibric [b,d], David Smithson [c]

[a] School of Electrical Engineering and Computer Science, The University of Newcastle, Australia
[b] Faculty of Psychology, University of Bergen, Norway
[c] School of Engineering and Computing Sciences, Durham University, UK
[d] Department of Psychology, Durham University, UK

ARTICLE INFO

Article history:

Keywords:
Cognitive training
Commercial software
N-back task
Fluid intelligence
Game reuse
User study

ABSTRACT

There is increasing interest in quantifying the effectiveness of computer games in non-entertainment domains. We have explored general intelligence improvements for participants using either a commercial-off-the-shelf (COTS) game, a custom do-it-yourself (DIY) training system for a working memory task or an online strategy game to a control group (without training). Forty university level participants were divided into four groups (COTS, DIY, Gaming, Control) and were evaluated three times (pre-intervention, post-intervention, 1-week follow-up) with three weeks of training. In general intelligence tests both cognitive training systems (COTS and DIY groups) failed to produce significant improvements in comparison to a control group or a gaming group. Also neither cognitive training system produced significant improvements over the intervention or follow-up periods.

1. Introduction

With the ever-rising popularity of commercial computer games, there is increasing interest in quantifying the degree to which such systems can be put to use in other areas, such as patient therapy (Robillard, Bouchard, Fournier, & Renaud, 2003), simulation (Lewis, Wang, & Hughes, 2007), training systems (Smith & Trenholme, 2009) and for research tools in basic and applied sciences (Frey, Hartig, Ketzel, Zinkernagel, & Moosbrugger, 2007; McMahan, Ragan, Leal, Beaton, & Bowman, 2011). The attraction of re-applying commercial game titles as experiment software (McMahan et al., 2011) comes from the relative ease with which a working, highly optimised game or game engine can be obtained and put to use at little cost (Bottino, Ferlino, Ott, & Tavella, 2007; Frey et al., 2007; Jayakanthan, 2002; Trenholme & Smith, 2008). In some cases, commercial computer games can be used 'as-is' without modification (McMahan et al., 2011; Whitlock, McLaughlin, & Allaire, 2012) as has been seen in a multitude of studies relating to cognitive and visuo-motor training (Bavelier, Dye, & Green, 2009; Green & Bavelier, 2006; Dye, Green, & Bavelier, 2009).

Very little existing literature, however, attempts to compare the effectiveness of commercial off-the-shelf (COTS) and custom-built (do-it-yourself or DIY) software when applied to the same task. In the current work, we have investigated the use of commercial and custom-built systems for cognitive training as there is ongoing debate in the effectiveness of such systems (Owen et al., 2010; Rabipour & Raz, 2012). If the widespread deployment of cognitive training systems was to be considered, any savings from reusing commercial software would need to be balanced against evidence of effectiveness.

As an example of commercial reuse for cognitive training, Hsiung, Kupferschmidt, Naus, Feldman, and Jacova (2009) applied *Brain Age*, a Nintendo DS title to 12 elderly participants with MCI (mild cognitive impairment; the precursor to dementia), with the results proving inconclusive as to the positive effects of the software. Similarly, Basak, Boot, Voss, and Kramer (2008) investigated the effects of training in a commercial real-time strategy title on patients' cognitive decline, reporting, however that participants trained on the software did show improvement in measures of working memory and task switching. Alternatively, new, custom-built software is often developed to test and engage particular cognitive processes or participant responses, for example to investigate sound and colour responses (Wolfson & Case, 2000), cognitive processing speed (Ball, Edwards, & Ross, 2007) and working memory (Jaeggi, Buschkuehl, Jonides, & Perrig, 2008).

Here, we have focused on cognitive training as represented in (i) the Nintendo DS title *Dr. Kawashima's Brain Training*, also known as *Brain Age*, the most widely used recreational software in this domain which claims to improve users' mental abilities through a series of semi-randomised puzzle games, and (ii) a custom-built training system aimed at exercising a user's working memory.

* Corresponding author. Address: School of Electrical Engineering and Computer Science, The University of Newcastle, Callaghan, NSW 2308, Australia. Tel.: +61 2 49216175.

E-mail address: shamus.smith@newcastle.edu.au (S.P. Smith).

0747-5632/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.chb.2013.05.014

Current research strongly suggests that the link between the engagement of working memory and improvement of fluid intelligence[1] test scores is highly tractable to tests of this scope and type. Several studies report a high degree of correlation between the two processes (Bühner, Kröner, & Ziegler, 2008; Jaeggi et al., 2008; Süß, Oberauer, Wittmann, Wilhelm, & Schulze, 2002) both in terms of working memory function as a predictor of fluid intelligence, and as a medium for its improvement. Similar to the puzzles in *Dr. Kawashima's Brain Training*, working memory can be engaged with any standard short-term memory puzzle such as recalling a list of words, shapes or images while processing these or other information. Our study uses the simpler *n-back* paradigm, in which participants must recall the nth term back in a sequence of stimuli, as used in (Jaeggi et al., 2008) to train working memory, and identified in (Shelton, Elliott, Hill, Calamia, & Gouvier, 2009) as not only one of the best measures of working memory available, but also one of the most reliable predictors of fluid intelligence.

We conducted a study to compare commercial and custom-build software developed to support cognitive training. In particular we examined changes in general intelligence before and after a three week intervention period and one week after the intervention period. We expected that (a) there would be performance improvements for participants actively using training software, either commercial or custom-build, in comparison to any control groups and (b) that there would be increased improvement for participants using the custom-built training software, e.g. as in (Jaeggi et al., 2008).

2. Developed custom training system

We developed training software that makes use of the n-back working memory paradigm. A n-back test involves the presentation of a series of stimuli such as letters, images or sounds at a predetermined rate or time interval, and a user is required to respond via an input device such as a computer keyboard, when the current entry in a sequence is the same as the entry *n* steps ago.

For every entry in the sequence, there are four relevant possibilities:

- a match – the n-back condition is TRUE and the user DOES respond.
- a miss – the n-back condition is TRUE but the user DOES NOT respond.
- a false match – the n-back condition is FALSE but the user DOES respond.
- a correct rejection – the n-back condition is FALSE and the user DOES NOT respond.

A more complex variation of the n-back paradigm, known as parallel or *dual n-back*, employs two separate sequences of different stimuli concurrently. For dual n-back, users are shown an image and played a sound in set sequences, and are able to respond on two separate inputs in the usual manner – one for when the n-back condition holds for the image stimulus, and another for the sound stimulus (see Fig. 1). Using two independent sequences of stimuli is considered a highly challenging test of working memory (Jaeggi et al., 2008, Jaeggi, Buschkuehl, Perrig, & Meier, 2010).

The dual n-back with image and sound stimuli is used in the training software we have developed and supports recording and maintaining users' performance data. The software automatically adjusts the level of n, the main factor determining the difficulty of the test, based on a user's performance in a set of trials. Thus the software maintains a level of challenge, and supports cognitive engagement, of the test (Jaeggi et al., 2008). Our dual n-back system was developed using Game Maker 7.0,[2] a rapid prototyping tool for the development of graphical applications for the Windows platform. Fig. 2 shows an example screen of the dual n-back training software. Audio matches are noted by the user by pressing the "L" key, while visual matches, the position of the white square around the centre cross, are noted by pressing the "A" key, on a standard computer keyboard.

3. Intelligence measures

A reliable external measure for fluid intelligence against which to test performance is required to assess the transfer of skill from use of any software intervention to cognitive functions. Raven's Progressive Matrices (RPM) (Raven, 2000) is a widely recognised test of general intelligence and is the measure used in this study. The RPM is made up of a series of diagrams with a part missing. The subject being tested is expected to select the correct part from examples given in order to complete the pattern (see Fig. 3 for an example). Progressive examples are designed to be increasingly difficult.

Jaeggi et al. (2008) used RPM as an external measure for working-memory training with an n-back program, similar to the method employed here. Shelton et al. (2009) also used a computer-administered RPM test in evaluating the effectiveness of working memory measures. Further support for the reliability of RPM can be found in (Williams & McCord, 2006) who report the equivalence of digital and hard-copy versions of the test.

For this study we have split the Advanced Progressive Matrices Set I and II into three tests for pretest (week 0), posttest (week 3) and delayed posttest (week 4) measures. Each test had 16 questions, was unique and had similar progressive difficulty through the test examples. Our main measures are therefore the final scores from the three tests. The maximum score for each test was 16 and a higher score indicates better performance.

4. Experimental procedure

We used an experimental-plus-control group design with two cognitive training groups and two control groups. Participants were randomly assigned to one of four groups and balanced for gender:

- COTS group were provided with a Nintendo DS game console and a copy of the *Dr. Kawashima's Brain Training* game.
- DIY group were provided with DELL notebooks running Windows XP with our custom dual n-back software installed.
- Gaming group were provided with a web link for *Phage Wars*, a web-based strategy game with no explicit working memory training-related content.[3] This game has eight variants and an expert player can complete one variant in ~20 minutes. This provided the Gaming group with enough activity variability over the evaluation period. The Gaming group will allow us to determine whether it is the specific cognitive training exercises that are beneficial in contrast to just a regular focused game activity.
- Control group were a no contact group.

[1] Research into general cognition (Gray & Thompson, 2004) increasingly points towards the existence of a concept termed *fluid intelligence* that embodies all transferable, task-independent reasoning/problem-solving abilities and their level of performance. This process is itself strongly related to the construct of *working memory* and differs from *crystallised intelligence* that represents the accumulation of general knowledge and experience.

[2] Available from YoYo Games: http://www.yoyogames.com/gamemaker/studio [Last access: 20/02/2013].

[3] Available at http://armorgames.com/play/2675/phage-wars [last access [20/02/2013].

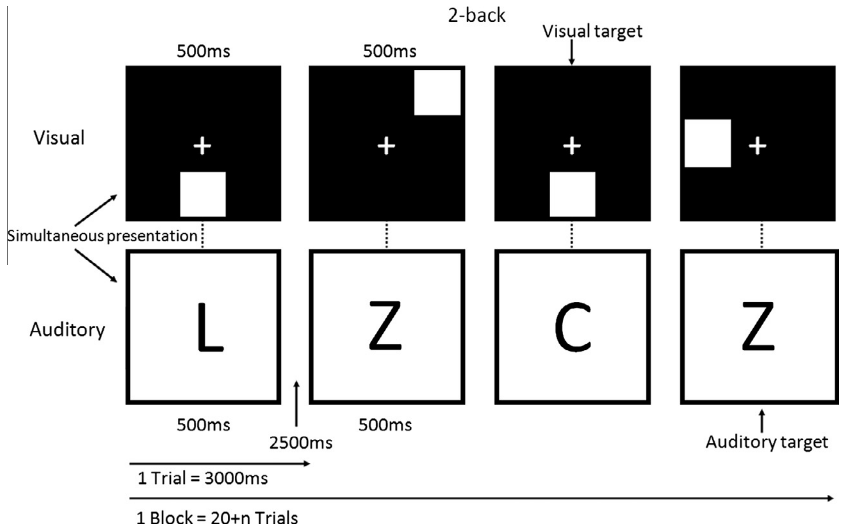

Fig. 1. Dual n-back training task with $n = 2$, i.e. 2-back condition. The letters were presented as spoken sound clips at the same rate as the spatial material is presented visually. Adapted from Jaeggi et al. (2008).

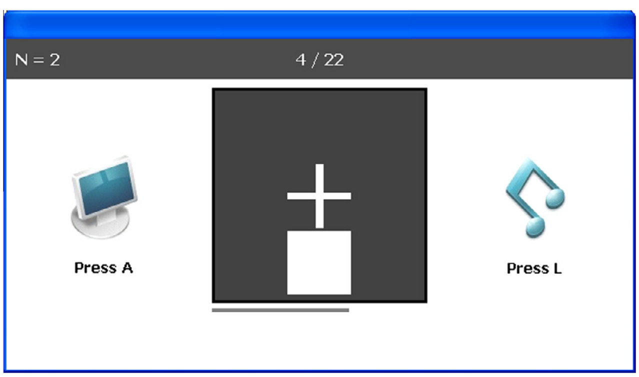

Fig. 2. Example screen from our dual n-back software. The value of N is displayed on the top left of the screen and the trial number and total number of trials is displayed in the top centre. The bar below the black square indicates the time left in the current trial.

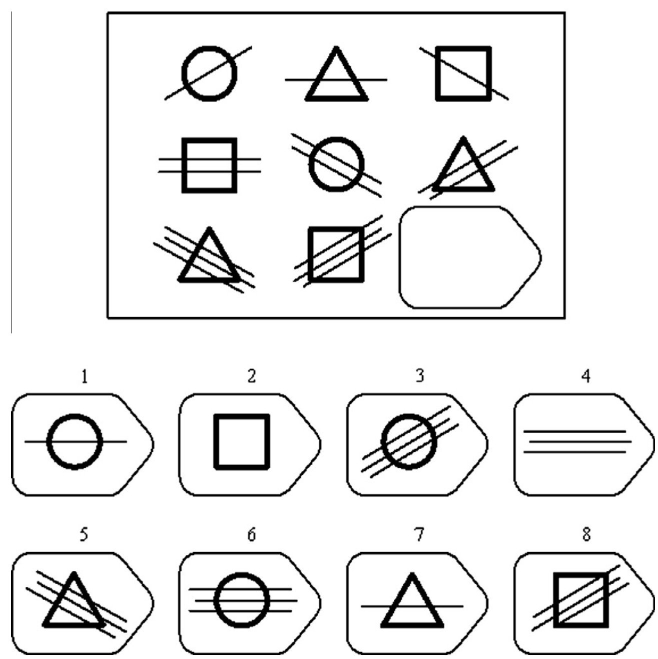

Fig. 3. Simulated item similar to those in the Raven's Progressive Matrices – Advanced Progressive Matrices.

Participants in the COTS, DIY and Gaming groups were requested to use their respective software for 20 minutes per day for at least 17 days of a 21 day evaluation period. Jaeggi et al. (2008) demonstrated that 17 days of n-back training was sufficient to observe significant participant differences and we have used this as our minimum benchmark for training. Our custom dual n-back software collected usage logs and performance metrics, for example session based n-back scores. These were used to validate training usage for the DIY participants. *Dr. Kawashima's Brain Training* software on the DS consoles collects usage data by day only, e.g. the days used for training are marked on a calendar. After the notebooks and DS consoles were returned, they were examined to verify that sufficient training sessions had been logged. Any participants without 17 days training were excluded from further participation in the study. Participant data was erased from devices between participants. Participants in the Gaming group were requested to keep a paper log of training days. This was also inspected at the end of the training period.

Forty university students (20 female, age range 18–34, evenly split over the four groups) completed the evaluation period. Participants were recruited from University notice board leaflets and an undergraduate participant pool from the Psychology Department. Psychology students received course credit for participating and all participants were entered into a prize draw at the end of the study.

All participants were students of a British university and were fluent at graduate level academic English. All participants used computers on a daily basis (on a rating of [Daily, Weekly, Monthly, Less frequently or Never]). Participants across groups also indicated similar levels of computer game experience (rated by [Master, Expert, Intermediate, Novice, None]) and computer game usage (rated by [Daily, Weekly, Monthly, Less frequently, Never]).[4]

[4] Computer game experience and usage did not differentiate between the groups, $F_{(3,38)} = 0.853$, $p = 0.474$ for computer game experience and $F_{(3,38)} = 0.474$, $p = 0.702$ for computer game usage.

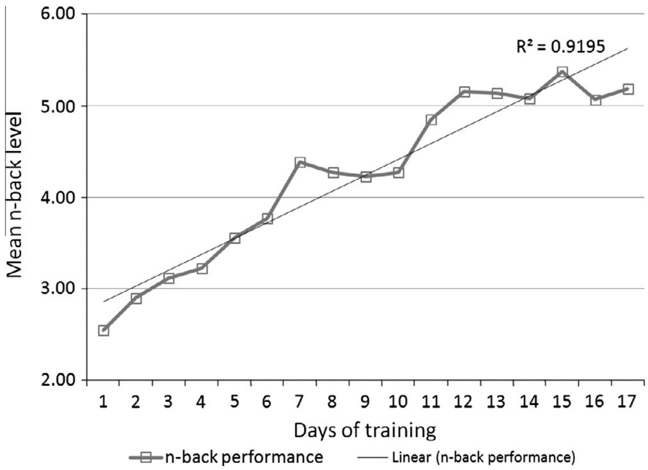

Fig. 4. Performance increase in the n-back training task. For each day/session, the mean level of *n* achieved by the DIY participants is presented. The level of *n* changes based on the participants' performance. Training gain is explained by a linear function: $R^2 = 0.9195$.

5. Results

Analysis of the training performance[5] for the DIY (i.e. n-back trained group) indicated that the participants improved their performance on the working memory task (See Fig. 4). The participants increased performance corresponds to a linear function, $R^2 = 0.9195$. This is comparable to the performance improvement results found with 17 days of n-back training by Jaeggi et al. (2008) (also explained with a linear function, $R^2 = 0.73$).

A repeated measures analysis of variance[6] with the between-subjects factor training intervention (Control, Gaming, COTS, DIY) and the within-subjects factor time point (pretest [week 0], posttest [week 3] and delayed posttest [week 4]) was conducted. The dependent variable was the RPM score at each time point. After screening for outliers, one participant in the control group was identified with a pretest RPM score outside of the threshold of −3 standard deviations from the mean. This participant was excluded from further analysis.

There was a significant interaction effect between group and test times (Wilks Lambda = 0.686, $F_{(6,70)} = 2.831$, $p = 0.016$, partial eta squared = 0.195). Therefore we examined differences between each testing time by group using pairwise comparisons with Bonferroni correction.

- Control group ($n = 9$): There were no significant differences between tests, pretest–posttest ($p = 0.352$), pretest-delayed posttest ($p = 0.762$), and posttest-delayed posttest ($p = 1$).
- Gaming group ($n = 10$): There were no significant differences between tests, pretest–posttest ($p = 1$) and pretest-delayed posttest ($p = 0.077$). However between posttest-delayed posttest the difference is significant ($p = 0.017$).
- COTS group ($n = 10$): There were no significant differences between tests, pretest–posttest ($p = 1$), posttest-delayed posttest ($p = 1$) and pretest-delayed posttest ($p = 0.837$).
- DIY group ($n = 10$): There were no significant differences between tests, pretest–posttest ($p = 1$), posttest-delayed posttest ($p = 0.412$) and pretest-delayed posttest ($p = 0.182$).

The posthoc tests showed that the only significant change in the RPM score was between the posttest and delayed posttest in the Gaming group. A summary of the average RPM mean scores can be seen in Table 1 and the RPM scores by group can be seen in Fig. 5.

6. Discussion

In the RPM tests both cognitive training systems (COTS/DIY groups) failed to produce significant improvements in comparison to the Control group or the Gaming group. The lack of significant improvements in the DIY group is particularly interesting. The dual n-back task and method applied was modelled on (Jaeggi et al., 2008) where statistically significant training differences were observed after 17 days, i.e. our minimum training requirement. One difference here was the location of the training. Jaeggi et al. (2008) used controlled experimental conditions for the training and were interested in identifying gains in intelligence over different training periods, namely 8, 12, 17 and 19 days of training. In comparison, our method only enforced a minimum training regime, at least 17 days training. Also participant training was done outside a controlled environment, i.e. in-the-field. Uncontrolled environments are an increasingly common problem when evaluating systems in natural settings (Rogers, Sharp, & Preece, 2011, p. 440). However, for studies over a number of weeks, it is difficult to control for all external factors, particularly when factor identification would require self monitoring by participants, which is itself a confounding factor (Wu & Clark, 2003). We therefore enforced minimum requirements for the training groups that we could validate, e.g. using automatic training logs on the COTS and DIY systems. However, the lack of controlled training conditions may have impacted the amount of time needed for significant improvements in the dual n-back training.

The participants in the DIY group did improve their performance in the n-back working memory training (see Fig. 4) which indicates that their working memory capacity increased. Also their RPM scores from test to test (see Fig. 5) improved. However the increased working memory capacity did not transfer to an increase in fluid intelligence (similar to results found by Redick et al. (2012)) within the study period. Therefore a useful follow-on study would be to compare training time requirements in environments that approximate real-life situations, e.g. with high *ecological validity*, as noted by McMahan et al. (2011) when considering controlled experiments with commercial computer games. For example, this might include the unsupervised use of cognitive training systems at home. However, caution is required as such *recreational training* may not be sufficiently specific or sufficiently long to show performance gains (Lorant-Royer, Munch, Mesclé, & Lieury, 2010).

There was also a lack of RPM score improvement for the COTS group. This is particularly problematic given the stated aims of the COTS system, i.e. a consumer product for brain training. This contributes to the ongoing debate on the benefits of cognitive training systems in general (see Owen et al., 2010; Rabipour & Raz, 2012) and the use of computer game based training systems. There is increasing evidence in the limited ability for commercial games to stimulate cognitive aptitudes. For example, a recent study (Lorant-Royer et al., 2010) noted weak positive effects in student's visuospatial, visuomotor and attentional capabilities after training with *Dr. Kawashima's Brain Training* or *New Super Mario Bros* game software. However, Rabipour and Raz (2012) observe that studies that do show improvements highlight that long-term exposure and application of training is needed for lasting results, i.e. needing a level of system accessibility that can be provided by commercial products. Thus the challenge may be in the selective use of particular game modes in commercial products to maximise any required benefits.

[5] Session performance data including n-back scores was automatically collected by our dual n-back system.
[6] All statistical results were generated with SPSS Statistics 20.

Table 1
RPM test scores for the four groups across three time periods.

Participant group	RPM test	Mean	Std deviation
Control (*n* = 9)	Pretest	10.200	1.69
	Posttest	11.900	1.58
	Delayed posttest	11.300	2.12
Gaming (*n* = 10)	Pretest	11.800	2.15
	Posttest	12.000	2.45
	Delayed posttest	10.300	3.06
COTS (*n* = 10)	Pretest	12.000	2.10
	Posttest	12.300	3.02
	Delayed posttest	12.300	1.83
DIY (*n* = 10)	Pretest	11.200	2.44
	Posttest	11.500	2.99
	Delayed posttest	12.600	2.41

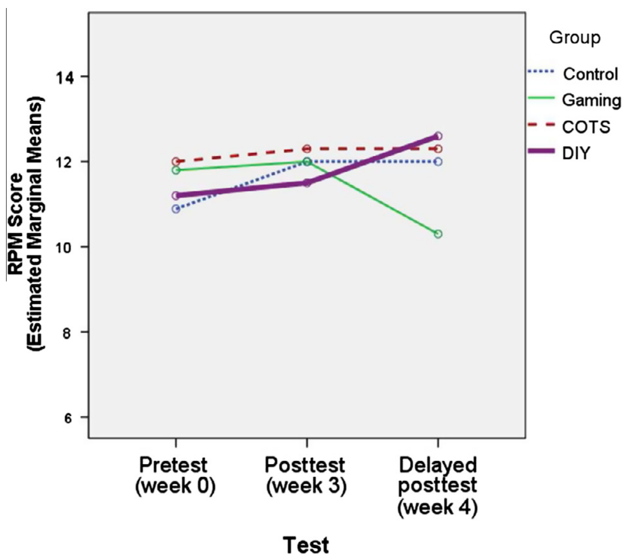

Fig. 5. RPM scores by group (Control, Gaming, COTS and DIY) at three testing times, pretest (week 0), posttest (week 3) and delayed posttest (week 4).

One interesting feature in Fig. 5 is the drop off in the final test scores (delayed posttest) for the Gaming group in comparison to the COTS and DIY scores, which were, respectively, stable or increasing. Also, for the Gaming group the difference from posttest to delayed posttest was statistically significant ($p = 0.017$). One explanation is that the COTS and DIY scores are a result of motivational factors. Green and Bavelier (2008) observe that improvement in performance is not always due to training-induced learning. Motivation can lead to temporary improvements in performance, or as Rabipour and Raz (2012) note, a *placebo* effect. All our study participants were aware that they were participating in a *cognitive training* experiment, as they were recruited for this task, read study instruction sheets and signed consent forms. The COTS and DIY participants were given obvious cognitive training systems to use and may have had expectations of cognitive benefit in the training. The Gaming group may have perceived no personal benefit from study participation and there was an overhead of daily activity. This may have negatively effected their performance in the final test e.g. a Hawthorne effect (Lied & Karzandjian, 1998). Again, not having significant improvements in the training groups makes this possible confound difficult to validate.

When considering results of this and similar studies, sample size should also be taken into account. Other cognitive training studies have had participant groups from extremely large, e.g. 11,430 participants in (Owen et al., 2010), to numbers similar to our study, e.g. 12 participants with mild cognitive impairment and 2 control participants in (Hsiung et al., 2009) and 16 participants in the 17 day training group in (Jaeggi et al., 2008). Further studies with larger sample size could generate stronger evidence of the transfer effects across the different types of cognitive training.

7. Conclusion

This paper presents a study comparing the effectiveness of using a commercial off-the-shelf game as opposed to developing dedicated software for experiments into cognitive training effects. Examples of training software were compared to control and general gaming groups. In general intelligence tests, utilising RPM, both cognitive training systems (COTS/DIY groups) failed to produce significant improvements in comparison to the Control group or the Gaming group. Also neither cognitive training system produced significant improvements over the intervention or follow-up periods.

We have demonstrated the limited effect cognitive training software can have on university age participants, even with tasks that have previously shown positive improvements (Jaeggi et al., 2008). This suggests caution in the over generalisation on the effectiveness of brain training systems with results from other demographic groups, for example school children (Robertson & Miller, 2009) or older adults (Basak et al., 2008; Whitlock et al., 2012).

However, we have identified a number of potentially confounding factors, beyond the contentious nature of brain training itself (Owen et al., 2010; Rabipour & Raz, 2012), that may have impacted our study. Firstly, although the training environments were uncontrolled, e.g. in-the-field, they were representative of the target deployment environments, e.g. with increased ecological validity (McMahan et al., 2011). Therefore there is scope to explore the impact that real world deployments have on both COTS and DIY solutions. Secondly, drop off effects in the Gaming group scores and increased (DIY) and stable (COTS) scores may have been affected by participant expectations on the value of participation in the study. Controlling any perceived lack of scientific rigor in testing conditions will be important if genuine comparison measures are to be collected. Thirdly, although our intervention period was guided by previous work (Jaeggi et al., 2008), it is possible that more training time and a larger participant pool are required to show improvements when using COTS cognitive products or when training is conducted in uncontrolled environments. Exploring these issues are all fruitful areas for future work.

Acknowledgements

This work was partially funded by the Wolfson Research Institute for Health and Wellbeing (Durham University). The authors wish to thank the participants for their perseverance with the training activities and Emma Mercier and Markus Hausmann for comments on an earlier draft of this article.

References

Ball, K., Edwards, J. D., & Ross, L. A. (2007). The impact of speed of processing training on cognitive and everyday functions. *Journals of Gerontology, 62*(1), 19–31.
Basak, C., Boot, W. R., Voss, M. W., & Kramer, A. F. (2008). Can training in a real-time strategy computer game attenuate cognitive decline in older adults? *Psychology and Aging, 23*, 765–777.
Bavelier, D., Dye, M., & Green, C. (2009). Exercising your brain: Training-related brain plasticity. In M. Gazzaniga (Ed.), *The cognitive neurosciences. Section plasticity* (Vol. IV, pp. 153–164). Lawrence MIT Press.
Bottino, R. M., Ferlino, L., Ott, M., & Tavella, M. (2007). Developing strategic and reasoning abilities with computer games at primary school level. *Computers & Education, 49*, 1272–1286.

Bühner, M., Kröner, S., & Ziegler, M. (2008). Working memory, visual-spatial-intelligence and their relationship to problem-solving. *Intelligence, 36*(6), 672–680.

Dye, M. W., Green, C. S., & Bavelier, D. (2009). The development of attention skills in action video game players. *Neuropsychologia, 47*(8–9), 1780–1789.

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 Behaviour, 23*, 2026–2039.

Gray, J. R., & Thompson, P. M. (2004). Neurobiology of intelligence: Science and ethics. *Nature Reviews Neuroscience, 5*, 471–482.

Green, C. S., & Bavelier, D. (2006). Enumeration versus multiple object tracking: The case of action video game players. *Cognition, 101*, 217–245.

Green, C. S., & Bavelier, D. (2008). Exercising your brain: A review of human brain plasticity and training-induced learning. *Psychology and Aging, 23*(4), 692–701.

Hsiung, G.-Y. R., Kupferschmidt, A., Naus, K., Feldman, H. H., & Jacova, C. (2009). Cognitive and clinical outcomes in a pilot study using a handheld computer game for cognitive training of patients with Mild Cognitive Impairment (MCI). *Alzheimer's & Dementia: The Journal of the Alzheimer's Association, 5*(4), P411.

Jaeggi, S. M., Buschkuehl, M., Jonides, J., & Perrig, W. J. (2008). Improving fluid intelligence with training on working memory. *Proceedings of the National Academy of Sciences (PNAS), 105*(19), 6829–6833.

Jaeggi, S. M., Buschkuehl, M., Perrig, W. J., & Meier, B. (2010). The concurrent validity of the *N*-back task as a working memory measure. *Memory, 18*(4), 394–412.

Jayakanthan, R. (2002). Application of computer games in the field of education. *The Electronic Library, 20*(2), 98–102.

Lewis, M., Wang, J., & Hughes, S. (2007). USARSim: Simulation for the study of human-robot interaction. *Journal of Cognitive Engineering and Decision Making, 1*(1), 98–120.

Lied, T. R., & Karzandjian, V. A. (1998). A Hawthorne strategy: Implications for performance measurement and improvement. *Clinical Performance and Quality Healthcare, 6*(4), 201–204.

Lorant-Royer, S., Munch, C., Mesclé, H., & Lieury, A. (2010). Kawashima vs "Super Mario"! Should a game be serious in order to stimulate cognitive aptitudes? *Revue Européenne de Psychologie Appliquée/European Review of Applied Psychology, 60*(4), 221–232.

McMahan, R. P., Ragan, E. D., Leal, A., Beaton, R. J., & Bowman, D. A. (2011). Considerations for the use of commercial video games in controlled experiments. *Entertainment Computing, 2*(1), 3–9.

Owen, A. M., Hampsihire, A., Grahn, J. A., Stenton, R., Dajani, S., Burns, A. S., et al. (2010). Putting brain training to the test. *Nature, 465*, 775–779.

Rabipour, S., & Raz, A. (2012). Training the brain: Fact and fad in cognitive and behavioral remediation. *Brain and Cognition, 79*(2), 159–179.

Raven, J. (2000). The Raven's Progressive Matrices: Change and stability over culture and time. *Cognitive Psychology, 41*(1), 1–48.

Redick, T. S., Shipstead, Z., Harrison, T. L., Hicks, K. L., Fried, D. E., Hambrick, D. Z., et al. (2012). No evidence of intelligence improvement after working memory training: A randomized, placebo-controlled study. *Journal of Experimental Psychology: General*. http://dx.doi.org/10.1037/a0029082.

Robertson, D., & Miller, D. (2009). Learning gains from using games consoles in primary classrooms: a randomized controlled study. *Procedia – Social and Behavioral Sciences, 1*(1), 1641–1644.

Robillard, G., Bouchard, S., Fournier, T., & Renaud, P. (2003). Anxiety and presence using VR immersion: A comparative study of the reactions of phobic and non-phobic participants in therapeutic virtual environments derived from computer games. *CyberPsychology & Behaviour, 6*(5), 467–475.

Rogers, Y., Sharp, H., & Preece, J. (2011). *Interaction design* (3rd ed.). UK: John Wiley and Sons.

Shelton, J. T., Elliott, E. M., Hill, B., Calamia, M. R., & Gouvier, W. D. (2009). A comparison of laboratory and clinical working memory tests and their prediction of fluid intelligence. *Intelligence, 37*(3), 283–293.

Smith, S. P., & Trenholme, D. (2009). Rapid prototyping a virtual fire drill environment using computer game technology. *Fire Safety Journal, 44*(4), 559–569.

Süß, H., Oberauer, K., Wittmann, W. W., Wilhelm, O., & Schulze, R. (2002). Working-memory capacity explains reasoning ability – And a little bit more. *Intelligence, 30*(3), 261–288.

Trenholme, D., & Smith, S. P. (2008). Computer game engines for developing first-person virtual environments. *Virtual Reality, 12*, 181–187.

Whitlock, L. A., McLaughlin, A. C., & Allaire, J. C. (2012). Individual differences in response to cognitive training: Using a multi-modal, attentionally demanding game-based intervention for older adults. *Computers in Human Behavior, 28*(4), 1091–1096.

Williams, J. E., & McCord, D. M. (2006). Equivalence of standard and computerized versions of the Raven progressive matrices test. *Computers in Human Behavior, 22*(5), 791–800.

Wolfson, S., & Case, G. (2000). The effects of sound and colour on responses to a computer game. *Interacting with Computers, 13*(2), 183–192.

Wu, K. D., & Clark, L. A. (2003). Relations between personality traits and self-reports of daily behavior. *Journal of Research in Personality, 37*(4), 231–256.