SUPPLEMENTARY MATERIAL

Action Video Games Do Not Improve The Speed of Information Processing in Simple Perceptual Tasks

Don van Ravenzwaaij[1], Wouter Boekel[2], Birte U. Forstmann[2], Roger Ratcliff[3], and Eric-Jan Wagenmakers[2]
[1]University of Newcastle [2]University of Amsterdam [3] Ohio State University

Correspondence concerning this article should be addressed to:
Don van Ravenzwaaij
University of Newcastle, Department of Psychology
University Drive, Aviation Building, room AVG11
Callaghan NSW 2308, Australia
Ph: (+61) 2–4921–5662
E–mail should be sent to don.vanravenzwaaij@newcastle.edu.au.

This document consists of three sections. The first contains the unpublished analyses of the data from the lexical decision task referred to in van Ravenzwaaij, Boekel, Forstmann, Ratcliff, and Wagenmakers (2013).[1] The second contains model predictives. Model predictives are a means to assess whether the model gives a good description of the data. The third section contains a power analysis of Experiment 2 from van Ravenzwaaij et al. (2013).

Lexical Decision Experiment

Material

In each of the six sessions, a lexical decision task was administered with 2 blocks of 200 trials each. We used the first 6 lists of the 11 lists constructed by Dutilh, Krypotos, and Wagenmakers (2011). The average frequency per million of the 1,200 words was 1.23 (SD = 0.81). Each list consisted of 40 four–letter words (mean frequency = 1.10, SD = 0.83), 80 five–letter words (frequency = 0.99, SD = 0.81), and 80 six–letter words (frequency = 1.54, SD = 0.69). For further details regarding stimulus–construction, please refer to Dutilh et al. (2011).

Participants responded by pressing one of two buttons on a response device attached to the computer's parallel port. Feedback and instruction to participants promoted accurate

[1]Data is available at `http://www.donvanravenzwaaij.com/Papers.html`.

We are indebted to Gilles Dutilh for providing us with the stimulus material for the lexical decision task.

but fast responding: responses slower than 2,000 ms were followed by the message "TE LANGZAAM" (i.e., "too slow"), and responses faster than 200 ms were followed by the message "TE SNEL" (i.e., "too fast"). Correct responses within the 200–2,000 ms window triggered no feedback at all, whereas error responses were followed by the message "FOUT" (i.e., "incorrect"). The duration of the feedback was 1,200 ms. Each trial started with a blank screen that was presented for 250 ms.

Behavioral Results

One participant withdrew from the experiment after the first session and was replaced. The data from session 3 from one participant was lost due to experimenter error. For each participant, we excluded all RTs below 275 ms, as these were likely to be guesses. This led to the exclusion of less than 0.01% of all RTs and did not affect the results.

Figure 1 shows the within–subject effects for mean RT and accuracy for word and for non–word stimuli. Across conditions, participants' mean RTs on word stimuli shortened in subsequent sessions, as confirmed by the presence of a negative linear trend over sessions ($F(1, 95) = 32.7$, $p < .001$, $BF_{01} = 4.4 \cdot 10^{-6}$). Thus, practice on the moving dot task resulted in faster responding. Importantly, this session effect for mean RT did not interact with gaming condition ($F(1, 95) = 0.0$, $p > .05$, $BF_{01} = 9.78$). From the first to the last session, the session effect on mean RT for word stimuli was 50 milliseconds for the action condition and 53 milliseconds for the cognitive condition.

Across conditions, participants' mean RTs on non–word stimuli shortened in subsequent sessions, as confirmed by the presence of a negative linear trend over sessions ($F(1, 95) = 54.2$, $p < .001$, $BF_{01} = 2.0 \cdot 10^{-9}$). Thus, practice on the moving dot task resulted in faster responding. Importantly, this session effect for mean RT did not interact with gaming condition ($F(1, 95) = 0.1$, $p > .05$, $BF_{01} = 9.30$). From the first to the last session, the session effect on mean RT for non–word stimuli was 96 milliseconds for the action condition and 89 milliseconds for the cognitive condition.

In addition to speeding up, participants also made more mistakes in subsequent sessions on word stimuli; there was a significant negative linear trend over sessions ($F(1, 95) = 8.0$, $p < .05$, $BF_{01} = 0.18$). There was no evidence for an interaction between session and gaming condition for accuracy on word stimuli ($F(1, 95) = 2.3$, $p > .05$, $BF_{01} = 3.01$). For the non–word stimuli, participants did not make more mistakes in subsequent sessions; there was no significant negative linear trend over sessions ($F(1, 95) = 0.6$, $p > .05$, $BF_{01} = 7.19$). There was no evidence for an interaction between session and gaming condition for accuracy on non–word stimuli ($F(1, 95) = 2.0$, $p > .05$, $BF_{01} = 3.50$).

In sum, practice on the lexical decision task decreased mean RT for both the action and the cognitive condition. Playing the action video game did not result in better performance compared to playing the cognitive video game. Response accuracy decreased slightly over sessions for the word stimuli, hinting at the possibility that participants became less cautious as they improve with practice (see also Dutilh, Wagenmakers, Vandekerckhove, & Tuerlinckx, 2009). In order to quantify the psychological factors that drive the observed effects we now turn to a diffusion model decomposition.

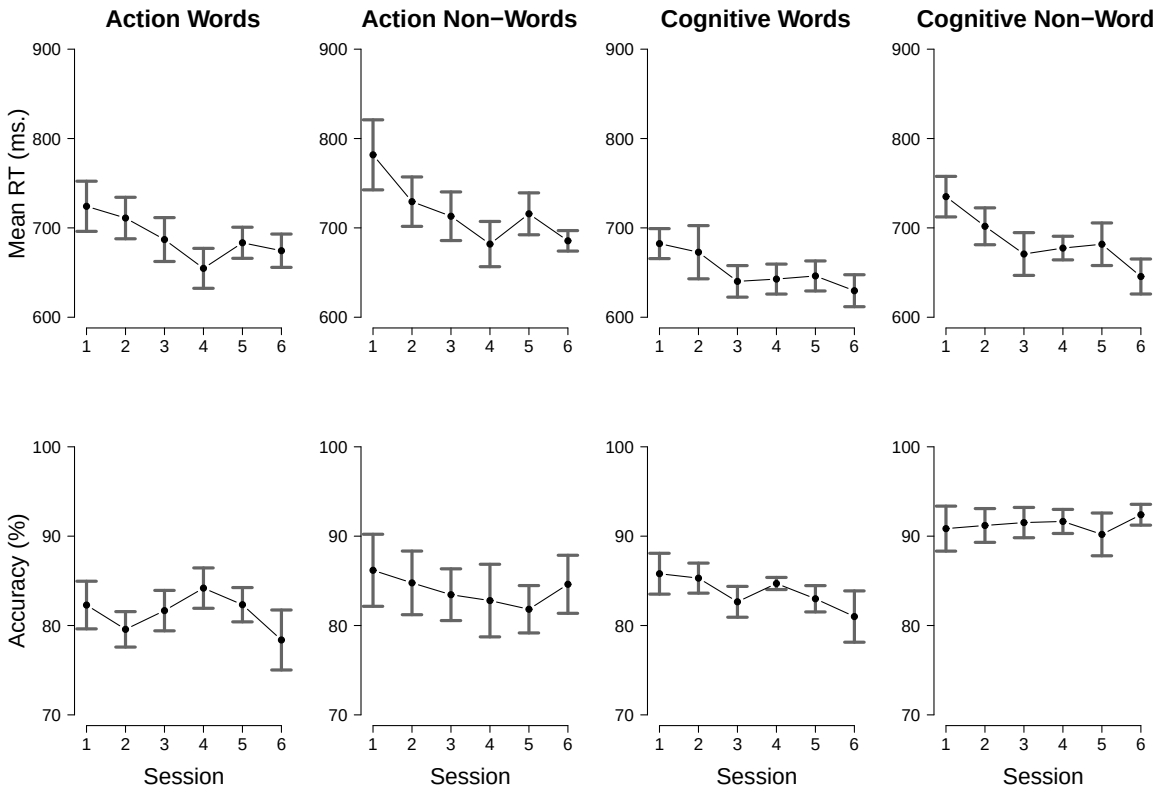

Figure 1. The within–subject effects for word and non–word stimuli in the action condition (left four panels) and for word and non–word stimuli in the cognitive condition (right four panels) on mean RT (top panels) and response accuracy (bottom panels) for the lexical decision task. Error bars represent 95% confidence intervals.

Diffusion Model Decomposition

The diffusion model was fit to the data using the DMAT software package (Vandekerckhove & Tuerlinckx, 2007), which minimizes a multinomial log–likelihood function. Each participant was fit separately. We allowed starting point z to vary freely to reflect possible a priori bias in favor of word or non–word responses. We estimated a separate mean word drift rate v_w, mean non–word drift rate v_{nw}, boundary separation a, and non–decision time T_{er} for each session. Furthermore, we constrained the standard deviation of drift rate η, range of starting point s_z, and range of non–decision time s_t to be equal across sessions.

Figure 2 shows the within–subject effects for word drift rate v_w, non–word drift rate v_{nw}, boundary separation a, and non–decision time T_{er}. Across conditions, the evidence for participants processing information on word stimuli faster in subsequent sessions is ambiguous ($F(1, 91) = 4.0$, $p < .05$, $BF_{01} = 1.26$). Importantly, there was no interaction between session and gaming condition for word drift rate ($F(1, 91) = 0.0$, $p > .05$, $BF_{01} = 9.69$). For non–word drift rate v_{nw}, participants did not process information faster in subsequent sessions ($F(1, 91) = 1.8$, $p > .05$, $BF_{01} = 3.79$). There was no evidence for an

interaction between session and gaming condition ($F(1, 91) = 0.1$, $p > .05$, $BF_{01} = 9.06$).

For boundary separation a, there was a linear trend over sessions across conditions, such that participants became less cautious in later sessions ($F(1, 91) = 62.4$, $p < .001$, $BF_{01} = 1.6 \cdot 10^{-10}$). There was no evidence for an interaction between session and gaming condition ($F(1, 91) = 0.6$, $p > .05$, $BF_{01} = 7.06$). For non–decision time T_{er} there was no evidence for the presence of a linear trend over sessions across conditions ($F(1, 91) = 0.6$, $p > .05$, $BF_{01} = 7.03$) and no evidence for an interaction between session and gaming condition ($F(1, 91) = 0.0$, $p > .05$, $BF_{01} = 9.57$).

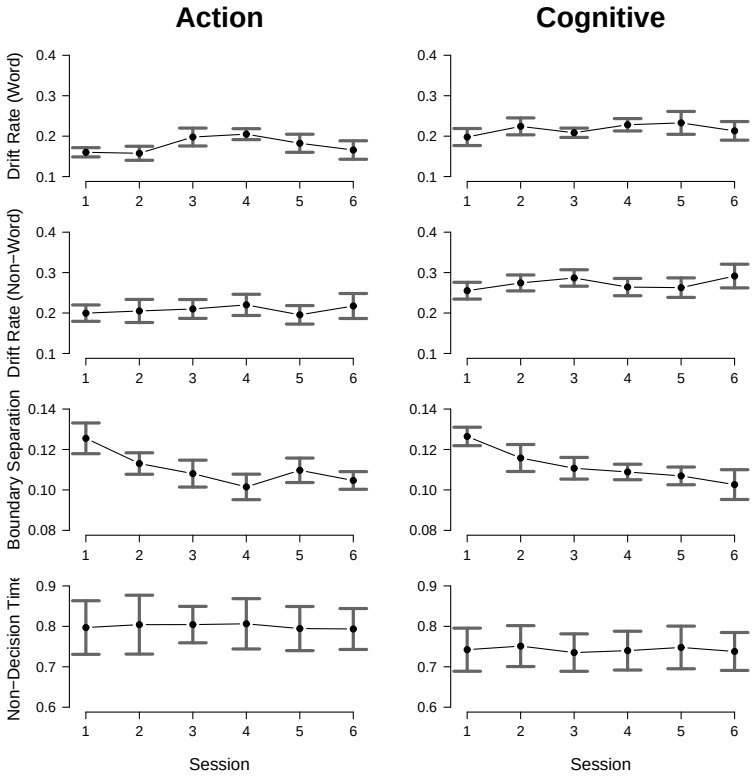

Figure 2. The within–subject effects of the action condition (top panels) and the cognitive condition (bottom panels) on drift rate v (left panels), boundary separation a (middle panels), and non–decision time T_{er} (right panels) for the lexical decision task from Experiment 1. Error bars represent 95% confidence intervals.

In sum, practice on an action video game did not lead to different diffusion model parameter estimates from playing a cognitive video game, as evidenced by the absence of any interaction effects. Hence, contrary to the results by Green, Pouget, and Bavelier (2010), playing action versus cognitive games did not yield a benefit on information processing.

Diffusion Model Fit

The adequacy of a model fit can be assessed by means of model predictives. Model predictives are simulated, synthetic, or fake data generated from the cognitive model with

parameter values that were obtained by fitting the real data. If the synthetic data generated from the model closely match the real data observed in the experiment, then the model fit is deemed adequate (e.g., Gelman & Hill, 2007, but see Bayarri & Berger, 1998).

To take into account parameter uncertainty, we drew 20 bootstrap samples from the real data set and generated diffusion model parameter estimates for each participant separately. We generated synthetic data using each set of parameter estimates per participant. To ensure that discrepancies between the real and synthetic data only reflect model misfit and no sampling error, we generated 10,000 trials per participant per condition. Then, we calculated the 0.1, 0.3, 0.5, 0.7, and 0.9 RT quantiles for each of the synthetic data sets and compared these to the RT quantiles for the real data. The real and synthetic RT quantiles, averaged across participants, are shown in Figure 3 (Experiment 1) and Figure 4 (Experiment 2).

Experiment 1

Figure 3 shows a quantile probability plot (e.g., Ratcliff, 2002; Ratcliff, Gomez, & McKoon, 2004), where the left–hand side represents error RTs for the five quantiles, and the right–hand side represents correct RTs for those same quantiles. The different symbol shapes indicate different sessions. The squares in the figure show the empirical data, and the circles show the simulated data that were generated using the DMAT parameter estimates for each bootstrap.

For response accuracy, the correspondence between the synthetic data and the empirical data can be judged by the horizontal disparity between the data points and the model points. The horizontal disparity between the squares and circles in Figure 3 shows that the diffusion model captured the error rate relatively poorly for the action condition and reasonably well for the cognitive condition. The relatively modest vertical disparity indicates that the diffusion model did capture the RTs well on average, though there was a discrepancy for some of the higher quantiles. This discrepancy for the high quantiles was largest for the error responses (shown in the left–hand side of each panel). There was a lot of uncertainty about the diffusion parameter estimates, as evidenced by the fact that the clouds of circles are very diffuse, rather than grouping closely together.

In sum, the model fit is acceptable for the correct RTs but off for the error rates and the slow error RTs. The misfit is likely due to the relatively low number of trials and the relatively low number of errors, a state of affairs that Experiment 2 sought to correct.

Experiment 2

Figure 4 shows that the diffusion model captured the data well. The small horizontal disparities between the squares and circles indicates that the model captured the error rates for all sessions in the action, cognitive, and control condition. The small vertical disparities between the squares and circles indicates that the model also captured the RT distributions, although for some of the highest quantiles slight discrepancies persist. Finally, the tightness of the circle clouds indicate that the model parameters were estimated with little uncertainty.

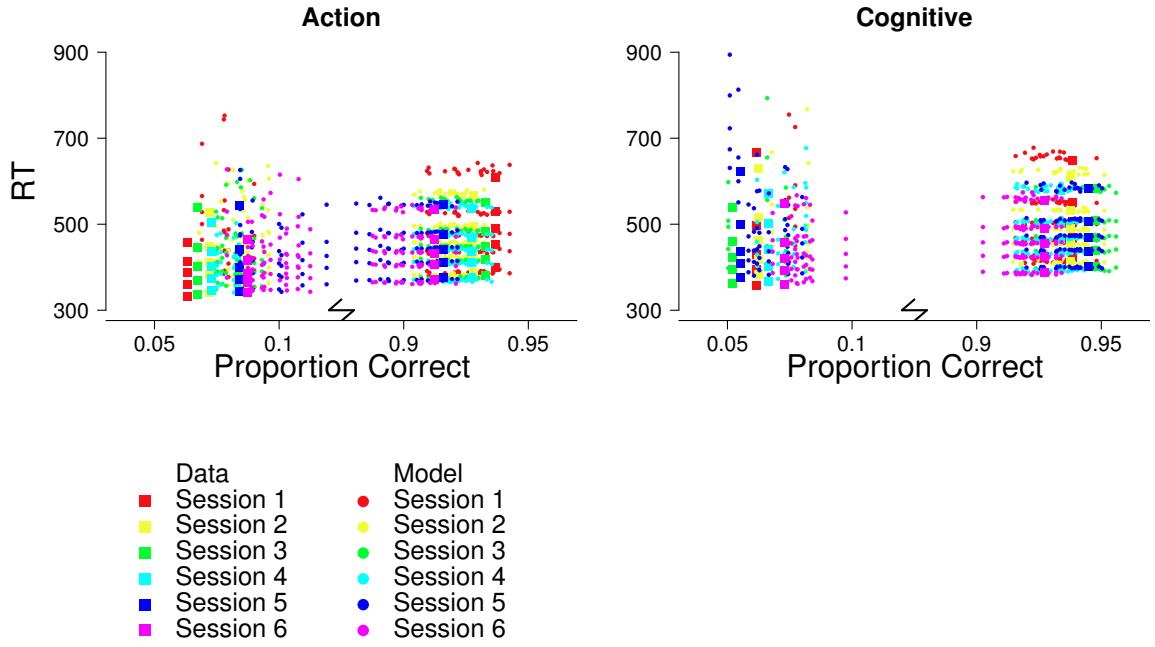

Figure 3. Comparison between synthetic data generated from the diffusion model and real data from Experiment 1. The model did not capture the error rates from the action game condition and also missed some of the higher RT quantiles, particularly for the error RTs. Squares = empirical data; circles = synthetic data, based on bootstrapped model parameter estimates.

Power Analysis Experiment 2

A potential worry the reader might have is that repeated testing of the moving dots task has reduced the power of our design. In order to alleviate these concerns, we conducted a power analysis, based on the effect size as reported by Green et al. (2010).

Green et al. (2010) conducted a pre–test and a post–test and reported a partial η^2 of .47. This quantity corresponds to a mean difference between the pre– and post–measure of 1.88 sd. Consider the following two scenarios. In the first scenario, we compare a pre–test to a post–test. We have three conditions, the no–game and the cognitive game condition improve 0.5 sd from pre– to post–test due to practice. The action game condition improves 1.88 sd + 0.5 sd from pre– to post–test due to gaming plus practice. When we generate data under these conditions 5000 times and test for an interaction effect between measure (pre– and post–) and condition, we get a power to detect the gaming effect of .957. Despite the fact that this are two measures only, the power is already quite high.

In the second scenario, we generate data for six sessions. We have three conditions, the no–game and the cognitive game condition improve 0.5 sd from pre– to post–test due to practice *per session*. Note that we assume that the practice effect is linear, a conclusion

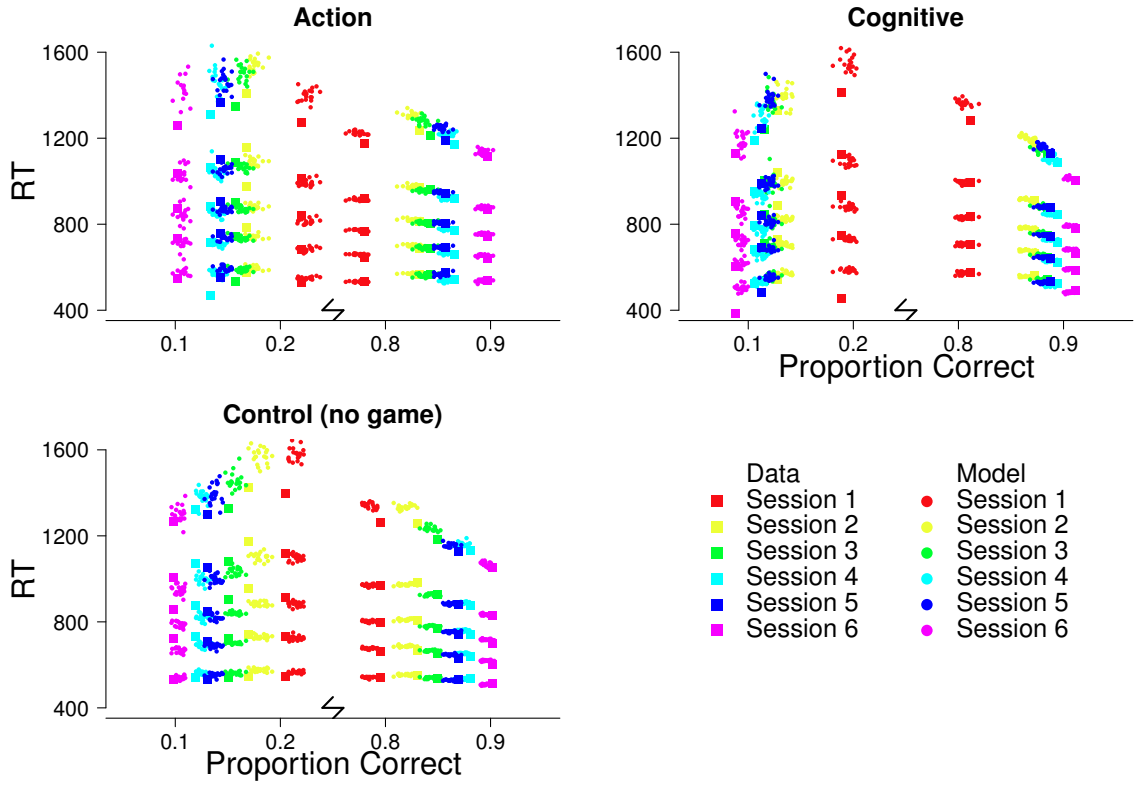

Figure 4. Comparison between synthetic data generated from the diffusion model and real data from Experiment 2. The model captured the data relatively well. Squares = empirical data; circles = synthetic data, based on bootstrapped model parameter estimates.

that is backed up by our behavioral results (see Figure 7 of van Ravenzwaaij et al., 2013). Thus, in total these conditions improve 2.5 sd (1–2, 2–3, 3–4, 4–5, and 5–6). The action game condition improves 1.88 sd + 2.5 sd from session 1 to session 6 due to gaming plus repeated practice. When we generate data under these conditions 5000 times and test for an interaction effect between session (1 through 6) and condition, we get a power to detect the gaming effect of .997. Unsurprisingly, this is substantially higher than the power for the two measurement design.

To conclude, the power of our design to detect an effect of playing action video games as reported by Green et al. (2010) is sufficiently high that even those who are uncomfortable with Bayesian statistics can be confident of the null results of Experiment 2.

References

Bayarri, M. J., & Berger, J. O. (1998). Quantifying surprise in the data and model verification. In *Bayesian statistics 6: Proceedings of the sixth Valencia international meeting* (pp. 53–82).

Dutilh, G., Krypotos, A.-M., & Wagenmakers, E.-J. (2011). Task–related versus stimulus–specific practice. *Experimental Psychology*, *58*, 434–442.

Dutilh, G., Wagenmakers, E.-J., Vandekerckhove, J., & Tuerlinckx, F. (2009). A diffusion model decomposition of the practice effect. *Psychonomic Bulletin & Review*, *16*, 1026–1036.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge: Cambridge University Press.

Green, C. S., Pouget, A., & Bavelier, D. (2010). Improved probabilistic inference as a general learning mechanism with action video games. *Current Biology*, *20*, 1573–1579.

Ratcliff, R. (2002). A diffusion model account of response time and accuracy in a brightness discrimination task: Fitting real data and failing to fit fake but plausible data. *Psychonomic Bulletin & Review*, *9*, 278–291.

Ratcliff, R., Gomez, P., & McKoon, G. (2004). Diffusion model account of lexical decision. *Psychological Review*, *111*, 159–182.

van Ravenzwaaij, D., Boekel, W., Forstmann, B., Ratcliff, R., & Wagenmakers, E.-J. (2013). Action video games do not improve the speed of information processing in simple perceptual tasks. Manuscript submitted for publication.

Vandekerckhove, J., & Tuerlinckx, F. (2007). Fitting the Ratcliff diffusion model to experimental data. *Psychonomic Bulletin & Review*, *14*, 1011–1026.