frontiers in
BEHAVIORAL NEUROSCIENCE

PERSPECTIVE ARTICLE
published: 04 February 2014
doi: 10.3389/fnbeh.2014.00028

Parameter optimization for automated behavior assessment: plug-and-play or trial-and-error?

Laura Luyten[1]*, *Natalie Schroyens*[2], *Dirk Hermans*[1] and *Tom Beckers*[1]

[1] *Psychology of Learning and Experimental Psychopathology, Faculty of Psychology and Educational Sciences, KU Leuven, Leuven, Belgium*
[2] *Experimental Neurosurgery and Neuroanatomy, Department of Neurosciences, KU Leuven, Leuven, Belgium*

Edited by:
Shigeyoshi Itohara, RIKEN Brain Science Institute, Japan

Reviewed by:
Tsuyoshi Miyakawa, Fujita Health University, Japan

***Correspondence:**
Laura Luyten, Psychology of Learning and Experimental Psychopathology, Faculty of Psychology and Educational Sciences, KU Leuven, Tiensestraat 102 box 3712, 3000 Leuven, Belgium
e-mail: laura.luyten@ppw.kuleuven.be

Behavioral neuroscience is relying more and more on automated behavior assessment, which is often more time-efficient and objective than manual scoring by a human observer. However, parameter adjustment and calibration are a trial-and-error process that requires careful fine-tuning in order to obtain reliable software scores in each context configuration. In this paper, we will pinpoint some caveats regarding the choice of parameters, and give an overview of our own and other researchers' experience with widely used behavioral assessment software. We conclude that, although each researcher should weigh the pros and cons of relying on software vs. manual scoring, we should be aware of possible divergence between both scores, which might be especially relevant when dealing with subtle behavioral effects, like for example in generalization or genetic research.

Keywords: freezing, fear conditioning, rats, automated measurements, parameter optimization, calibration, manual scoring, VideoFreeze

INTRODUCTION

Over the years, fear conditioning research in rodents has moved from purely "manual" scoring of freezing behavior by human observers to mainly automated measurements. Since a few years, conditioning researchers have (re)focused on generalization, including generalization of context conditioning, because of its relevance to the study of learning and memory, and to the development and maintenance of anxiety disorders (Wang et al., 2009; Wiltgen et al., 2010; Hermans et al., 2013). Contextual generalization/discrimination research necessitates that freezing in several contexts is compared directly (e.g., Wang et al., 2009; Wiltgen et al., 2010; Yu et al., 2010). Counterbalancing of contexts is not always possible, e.g., in case of a generalization gradient with some contexts having a grid and others a plastic floor (Luyten et al., 2013; Poulos et al., 2013). Furthermore, the differences in freezing between the originally trained and a similar context are often modest, and not as clear-cut as between the original and a dissimilar context (unpublished data), resulting in limited effect sizes. Concurrently, there is an increasing interest in transgenic animals, to investigate the role of certain genes in discrimination and generalization (e.g., Yu et al., 2010; Tayler et al., 2011; Cushman et al., 2012). Genetic modifications may, however, result in phenotypes with only small behavioral deficits.

Taken together, this growing domain is confronted with the challenge to distinguish subtle behavioral effects in different contexts. Our data show that automated behavioral measurements may not always be appropriate for these purposes, or that, at least, they should be implemented and interpreted with great care.

To our knowledge, there is only one paper that has systematically validated settings for a widely used system to detect freezing in rodents (i.e., Med Associates, Inc. Video Fear Conditioning System, with VideoFreeze® software). The optimized software parameters for freezing mice [motion index threshold 18 and minimum freeze duration 30 frames (1 s)], were published in this journal (Anagnostaras et al., 2010). One of the challenges of parameter optimization is finding a balance between the detection of "non-freezing" (e.g., tail) movements, while at the same time ignoring respiratory and cardiac motion during freezing. VideoFreeze is also frequently used to assess rat freezing behavior and others have reported their validated settings for rats (motion threshold 50) (Zelikowsky et al., 2012a). It makes sense to use a higher activity threshold for rats, as they are bigger animals and respiratory movements—which do not preclude freezing—will result in more pixels changing than for mice. Those are also the settings that we used in our studies.

MATERIALS AND METHODS

Experiments were conducted on 48 male Wistar rats (±275 g), in three replications of 16 rats each (eight rats per group). All sessions were meticulously scheduled using free ExpTimer software (Luyten and Van Cappellen, 2013). All experiments were approved by the KU Leuven animal ethics committee, in accordance with the Belgian Royal Decree of 29/05/2013 and European Directive 2010/63/EU.

On the first day, rats were trained in context A (**Figure 1A**). Four minutes after the start of the session, rats received five unsignaled footshocks (0.8 mA, 1 s), separated by 90 s. One minute after the last shock, animals were returned to their home cage. Twenty-four hours later, half of the rats were tested in context A and the other half in similar context B (**Figure 1B**). During this test, rats were exposed to the context for 8 min, without shocks. We hypothesized that there would be significantly less freezing (unpaired *t*-test) in context B than in A on day 2, because of a generalization decrement. The results presented here are part of an ongoing study, including other control groups, which will be discussed as a whole elsewhere. Freezing

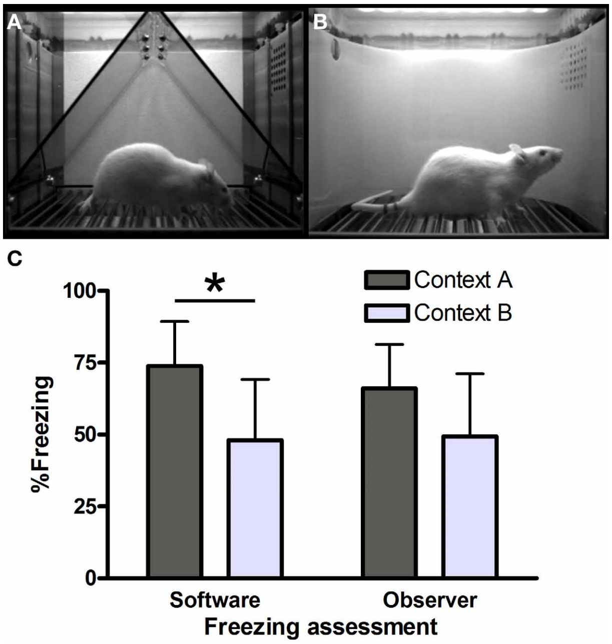

FIGURE 1 | Freezing in contexts (A) and (B). Screenshots from VideoFreeze software in **(A)** context A and **(B)** context B. Freezing **(C)** in contexts A and B in the third series of rats ($n = 8$ per group) as scored by software and human observers. *Significantly different, unpaired *t*-test ($p = 0.01$).

[absence of movement of the body and whiskers with the exception of respiratory motion (Fanselow, 1982)] was measured both manually [continuous measurement with a stopwatch from video recordings, cf. (Luyten et al., 2011, 2012)] and with VideoFreeze software (motion threshold 50, minimum freeze duration 30 frames) (Zelikowsky et al., 2012a). Percentage freezing was calculated as the percentage of time the rat was freezing during the 8-min test on day 2.

Context A (**Figure 1A**) consisted of a standard chamber (Med Associates), with a standard grid floor, a black triangular "A-frame" insert, illuminated by infrared and white light (intensity level 5) and cleaned and scented with a household cleaning product. Context B (**Figure 1B**) consisted of a standard chamber, with a staggered grid floor, a white plastic curved back wall insert, infrared light only and was cleaned and scented with another cleaner. Each chamber was located in one of two identical sound-attenuating boxes.

Rats were always trained in context A, and tested in either context A or context B. Previous (unpublished) data indicated that counterbalancing contexts A and B was not advisable because of different immediate post-shock freezing values (calculated by VideoFreeze) in both contexts when using the training protocol described above [52 rats trained in context A (average post-shock freezing 50%) vs. 24 rats trained in context B (31%), unpaired *t*-test $t_{(74)} = 4.98, p < 0.0001$]. The divergent freezing scores might be due to the different grids delivering the shocks. Given the findings in this paper, the difference may also be partially explained by a software scoring deviation.

RESULTS AND DISCUSSION

Three consecutive series of rats (16 each, eight per group) were compared using software measurements and showed significant contextual discrimination between contexts A and B [A > B; series 1: $t_{(14)} = 2.65$, $p = 0.02$; series 2: $t_{(14)} = 3.79$, $p < 0.01$; series 3: $t_{(14)} = 2.79, p = 0.01$].

Series 3 (**Figure 1C**) was also scored manually, to allow comparison with yet another context (not described here) which was not located in a Med Associates box. Manual scoring was done by two independent observers, blind to the software scores. Surprisingly, hand-scored freezing (average of observers 1 and 2) did not yield significant differences between contexts A and B [$t_{(14)} = 1.78$, $p = 0.10$]. Moreover, software scores were significantly higher than manual scores in context A [74 vs. 66%, i.e., a 8% difference, paired *t*-test $t_{(7)}= 3.93, p < 0.01$], while there was virtually no difference between the software and manual scores in context B [48 vs. 49%, paired *t*-test $t_{(7)} = -1.28, p = 0.24$].

Because of this finding, we decided to examine inter-rater agreement and to retrospectively reevaluate freezing, also manually, in the two previous series of rats.

First, we investigated our findings in series 3 more thoroughly. The agreement between both human observers was substantial (Landis and Koch, 1977) [Cohen's kappa, a statistic to assess rater concordance, here using 20 ordered categories (0–5% freezing, 6–10%, etc.) = 0.65], with an average difference of 2.6% freezing. Therefore, we combined both ratings and used the average as the manual score. Correlations between software and hand-scored measurements were high in both contexts (93% in A and 99% in B), but while agreement in context B was substantial (kappa 0.71), it was only poor in context A (kappa 0.05). Additional analyses comparing software scores with those of a human observer for the two previous series of rats (16 rats per context) led to similar conclusions: only slight agreement in context A (kappa 0.18), but moderate agreement in context B (kappa 0.45).

To conclude, we find good agreement between software and manual scores in context B, but not in context A, while using identical software settings.

We therefore decided to reevaluate our camera calibration. Before the start of our studies, both cameras were calibrated in the base setup (without inserts) as described in the manual and discussed with experienced VideoFreeze users, and both cameras showed crisp and well-contrasted images (**Figure 1**). In addition, both cameras were calibrated using the "Calibrate-Lock" function before each rat. However, it is possible that differences in camera white balance (because of different plastic inserts used in both contexts) caused the observed discrepancies. To investigate this, we adapted the camera white balance in context A until it matched the other context (average grayscale intensity of 119, higher values led to very overexposed images), resulting in a slightly whiter image.

We trained and tested seven more naïve rats in context A and compared software scores with those of two human observers. Two additional rats were excluded from further analyses due to extremely low freezing (≤5%) on day 2. Unfortunately, adapting the white balance did not resolve the discrepancy between the software and observers' scores. Software scores were on average still 3% higher than human scores and when removing one outlying case (Grubb's test, $p < 0.05$) with an average observer score that was 23% higher than the software measure, this difference was even 8%. Analyses indicated that there was still poor agreement between software and observer scores (kappa 0.06).

In conclusion, changing the white balance did not improve agreement between software and manual scores in context A.

Given these findings, we decided to probe into other researchers' experiences and contacted seven research groups who recently (2011–2013) used the Med Associates software and setup for rat research and who implemented different context configurations in their papers. We asked which software settings they used and how they calibrated their systems. It turns out that there are considerable differences between various labs, as to how they define what the software should consider as freezing (**Table 1**).

All studies used several floors (different grids and/or plastic floors) and inserts (e.g., black A-frame and/or white curved back wall), as in our own experiments. The applied software settings were quite variable, with motion thresholds ranging from 18 to 150, and a minimum freeze duration of less than 1 s up to 3 s. Some authors used settings that were previously optimized for mice (Anagnostaras et al., 2010; Halladay et al., 2012; Beeman et al., 2013; Broadwater and Spear, 2013a,b), while others performed their own validations for rats (Zelikowsky et al., 2012b). Although not mentioned in their papers, several researchers reported to us that they optimized their parameters as well. J. Long used a 50 or 100 motion threshold depending on the context and in K. Goosens' lab, the motion threshold usually ranges from 100 to 150, depending on the context configuration and size and strain of the animal (but kept constant for all animals in a certain context on a given test day), and is determined by the experimenter (personal communication).

With regard to the calibration procedure, there was no uniformity either. While half of the research groups (including ours) calibrated the camera before the start of the experiments using a base context, the other half readjusted brightness, gain, and shutter in each context. A quick survey among two labs using VideoFreeze for mice gave a similar picture, with one of both calibrating in each context (Tayler et al., 2011) and the other at initial setup (McDermott et al., 2012). Although some authors mentioned to us that they found that camera calibration can greatly influence the measurements, we did not find meaningful improvements when adapting white balance in our experiments.

Table 1 | Software settings and calibration procedures (adjustment of brightness, gain, and shutter) in seven research groups who use VideoFreeze software and several context configurations for rat conditioning studies.

Motion threshold	Minimum freeze duration (seconds)	Camera calibration	References
18	1	In each context	Beeman et al., 2013
18	1	In standard context	Broadwater and Spear, 2013a,b
20	1.07	In standard context	Moffett et al., 2011
50#	1#	In standard context	Zelikowsky et al., 2012b
50 or 100	0.77	In each context	Long et al., 2011
120	1	In each context	Vander Weele et al., 2013
150#	3#	In each context	Sticht et al., 2012

Most information was obtained through personal communication with the authors. Settings indicated with # were mentioned in the corresponding publications.

Taken together, there is considerable variability in the settings that are being applied by various research groups using the same software and equipment. It is somewhat surprising that rather divergent parameters are being put forward as optimal settings, although this might partially depend on the hand-scoring technique that was used for validation. A recent paper (Shoji et al., 2014) describing a new software package provides hints on how to determine motion thresholds and calibrate in each context. Nevertheless, the question remains whether studies and contexts become more comparable when using different, "optimized" settings in each context or if, conversely, this results in less commensurable measurements. Our finding that the agreement between manual and software measures varied substantially across contexts when using identical settings, is quite alarming. Thus, researchers should be aware of a potential divergence of agreement in different contexts when using a fixed motion threshold and minimum freeze duration. On the one hand, direct comparison of different contexts, e.g., in contextual generalization research, may lead to biased conclusions when using software measurements (in our case artificially increasing differences between contexts A and B). On the other hand, it is self-evident that manual scores by human observers also have drawbacks compared to the objectivity and time-efficiency of automated measurements. We feel that researchers should carefully compare both measurements and decide on the best choice for their research question.

Finally, we would like to stress that in the papers mentioned in **Table 1**, we see no interpretation problems, as most of these studies did not directly compare freezing between several contexts, and in case they did (Zelikowsky et al., 2012b), contexts were counterbalanced. Note however that, theoretically, even limited measurement deviations between contexts may induce heightened variability when counterbalancing contexts, thereby decreasing the chance of finding significant effects.

CONCLUSION

While each researcher should balance the (dis)advantages of automated and manual scoring against one another, we believe that caution is required when using software measurements, particularly when comparing different context configurations or in case of subtle behavioral effects.

AUTHOR CONTRIBUTIONS

Laura Luyten designed the experiments, partially conducted them, analyzed and interpreted all data and wrote the manuscript. Natalie Schroyens carried out part of the studies and contributed to the analyses and manuscript draft. Laura Luyten and Natalie Schroyens manually scored the freezing videos. Dirk Hermans and Tom Beckers contributed to the conceptual design and manuscript draft. All authors approved the final version of the manuscript.

ACKNOWLEDGMENTS

This work was supported by the Research Foundation—Flanders (FWO) Project G072909N, Research Grant (to Laura Luyten) 1504614N and KU Leuven grant PF/10/005. Laura Luyten is a postdoctoral fellow of the FWO. The authors would like to thank all researchers who provided information about their methods.

REFERENCES

Anagnostaras, S. G., Wood, S. C., Shuman, T., Cai, D. J., Leduc, A. D., Zurn, K. R., et al. (2010). Automated assessment of pavlovian conditioned freezing and shock reactivity in mice using the video freeze system. *Front. Behav. Neurosci.* 4:158. doi: 10.3389/fnbeh.2010.00158

Beeman, C. L., Bauer, P. S., Pierson, J. L., and Quinn, J. J. (2013). Hippocampus and medial prefrontal cortex contributions to trace and contextual fear memory expression over time. *Learn. Mem.* 20, 336–343. doi: 10.1101/lm.031161.113

Broadwater, M., and Spear, L. P. (2013a). Age differences in fear retention and extinction in male Sprague-Dawley rats: effects of ethanol challenge during conditioning. *Behav. Brain Res.* 252, 377–387. doi: 10.1016/j.bbr.2013.06.029

Broadwater, M., and Spear, L. P. (2013b). Consequences of ethanol exposure on cued and contextual fear conditioning and extinction differ depending on timing of exposure during adolescence or adulthood. *Behav. Brain Res.* 256, 10–19. doi: 10.1016/j.bbr.2013.08.013

Cushman, J. D., Maldonado, J., Kwon, E. E., Garcia, A. D., Fan, G., Imura, T., et al. (2012). Juvenile neurogenesis makes essential contributions to adult brain structure and plays a sex-dependent role in fear memories. *Front. Behav. Neurosci.* 6:3. doi: 10.3389/fnbeh.2012.00003

Fanselow, M. S. (1982). The postshock activity burst. *Anim. Learn. Behav.* 10, 448–454. doi: 10.3758/BF03212284

Halladay, L. R., Zelikowsky, M., Blair, H. T., and Fanselow, M. S. (2012). Reinstatement of extinguished fear by an unextinguished conditional stimulus. *Front. Behav. Neurosci.* 6:18. doi: 10.3389/fnbeh.2012.00018

Hermans, D., Baeyens, F., and Vervliet, B. (2013). "Generalization of acquired emotional responses," in *Handbook of Cognition and Emotion,* eds M. Robinson, E. Watkins, and E. Harmon-Jones (New York, NY: Guilford Press), 117–134.

Landis, J. R., and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics* 33, 159–174. doi: 10.2307/2529310

Long, J. M., Lee, G. D., Kelley-Bell, B., Spangler, E. L., Perez, E. J., Longo, D. L., et al. (2011). Preserved learning and memory following 5-fluorouracil and cyclophosphamide treatment in rats. *Pharmacol. Biochem. Behav.* 100, 205–211. doi: 10.1016/j.pbb.2011.08.012

Luyten, L., Casteels, C., Vansteenwegen, D., Van Kuyck, K., Koole, M., Van Laere, K., et al. (2012). Micro-positron emission tomography imaging of rat brain metabolism during expression of contextual conditioning. *J. Neurosci.* 32, 254–263. doi: 10.1523/JNEUROSCI.3701-11.2012

Luyten, L., Fanselow, M. S., Vansteenwegen, D., Nuttin, B., and Hermans, D. (2013). "Broad contextual generalization gradients in rats - Optimization of a behavioral protocol," in *Scientific Advisory Board Meeting of the Center of Excellence on Generalization Research* (Leuven).

Luyten, L., and Van Cappellen, F. (2013). ExpTimer: timer software to facilitate complex, multi-step procedures. *J. Open Res. Softw.* 1, e2. doi: 10.5334/jors.ab

Luyten, L., Vansteenwegen, D., Van Kuyck, K., Deckers, D., and Nuttin, B. (2011). Optimization of a contextual conditioning protocol for rats using combined measurements of startle amplitude and freezing: the effects of shock intensity and different types of conditioning. *J. Neurosci. Methods* 194, 305–311. doi: 10.1016/j.jneumeth.2010.11.005

McDermott, C. M., Liu, D., and Schrader, L. A. (2012). Role of gonadal hormones in anxiety and fear memory formation and inhibition in male mice. *Physiol. Behav.* 105, 1168–1174. doi: 10.1016/j.physbeh.2011.12.016

Moffett, M. C., Schultz, M. K., Schwartz, J. E., Stone, M. F., and Lumley, L. A. (2011). Impaired auditory and contextual fear conditioning in soman-exposed rats. *Pharmacol. Biochem. Behav.* 98, 120–129. doi: 10.1016/j.pbb.2010.11.022

Poulos, A. M., Reger, M., Mehta, N., Zhuravka, I., Sterlace, S. S., Gannam, C., et al. (2013). Amnesia for early life stress does not preclude the adult development of posttraumatic stress disorder symptoms in rats. *Biol. Psychiatry*. doi: 10.1016/j.biopsych.2013.10.007. Available online at: http://www.sciencedirect.com/science/article/pii/S000632231300913X

Shoji, H., Takao, K., Hattori, S., and Miyakawa, T. (2014). Contextual and cued fear conditioning test using a video analyzing system in mice. *J. Vis. Exp.* e50871. doi: 10.3791/50871. Available online at: http://www.jove.com/video/50871/contextual-cued-fear-conditioning-test-using-video-analyzing-system?status=a52877k

Sticht, M. A., Long, J. Z., Rock, E. M., Limebeer, C. L., Mechoulam, R., Cravatt, B. F., et al. (2012). Inhibition of monoacylglycerol lipase attenuates vomiting in Suncus murinus and 2-arachidonoyl glycerol attenuates nausea in rats. *Br. J. Pharmacol.* 165, 2425–2435. doi: 10.1111/j.1476-5381.2011.01407.x

Tayler, K. K., Lowry, E., Tanaka, K., Levy, B., Reijmers, L., Mayford, M., et al. (2011). Characterization of NMDAR-independent learning in the hippocampus. *Front. Behav. Neurosci.* 5:28. doi: 10.3389/fnbeh.2011.00028

Vander Weele, C. M., Saenz, C., Yao, J., Correia, S. S., and Goosens, K. A. (2013). Restoration of hippocampal growth hormone reverses stress-induced hippocampal impairment. *Front. Behav. Neurosci.* 7:66. doi: 10.3389/fnbeh.2013.00066

Wang, S. H., Teixeira, C. M., Wheeler, A. L., and Frankland, P. W. (2009). The precision of remote context memories does not require the hippocampus. *Nat. Neurosci.* 12, 253–255. doi: 10.1038/nn.2263

Wiltgen, B. J., Zhou, M., Cai, Y., Balaji, J., Karlsson, M. G., Parivash, S. N., et al. (2010). The hippocampus plays a selective role in the retrieval of detailed contextual memories. *Curr. Biol.* 20, 1336–1344. doi: 10.1016/j.cub.2010.06.068

Yu, T., Li, Z., Jia, Z., Clapcote, S. J., Liu, C., Li, S., et al. (2010). A mouse model of Down syndrome trisomic for all human chromosome 21 syntenic regions. *Hum. Mol. Genet.* 19, 2780–2791. doi: 10.1093/hmg/ddq179

Zelikowsky, M., Bissiere, S., and Fanselow, M. S. (2012a). Contextual fear memories formed in the absence of the dorsal hippocampus decay across time. *J. Neurosci.* 32, 3393–3397. doi: 10.1523/JNEUROSCI.4339-11.2012

Zelikowsky, M., Pham, D. L., and Fanselow, M. S. (2012b). Temporal factors control hippocampal contributions to fear renewal after extinction. *Hippocampus* 22, 1096–1106. doi: 10.1002/hipo.20954

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 28 November 2013; accepted: 20 January 2014; published online: 04 February 2014.

Citation: Luyten L, Schroyens N, Hermans D and Beckers T (2014) Parameter optimization for automated behavior assessment: plug-and-play or trial-and-error? Front. Behav. Neurosci. ***8****:28. doi: 10.3389/fnbeh.2014.00028*

This article was submitted to the journal Frontiers in Behavioral Neuroscience.