Do Behavioral Test Scores Represent Repeatable Phenotypes of Female Mice?

Nadav Menashe, OMS-II[1*], Youstina Salama, OMS-II[2*], Johannie Spaan, PhD[3], Michelle Steinauer, PhD[3]

[1]Western University of Health Sciences COMP-Northwest, Lebanon, OR

[2]Western University of Health Sciences COMP, Pomona, CA

[3]Department of Basic Medical Sciences, Western University of Health Sciences COMP-Northwest, Lebanon, OR

OBJECTIVE

The Marble Bury test is a classic animal behavior test used to model compulsive disorders. It has been used as a tool for quantifying compulsive phenotypes for several applications, most notably for screening pharmaceuticals for disease treatment in humans. Recently, its validity has been questioned for multiple reasons. We hypothesized that if the Marble Bury test quantified a unique phenotype of a mouse, an individual's scores would be repeatable over time.

INTRODUCTION

Prior studies (Gyertyán, 1995; Poling et al., 1981; Wolmarans et al., 2016) have partially tried to address our objective in terms of evaluating habitation to the test but have not explicitly tested repeatability. Also, the timeframes between trials were relatively short with only hours to days between trials. With our study, we investigated repeatability over a longer time frame in mice from four weeks of age to 16 weeks of age. Determining the consistency of phenotypes across longer timeframes will enable its use for chronic conditions or longer-term interventions. Additionally, if these scores represent repeatable phenotypes, experimental designs with pre-intervention and post-intervention tests would be a valid approach, adding statistical power and perhaps reveal novel information regarding follow-ups post intervention.

STUDY DESIGN

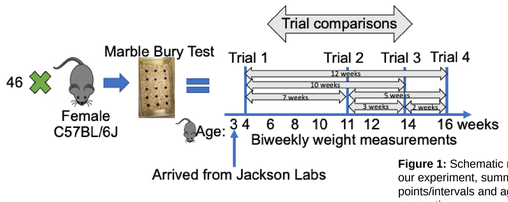
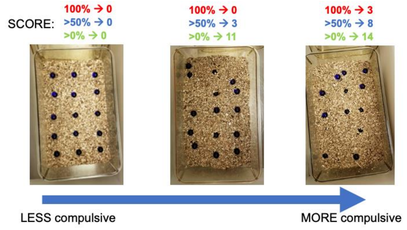


Figure 1: Schematic representation of our experiment, summarizing the time-points/intervals and age (in weeks) across time

Marble burying test

Mice were individually placed in the center of the case for 15 minutes and then removed. We scored the marble burying test in three ways: 1.) counting only marbles that were 100% buried (completely hidden) after the trial, 2.) counting only marbles that were >50% buried (diameter appears smaller when viewed from above) after the trial, and 3.) counting marbles that showed any degree of burying (>0%). Counting onsite was done in 2-3 teams, with two individuals in each team. A consensus was reached among each team, and an average among teams was calculated.



Descriptive Statistical Analysis

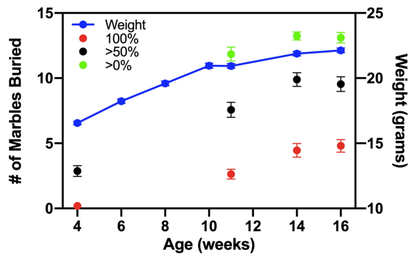
To get an overview of the data set, the mean and standard deviation ($\overline{X} \pm SD$) of marbles buried for each marble bury score at the four repeated trial timepoints were calculated and differences between trials were analyzed with a Friedman rank sum test, followed by a pairwise fashion using the Wilcoxon signed rank test with Bonferroni correction. Mean and standard deviation ($\overline{X} \pm SD$) of body weight were also calculated and the differences between the trial time points were analyzed with a one-way repeated measures ANOVA

Repeatability Statistical Analysis

A repeatability analysis with a single grouping variable was conducted over the 12-week experimental period. Adjusted repeatability value *R* for all three marble bury scores (100%, >50%, and >0%) were calculated including animal ID as a random effect and trial as a fixed effect. The repeatability estimates were conducted with a generalized linear mixed model based on Poisson distribution and log link function, due to non-normal, count data (marble bury scores). Significance testing, *P*-values was based on the likelihood ratio test and the 2.5% and 97.5 % confidence intervals resulting from 1000 bootstrapping runs. The likelihood ratio test obtains statistical significance for repeatability by testing the between-group variance = 0 against the between-group variance >0 . The repeatability estimates range between 0 (low repeatability) and 1 (high repeatability).

RESULTS

Average marbles buried and weight per mouse in weeks



Interpretation: the pattern of marble burying of the mice (N=46) increased across trials and was remarkably similar to their growth curve

Figure 2: The average number of marbles buried for the three different marbly burying scores, 100% buried (red), >50% buried (black), and >0% buried (green) during weeks 4, 11, 14, and 16 of age (trials 1, 2, 3, and 4, respectively; *N* = 46 mice). The z-axis indicates the average body weight (blue) in grams during weeks 4, 6, 8, 10, 11, 14, and 16 of age. Error bars represent the standard error

Are the number of marbles 100%, >50%, and >0% buried repeatable for an individual mouse across trials?

Table 1: Summary of the link-scale adjusted repeatability approximation estimates among individual mice (*N* = 46) for each marble bury score (100%, >50%, and >0%), while accounting for the fixed effect, trial

	Link-scale adjusted repeatability approximation for animal ID (*N* = 46)		
	Adjusted *R*	CI (2.5 – 97.5 %)	*P*-value
Marbles buried 100%	0.52	0.32 – 0.67	< 0.0001
Marbles buried > 50%	0.38	0.21 – 0.54	< 0.0001
Marbles buried > 0%	0	0 – 0.12	1

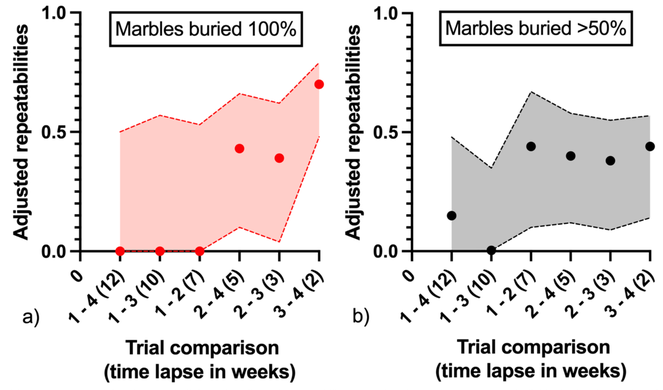


Figure 3: Adjusted repeatability estimates among individual nice (N = 46) at time intervals of 12 weeks (trial 1 vs 4), 10 weeks (trial 1 vs 3), 7 weeks (trial 1 vs 2), 5 weeks (trial 2 vs 4), 3 weeks (trial 2 vs 3), and 2 weeks (trial 3 vs 4) for marbles buried 100% (red dots) and **>50% (black dots).** Shaded areas represent the 2.5% and 97.5% confidence intervals

Table 2: Summary of the changes in link-scale adjusted repeatability approximation estimates among individual mice (*N* = 46) and all possible trial comparisons (i.e., time lapse between trials) for marble burying scores (100% and >50%), while accounting for the fixed effect, trial

	Link-scale adjusted repeatability approximation for animal ID (*N* = 46)		
	Adjusted *R*	CI (2.5 – 97.5 %)	*P*-value
Marbles buried 100% (time lapse in weeks):			
Trial 1 and 4 (12)	0	0 – 0.50	0.5
Trial 1 and 3 (10)	0	0 – 0.57	1
Trial 1 and 2 (7)	0	0 – 0.53	1
Trial 2 and 4 (5)	0.43	0.10 – 0.66	0.0050
Trial 2 and 3 (3)	0.39	0.04 – 0.62	0.0100
Trial 3 and 4 (2)	0.70	0.48 – 0.79	<0.0001
Marbles buried >50% (time lapse in weeks):			
Trial 1 and 4 (12)	0.15	0 – 0.48	0.1750
Trial 1 and 3 (10)	0.004	0 – 0.35	0.4880
Trial 1 and 2 (7)	0.44	0.10 – 0.67	0.0039
Trial 2 and 4 (5)	0.40	0.12 – 0.58	0.0060
Trial 2 and 3 (3)	0.38	0.09 – 0.55	0.0072
Trial 3 and 4 (2)	0.44	0.14 – 0.57	<0.0001

Interpretation: For the 100% marble buried score. 70% (R = 0.70, P< 0.0001), 39% (R = 0.39, P = 0.0100), and 43% (R= 0.39, P= 0.0050) of the variance can be explained by individual mice when compared at 2-, 3-, and 5-week intervals, respectively. For the > 50% buried score 44% (R = 0.44, P< 0.0001), 38% (R = 0.38, P = 0.0072), 40% (R =0.40, P= 0.0060), and 44% (R =0.44, P= 0.0039) of the variance can be explained by individual mice when compared at 2-, 3-, 5-, and 7-week intervals, respectively

DISCUSSION

Overall, counting the number that were 100% buried resulted in the highest repeatability score across all trials; however, it was not an adequate measure for the first trial when mice were only 4 weeks of age. During this trial, very few mice fully buried any marbles (5 mice had a score of 1, the remainder 0), thus the 100% bury measurement was not able to capture the phenotype under our experimental conditions at this early time point. It is possible that allowing a greater time interval for the mice to bury the marbles could have remedied this problem; however, it is possible, then, at the 16-week mark, the counts would "hit a ceiling" and be uninformative if the majority of mice buried all the marbles. Thus, when using the marble bury test across this wide range of growth, the "floor and ceiling effect" may be problematic (Deacon, 2006). The "greater than 50% buried" score is less problematic with regard to the floor and ceiling effect, but had somewhat lower repeatability in our study, which, in part, could be due to observer error as judging >50% buried is arguably more difficult than judging 100% buried. The low marble bury scores during the first trial prompted the addition of another score for the remaining three trials in which we counted any marbles that appeared to be buried (>0%). This measurement had low repeatability among the remaining three trials. One reason is that the data are highly skewed toward the maximum count especially for trials 3 and 4, and thus a "ceiling" is reached. Another reason is that this measurement likely is prone to a wider range of error because it relies more heavily on human judgement.

Our study is unique in that it is the first to show repeatability of marble burying scores across time frames as long as 12 weeks and is the first to use repeatability statistics to show relationships of each individual's scores among trials. Repeatability analysis differs from the normal measures that compares group means of independent (e.g., ANOVA, t-test) or dependent (e.g., repeated ANOVA, paired t-test) response variables, in that it aids in identifying the source of variation within the data and to quantify the constancy of phenotypes (Nakagawa & Schielzeth, 2010). This method has been successfully applied to similar behavioral tests (open field, novel object, and Y-maze behavioral tests; Rudeck et al., 2020; Schuster et al., 2017), and is a powerful way to measure constancy of phenotypes because it measures the between-individual differences in behavior within the same experimental design rather than comparing group means.

CONCLUSION

Our data supports the hypothesis that the scores from the marble burying test represent a measurable phenotype of the animal that stays consistent over-time. We performed four identical trials of marble burying on the same population of mice from 4-16 weeks of age and the data showed statistically significant repeatability in an individual's marble bury scores across the four trials. Another important finding is that the number of marbles buried increased across trials, and the pattern was strikingly consistent with the growth curve of the mice. Thus, the phenotype increased in a predictable way, presumably with growth and the ability of the mice to bury the marbles. The highest repeatability between trials occurred between trials 3 and 4, which occurred when the mice had reached the plateau of the growth curve and there was no difference in weight between trials. This interval also marked the shortest time span between trials (2 v. 3 or 7 weeks); thus, the high repeatability score could be due to a relatively short time span between trials. Further experiments are needed to untangle the effects of growth and time intervals between trials.

REFERENCES

- Deacon, R. M. J. (2006). Burrowing in rodents: A sensitive method for detecting behavioral dysfunction. *Nature Protocols*, *1*(1), 118–121. https://doi.org/10.1038/nprot.2006.1
- Gyertyán, I. (1995). Analysis of the marble burying response: Marbles serve to measure digging rather than evoke burying. *Behavioural Pharmacology*, *6*(1), 24–31. https://doi.org/10.1097/00008877-199505001-00026
- Nakagawa, S., & Schielzeth, H. (2010). Repeatability for Gaussian and non-Gaussian data: A practical guide for biologists. *Biological Reviews*, *85*(4), 935–956. https://doi.org/10.1111/j.1469-185X.2010.00141.x
- Poling, A., Cleary, J., & Monaghan, M. (1981). Burying by Rats in Response to Aversive and Nonaversive Stimuli. *Journal of the Experimental Analysis of Behavior*, *35*(1), 31–44. https://doi.org/10.1901/jeab.1981.35-31
- Rudeck, J., Vogl, S., Banneke, S., Schönfelder, G., & Lewejohann, L. (2020). Repeatability analysis improves the reliability of behavioral data. *PLOS ONE*, *15*(4), e0230900. https://doi.org/10.1371/journal.pone.0230900
- Schuster, A. C., Carl, T., & Foerster, K. (2017). Repeatability and consistency of individual behaviour in juvenile and adult Eurasian harvest mice. *Die Naturwissenschaften*, *104*(3), 10. https://doi.org/10.1007/s00114-017-1430-3
- Wolmarans, D. W., Stein, D. J., & Harvey, B. H. (2016). Of mice and marbles: Novel perspectives on burying behavior as a screening test for psychiatric illness. *Cognitive, Affective, & Behavioral Neuroscience*, *16*(3), 551–560. https://doi.org/10.3758/s13415-016-0413-8

ACKNOWLEDGEMENTS

We would like to thank the *Journal of Pharmacological and Toxicological Methods* for publishing an article based on this poster in their journal. IACUC protocol number: R18IACUC010