Do Behavioral Test Scores Represent Repeatable Phenotypes of Test Subjects?

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

Western University of Health Sciences
College of Osteopathic Medicine of the Pacific
COMP-Northwest

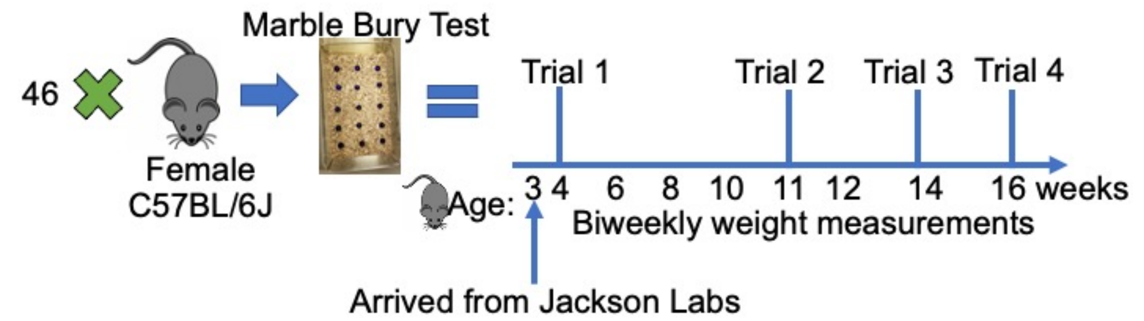
OBJECTIVE

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

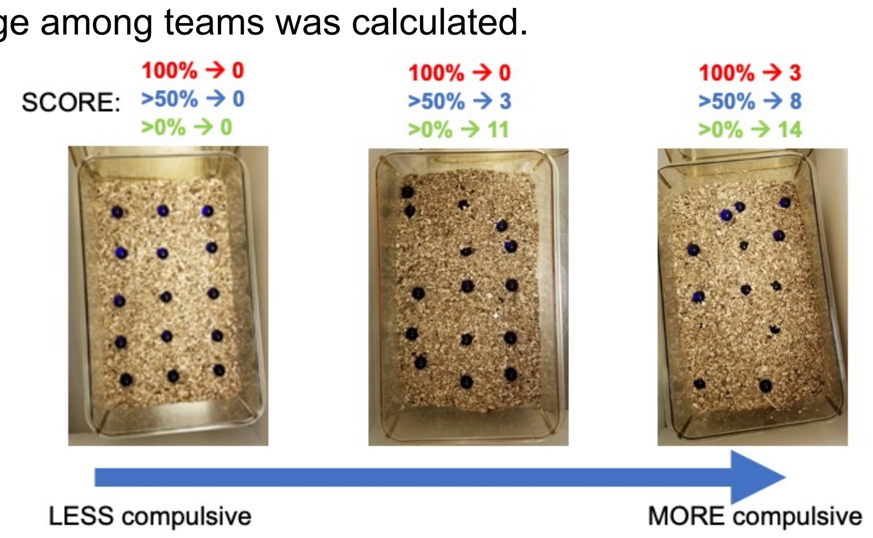
Prior studies (Gyertyán, 1995; Poling et al., 1981; Wolmarans et al., 2016) have partially tried to address our objective in terms of evaluating habitation to the test but have not explicitly tested repeatability. Also, the timeframes between trials were relatively short with only hours to days between 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



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.

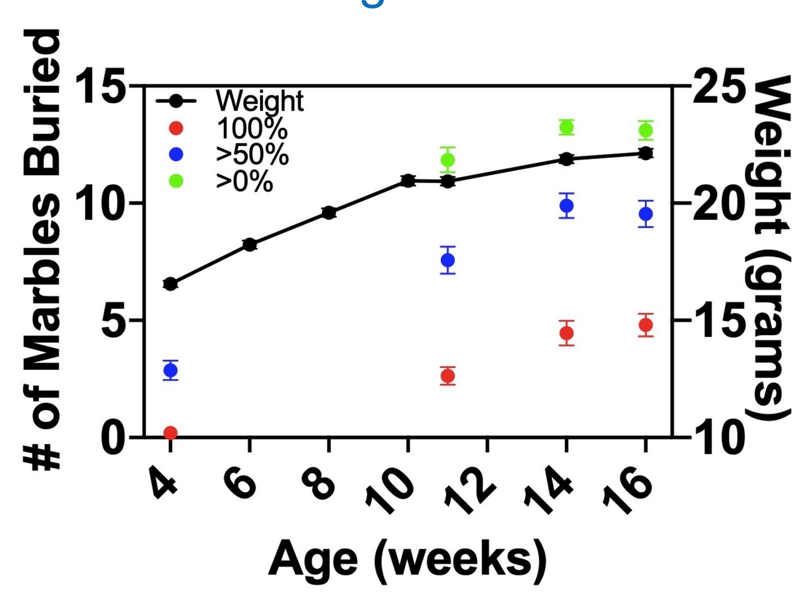


Statistical Analysis

To address whether an individual's marble burying score is predictive of its future scores when the test is repeated, we used a combination of post hoc pairwise comparisons (after significant one-way ANOVA's) and adjusted repeatabilities analysis (removes the variance explained by repeated trials from the repeatability estimation). The statistical package, rptR (Stoffel et al., 2017) was used for estimating adjusted repeatabilities among individual mice, while controlling for repeated trials at different weekly intervals. Poisson distributed count data (marble burying scores) were analyzed with log link, 1000 parametric bootstrap interactions for estimating confidence intervals and 1000 permutations for significance testing.

RESULTS

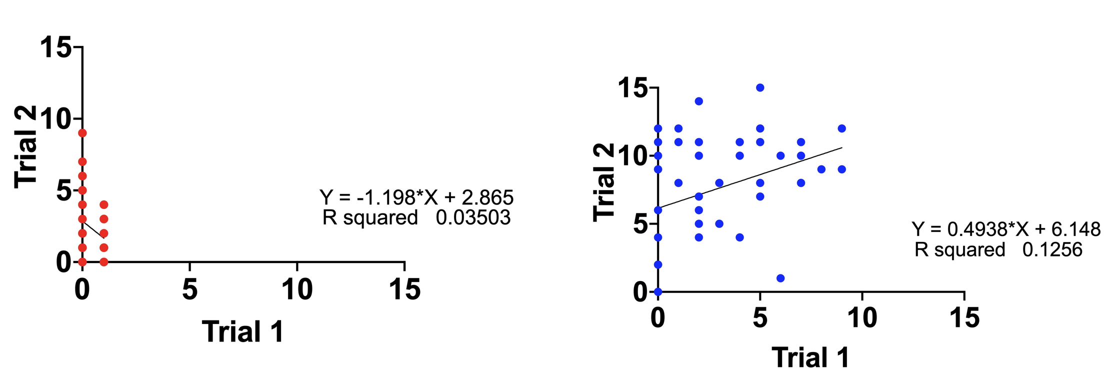
Average marbles buried and weight per mouse in weeks



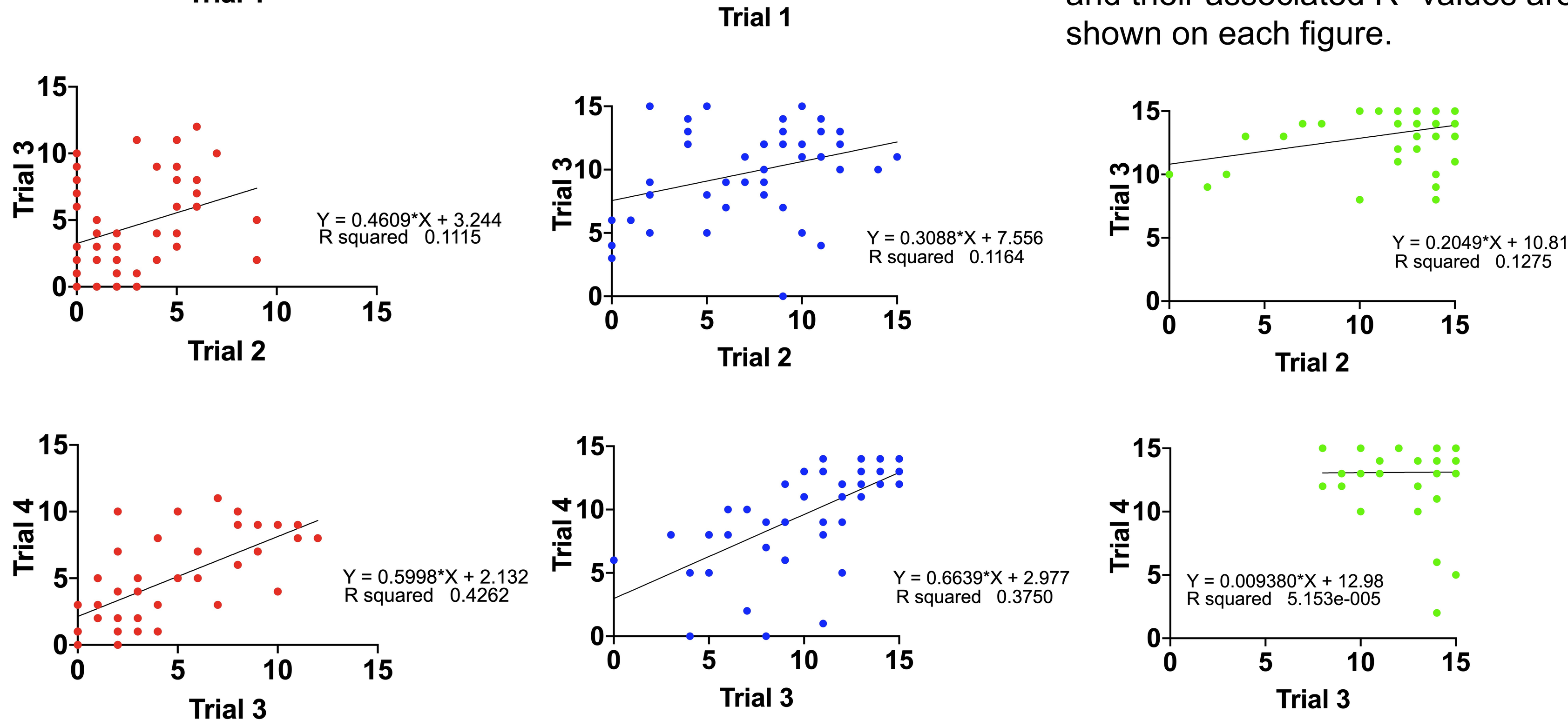
Interpretation: the pattern of marble burying of the mice (N=46) increased across trials and was remarkably similar to their growth curve

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

	Adjusted *R*	SE	CI (2.5 – 97.5 %)	*P*-value
a) Marbles 100% buried	0.52	0.09	0.31 – 0.67	**< 0.0001**
Trial 1 and 2	0	0.14	0 – 0.48	1
Trial 2 and 3	0.39	0.15	0.07 – 0.67	**0.01**
Trial 3 and 4	0.70	0.09	0.46 – 0.81	**< 0.0001**
b) Marbles buried >50%	0.38	0.09	0.19 – 0.53	**< 0.0001**
Trial 1 and 2	0.44	0.14	0.13 – 0.68	**0.0039**
Trial 2 and 3	0.38	0.12	0.08 – 0.55	**0.0072**
Trial 3 and 4	0.44	0.11	0.14 – 0.57	**< 0.0001**
c) Marbles buried >0%	0	0.03	0 – 0.12	1
Trial 2 and 3	0	0.05	0 – 0.16	0.5
Trial 3 and 4	0	0.05	0 – 0.15	0.5



Figures show the correlation between the 100% (red dots), >50% (blue dots), and >0% (green dots) marble bury score between various trials. Each dot represents a single mouse (N=46). Linear regression equations and their associated R^2 values are shown on each figure.



Interpretation: scores obtained for marbles buried 100% and >50% at four repeated trials was significantly repeatable for a given mouse but not for the scores obtained by marbles buried >0%

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 seven weeks and is the first to use repeatability statistics to show relationships of each individual's scores among trials. Repeatability analysis goes beyond the normal measures that compares means of independent (e.g., ANOVA, t-test) or dependent (e.g., repeated ANOVA, paired t-test) variables, in that it measures the closeness (repeatability) between independent tests obtained within the same experimental design at repeated time intervals. It also describes and explains the source of variation to be expected between repeated measures of the same test subject (Nakagawa & Schielzeth, 2010). Not only was this method successfully applied to the open field behavioral test (see Rudeck et al., 2020), but being able to direct the statistics towards intraindividual variability rather than group comparisons makes this method more powerful.

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
- 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 mice who were a part of this study. IACUC protocol number: R18IACUC010