Assessing the Quality of Behavioral Data Obtained by Human Observers Using Cohen's Kappa and Accessory Metrics: Development of the Algorithms and an Open-Source Library

João Antônio Marcolan(✉), Jefferson Luiz Brum Marques, and José Marino-Neto

Institute of Biomedical Engineering -IEB -UFSC, Department of Electrical Engineering (EEL), Federal University of Santa Catarina, 88040-900 Florianópolis, SC, Brazil
jamarcolan@gmail.com

Abstract. Behavioral recordings made by human observers (HOs) are central to animal pre-clinical behavioral models (ABM) of neurobiological diseases, where behaviors (e.g., swimming or immobility) are transcripted from video recordings of experi-ments by HOs. These models face criticism due to their vulnerability to reproductibility issues; evaluation of HO's reliability during training can help to control this source of error. Here, we propose and test algorithms for estimation of Co-hen's Kappa (K) index and accessory measures (maximum K, prevalence, bias) associated with bootstrapping (BS) of behavioral ratings produced during a real experiment using the rat's Forced Swimming Test (FST), to evaluate intra-Hos reliability for the recorded categories. Present results indicate that the use of repli-cas after BS faithfully mirrors most of the concordance attributes of the original transcripts while allowing a statistical evaluation of intra-HO's reliability, and their differences concerning the maximum agreement (Kmax), and the probabili-ties of under-or overestimation of K (bias and prevalence). The use of these tools can inform and optimize the performance of HOs in the use of ABM, without requiring time-intensive re-testing, favoring the reproducibility of the data obtained by these procedures.

Keywords: Intra-observer reliability tests · Behavioral models · Bootstrap · Confidence intervals · Cohen's K

1 Introduction

Behavioral recordings by human raters are central to preclinical animal behavioral models of neuro-biological diseases, such as anxiety and depression, as well as to the study of drugs potentially useful to treat these conditions ([19]). In these models, behavioral categories (like swimming or eating) are recorded (transcripted from video recordings) by human observers (HOs). These models

J. L. B. Marques et al. (Eds.): CLAIB 2022/CBEB 2022, 101, pp. 413–423, 2024.
https://doi.org/10.1007/978-3-031-49410-9_43

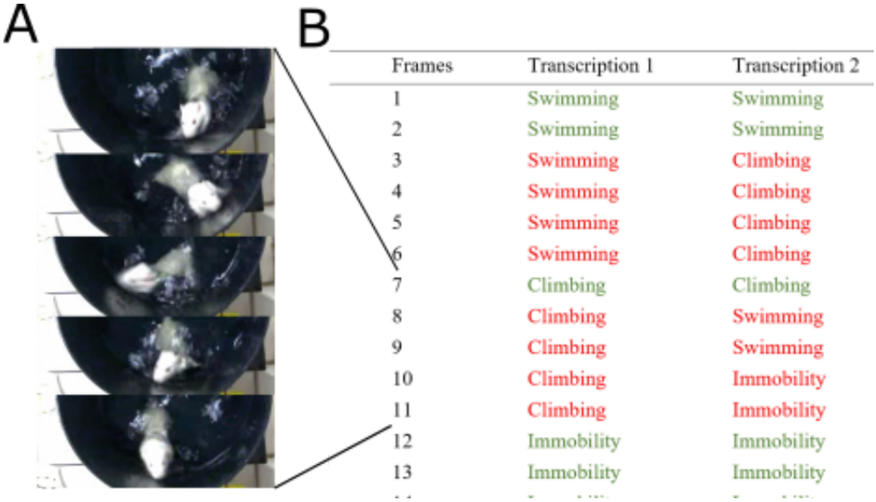
are facing considerable criticism regarding their vulnerability to reproducibility issues (e.g., [17]), arising from multiple methodological causes (e.g., [16]). These causes may include failure to assess and control for intra- or inter-observer reliability (e.g., using analyses such as Cohen's Kappa(K) analyses and its associated indexes; [2,6]). As suggested by a recent systematic review on the use and report of reliability assessment on the Forced Swimming Test literature (FST, a relevant model in behavioral pharmacology for the study of antidepressant drugs), these tests are rarely employed [10]. Despite their relevance for quality of data collection and long-standing literature, there is a lack of accessible, open-source tools to estimate intra- and inter-rater reliability, available to neurobehavioral researchers.

Here, we propose and test algorithms for the calculation of K and associated measures (Maximum K achievable, or Kmax, Prevalence (or P), and Bias (or B; [15]) of behavioral ratings produced during a real experiment using the rat's FST Test, aimed to estimate and evaluate inter-and intra-HO reliability for all and each of the recorded behavioral categories. Estimation and assessment of these indexes are laborious, time-consuming, and demand repeated ratings of multiple samples by multiple researchers. To improve the suitability of these tests without sacrificing their meaning and precision, we also propose and test an approach to infer the population's K, Kmax and Cohen's "d" using bootstrapping of pairs of real observations This method is preferred to estimate confidence intervals (CI) of reliability indexes like K because its results tolerate non-Gaussian distributions of data [8,20]. The algorithms presented here were coded in a freely available open-source stand-alone application and library, (https://github.com/EthoWatcher/Reliability).

2 Materials and Methods

The tests and algorithms described below were coded using C++ 11, the resulting images were developed in Python 3.6.13, run in Windows 10, and are available on a Github repository (https://github.com/EthoWatcher/Reliability). These tests were carried out using transcriptions of a video taken from a real rat FST experiment (one adult male rat, recorded for 4 min), performed by a single HO in 2 sessions 15 days apart (as part of a study approved by the Ethics Committee of the Federal University of Santa Catarina; CEUA-UFSC, PP00764). HO was trained to use the transcription tool ([11,12], https://github.com/EthoWatcher/ethowatcher) and a catalog of 6 categories usual for FST studies [4,9]. Each transcription resulted in a sequence of 7200 annotations (one for every frame in the video, Fig 1). For the bootstrapping (BS) tests, these 2 samples were paired and resampled 1667 times by a BS algorithm developed and coded after [13,18]. Original or resampled (post-BS procedures) pairs of transcriptions were used to build intra-HO catalog agreement matrices (CAM; Table 1A) and catalog maximum agreement matrices (CAMmax; Table 1B), that were developed according to an adaptation of the AMmax of Sim and Wright (2005). Categorical agreement matrices (CTAM, Table 1C) and categorical maximum agreement matrices

(CTAMax) for each behavioral category in the catalog were also built. K and related indexes (Kmax, B, P and Cohen's "d", see below) were estimated from CAM, CTAM, CAMmax and CTAMax, to determine the intra-HO's reliability in using the entire catalog, as well as HO's performance for each category.



Frames	Transcription 1	Transcription 2
1	Swimming	Swimming
2	Swimming	Swimming
3	Swimming	Climbing
4	Swimming	Climbing
5	Swimming	Climbing
6	Swimming	Climbing
7	Climbing	Climbing
8	Climbing	Swimming
9	Climbing	Swimming
10	Climbing	Immobility
11	Climbing	Immobility
12	Immobility	Immobility
13	Immobility	Immobility

Fig. 1. A. Video frames of an experimental FST in rats. B. Resulting transcription of corresponding frame pairs side-by-side; behaviors in green are agreements, and in red indicates disagreements between transcriptions 1 and 2 of the same video sample taken by the same HO 15 days apart.

The first step was to transform CAM (Table 1A, N x N, for N behaviors) into a CTAM (Table 1C) for each behavior. The algorithm arranges the CAM so that the selected behavior is positioned in its 1st row and its 1st column of CTAM. Using the notation indicated in Table 1A, the CAM cell a indicates the frequency of intra-HO agreements on "Swimming" (S) occurrence in the 2 transcriptions; the agreement on Not-S occurrence in both transcriptions is the average between cells e and i; 3) In the CTAM, disagreements (cells b and c, Table 1C) are the average of cells b and c, and of the cells d and g respectively of the Table 1A. From the CTAM, reliability indexes for the HO for each behavior can be then estimated (see below).

In addition to the CTAM and the CAM, a CTAMmax and a CAMmax were obtained by changing the main diagonal cells so that they contain the highest possible values of agreement frequencies while keeping the marginal values (Kmax; [15]). K max can be compared with K to estimate if there is (or not) room for performance improvement [15]. The higher the difference between the K and the corresponding Kmax, the higher the chance of improving the HO's performance (e.g., by increasing training or refining the catalog).

The second step is to find the CAMmax and CTAMax for each behavior in the catalog. CTAMmax was determined by a modification of the methodology described by Sim and Wright (2005), to allow finding the CAMmax. This modification establishes the agreement diagonal as the minimum value between the marginals of the respective lines and columns [15]. Then a backtracking algorithm [14] is used to find the values of the other cells and to calculate the marginals. The backtracking algorithm selects a value for the cell and tests if it is valid. Validity occurs when a) the value of the sum of the column or row cells must result exactly in the marginal values (if the number is in the last cell of the row and/or column) or when b) the value is any number that, when added to the other values selected of the row or column, is lower than or equal to the values of the marginal of the corresponding column or row (if there are remaining cells to be determined). When the selected number is valid by these criteria, the algorithm repeats these procedures for the next cell, and so on.

After determining the CAM, CAMmax, CTAM, and CTAMmax we calculate the K, for each matrix, using formula (1).

$$K = \frac{Po - Pc}{Pc - 1} \tag{1}$$

where $Po =$ proportion of observed agreements, $Pc =$ proportion of agreements expected by chance. Using Table 1A, the Po can be calculated using formula (2) and Pc using formula (3).

$$Po = \frac{a + e + i}{n} \tag{2}$$

where a, e, and i are cells located in the main diagonal and $n =$ number of paired rating

$$Pc = \sum_{n=1}^{q\_c} \frac{PiPj}{n} \tag{3}$$

where q_c = number of behavioral categories, Pj indicates the marginals of transcription 2 and Pi is the marginals of transcription 1.

K ranges from −1 (total disagreement) to 1 (complete agreement), while $K = 0$ indicates random agreement. The benchmark for Cohen's K interpretation proposed by Landis and Koch (1977), determined that $K < 0$ corresponds to a poor agreement, while K ranging 0.01–0.2 = slight agreement; 0.21–0.04 = fair agreement; 0.41–0.6 = moderate agreement; 0.61–0.80 = substantial agreement; 0.81–1 = almost perfect agreement [7]. The Prevalence index (P) indicates the homogeneity of the frequencies of the categories on which the transcriptions agree, and using the notation in Table 1C, can be calculated by the formula (4).

$$Prevalence = \frac{abs(a - d)}{n} \tag{4}$$

where $abs(a - d)$ = absolute value of the difference between the frequencies of these cells and $n =$ number of paired ratings. The higher the P-value, the more the measured K can be undervalued in the performance of the observer. As an

Table 1. An example of an intra-observer (A) catalog agreement matrix (CAM), (B) catalog maximum agreement matrix (CAMax); and (C) categorical agreement matrix (CTAM). Numbers refer to the frequency of agreements for each category. Note that the superscript right letters in the upper right-hand corners of the cells indicate the notation used in respective agreement formulas

A (CAM)		Trasncription 2			Total
		Swimming	Climbing	Immobility	
Transcription 1	Swimming	2^a	4^b	0^c	6
	Climbing	2^d	1^e	2^f	5
	Immobility	0^g	0^h	3^i	3
Total		4	5	5	14^n

B (CAMmax)		Trasncription 2			Total
		Swimming	Climbing	Immobility	
Transcription 1	Swimming	4^a	0^b	2^c	6
	Climbing	0^d	5^e	0^f	5
	Immobility	0^g	0^h	3^i	3
Total		4	5	5	14^n

C (CTAM)		Trasncription 2	
		Swimming	Not Swimming
Transcription 1	Swimming	2^a	2^b
	Not Swimming	1^c	2^d

example, a "slight" K value associated with a high P suggests that the observer can be better than the K indicates. P values near zero suggest that the estimated K is accurately indicating the observer's (bad or good) performance. So, the more the P approaches zero, the more reliable is the K values in estimating agreement.

Bias (B) indicates the homogeneity of frequencies at which transcriptions disagree. Using the notation in Table 1C. B can be calculated by the formula (5).

$$Bias = \frac{abs(b-c)}{n} \tag{5}$$

where $abs(a-c)$ = absolute value of the difference between the frequencies of these cells and the n = number of paired ratings. In contrast to P, the higher the B, the more the observed K value overestimates the real agreement between the transcriptions. B values near zero suggest that the estimated K is accurately indicating the observer's (bad or good) performance [15].

B and P are calculated for CAM and CAMmax, using an approach modified here from that was proposed [15] to allow for calculations for N x N matrices, for N >2 (N = number of categories in the catalog). This was achieved by (1) selected the cells located on the central diagonal, for the calculation of the prevalence, or cells located outside the central diagonal, to calculate the bias; (2) constructed for each selected cell a vector with origin at (0, 0), directed along the X-axis, and with intensity determined by its value, Fig. 2A; (3) applied to each of the vectors a linear transformation that rotated them so that the angular difference between the vectors was uniform, Fig. 2B; (4) summing the positioned vectors, and with the module of this vector divided by the n=number of paired ratings, prevalence or bias was calculated. We repeat this operation for every possible combination of vectors and extract the largest value.

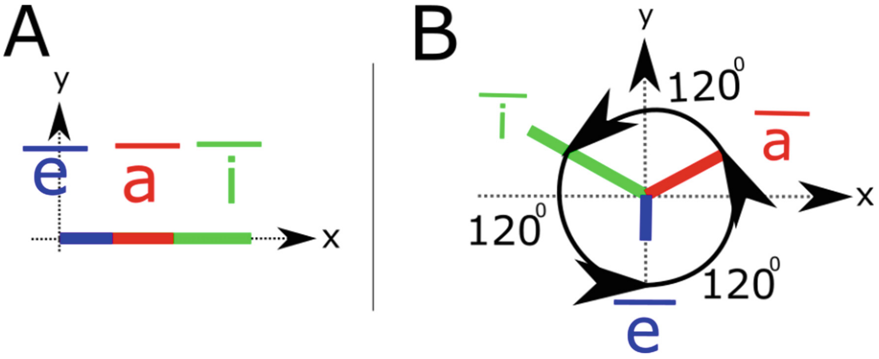


Fig. 2. Example of calculation of P of CAM from Table 1A. A) the first step of the algorithm that positions the cells values on the cartesian plane; B) the second step of the algorithm that rotates the vector in a way that all have evenly angles, The length of blue bars represents the 'e' cell value, red bars represent the 'a' cell value and green bars represent 'i' cell value from CAM matrix of Table 1A. The black solid lines represent the angle among vectors.

After calculating the K, P, and B of the CAM, CAMmax, CTAM and CTAMmax constructed for the original pair of transcriptions from the FST video, a set of new agreement matrices were built from those 1667 pairs of transcriptions obtained after the BS procedures. To build these new transcriptions and their corresponding agreement matrices, a list of random numbers with replacement (ranging from 1 to 7200, the total number of frames in each transcription) was generated and associated with each frame of the paired original transcription (according to [13,18]). With each bootstrap replica was calculated the K, P, and B from the CAM, CAMmax, CTAM and CTAMmax, in a way that was possible to generate a frequency distribution for each agreement descriptor.

The frequency distributions of K and Kmax were used to calculate the distance (Cohen's d) between these descriptors using the formula (6) [5].

$$K = \frac{M_{max} - M_o}{Pooled\ sample\ SD} \frac{N-3}{N-2.25} \sqrt{\frac{N-2}{N}} \tag{6}$$

where M$_{max}$ = average of Kmax from the bootstrap replicas, M$_o$ = average of K from the bootstrap replicas, N = number of bootstrap replicas. The standard deviation (SD) pooled for all the BS replicas was estimated by formula (7).

$$Pooled\ sample\ SD = \sqrt{\frac{SD_max^2 + SD_o^2}{2}} \tag{7}$$

where SD$_{max}$ = standard deviation of Kmax for the BS replicas and SD$_o$ standard deviation of K from the BS replicas. Cohen's d between the K and Kmax allows an estimate of the distance between the means of the distributions in SD (e.g. d = 2 indicates that there is a distance of 2 SD between K and Kmax distributions). The higher the d value, the higher the possibility of improvement in the HO's performance (e.g., through more training). Furthermore, a 95% CI was calculated using the percentile interval technique ([18]) for the frequency distribution of each agreement descriptor, which is useful to estimate the range of possible values for each descriptor. The 95% CI of K and Kmax can be used to test for significant distances between them: if they don't overlap, it is possible to affirm that they are significantly different, assuming a 5% error of type I in this conclusion [1,3]. Thus, by observing the d value between K and Kmax 95% CIs it is possible to assess which categories can be refined, or better trained.

3 Results and Discussion

The frequency distribution of K and Kmax values generated from 1667 bootstrapped replicas from each category of the catalog and for the overall catalog is compared to the same values estimated for the original pair of transcriptions as shown in Fig. 3. K values (Fig. 3A–E) indicate that the HO's reliability in the FST video was 0.77 (a K value indicative of "substantial" intra-observer agreement, according to [7]). That contrasts with the almost perfect agreement (by the same benchmark) shown when recording some behaviors (e.g., swimming or immobility) or just a fair performance when rating headshaking.

BS procedures preserved the mean K values essentially similar to those observed for the original pair of transcriptions (as indicated by the K deltas); their narrow frequency distributions and C.I. values are confined to the same benchmark "diagnoses" found for original transcriptions. Interestingly, BS procedures overestimated the K values for the poorest performances (those showing the widest distributions: diving and headshaking). Kmax values and distributions (Figs. 3 and 4) can further refine the reliability analysis: significant differences between BS replicas average K and the average Kmax suggest that performance in the overall catalog can be improved (by training or better defining the categories) from 0.77 (substantial) to 0.99 (nearly perfect, the same is valid for swimming or immobility).

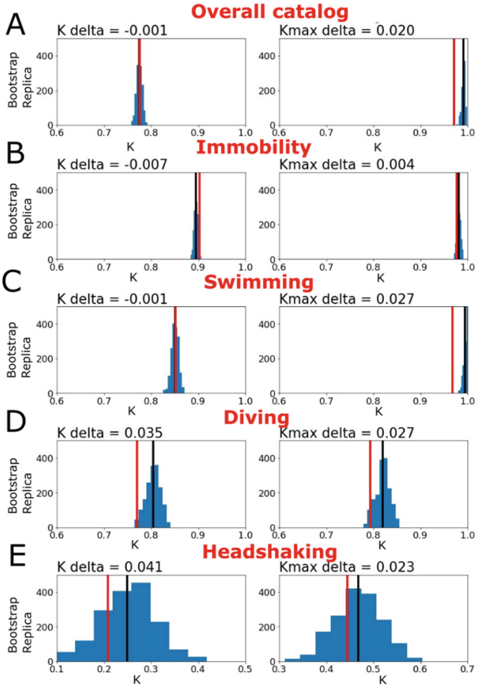


Fig. 3. Frequency distribution (n = 1667 bootstrap replicas) of K and Kmax for the overall catalog and of each behavioral category. Red lines represent K (left column of figures) or Kmax (right column of graphs) observed for the original pair of transcriptions, while black lines represent the average of the bootstrapped replicas. Kδ and Kmax δ are the differences between the K and Kmax values of the original pair of transcriptions and the mean value obtained for the bootstrapped replicas

The substantial K obtained for diving is too close to its maximum achievable (Fig. 4D), suggesting that there is no room for further improvement in the performance of the observer regarding this behavior. This analysis also indicates that while headshaking reliability can increase, the maximum K achievable remains at the "moderate" levels.

Analysis of B and P values of bootstrapped replicas (Fig. 4A–E) can further refine the interpretation of the K and Kmax. Because it is possible to evalu-

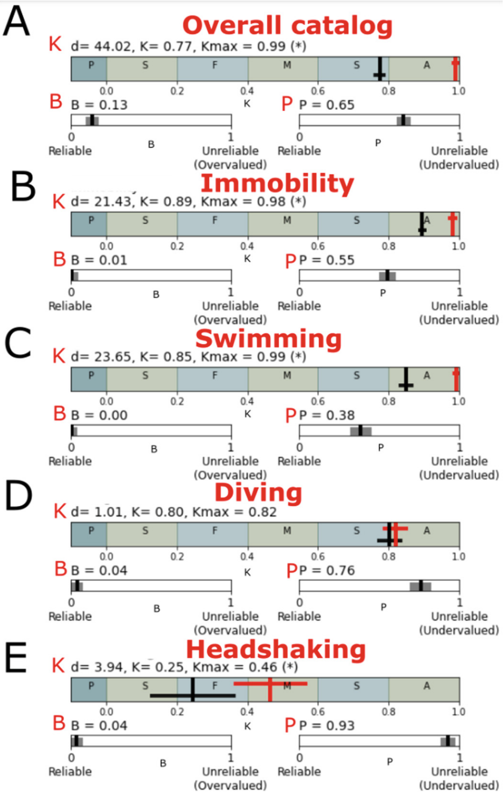


Fig. 4. A summary of the K, P, B and the K-Kmax distance for the overall catalog (A) and single behavioral categories (B–E) for the 1667 replicas obtained by the BS algorithms. The average K values (black vertical bars) and their respective confidence intervals (black horizontal bars) are represented with the respective Kmax averages (red vertical lines) and their confidence intervals (red horizontal lines). d = Cohen's d between Kmax and K in Standard deviations. (*) indicates that K and K max are significantly different ($p < 0.05$; according to [1,3]). The gray rectangles in this graph indicate the K level according to the benchmark of [7] for K so that P indicates $K < 0$ (poor agreement), S for K of 0.01–0.2 (slight agreement); F for K of 0.21–0.04 (fair agreement); M for K of 0.41–0.6 (moderate agreement); S for K of 0.61–0.80 (substantial agreement); A for K of 0.81–1 (almost perfect agreement). In the B graph, the average bias (or B) of the BS replicas values (black vertical bars) and their respective confidence intervals (black horizontal bars) are represented. In the P graph, the average prevalence (or P) of the BS replicas values (black vertical bars) and their respective confidence intervals (black horizontal bars) are represented

ate if the K calculated is unreliable and with the direction they are skewed. B indicates that none of K calculated from the catalog and for each category are skewed towards overvaluation. Implying that all K values estimated are at least conservative. Complementing, the P values indicate if the K calculated are reliable or skewed towards undervaluation. Using this as criteria the most unreliable K value was for Headshaking, implying that the K calculated for this category is better than shown in Fig. 4E. P informs that every category is skewed toward undervaluation, meaning that the K values calculated are conservatives.

4 Conclusions

Present results indicate that the use of replicas after BS faithfully mirrors most of the concordance attributes of the original transcripts while allowing a statistical evaluation of intra-HO's reliability, their differences in relation to the maximum agreement (Kmax), and the probabilities of under-or overestimation of K (bias and prevalence). The use of these tools can inform and optimize the performance of HOs in the use of ABM, without requiring time-intensive re-testing, favoring the reproducibility of the data obtained by these procedures.

References

1. Austin, P.C., Hux, J.E.: A brief note on overlapping confidence intervals. J. Vasc. Surg. **36**(1), 194–195 (2002). PMID: 12096281. https://doi.org/10.1067/mva.2002.125015
2. Chaturvedi, S.R.B.H., Shweta, R.C.: Evaluation of inter-rater agreement and inter-rater reliability for observational data: an overview of concepts and methods (2015)
3. Cumming, G.: Inference by eye: reading the overlap of independent confidence intervals. Stat. Med. **28**(2), 205–220 (2009)
4. Domingues, K., Lima, F.B., Linder, A.E., Melleu, F.F., Poli, A., Spezia, I., Lino de Oliveira, C.: Sexually dimorphic responses of rats to fluoxetine in the forced swimming test are unrelated to the function of the serotonin transporter in the brain. Synapse **74**(1), e22130 (2020)
5. Durlak, J.A.: How to select, calculate, and interpret effect sizes. J. Pediatr. Psychol. **34**(9), 917–928 (2009)
6. Gisev, N., Bell, J.S., Chen, T.F.: Interrater agreement and interrater reliability: key concepts, approaches, and applications. Res. Social Adm. Pharm. **9**(3), 330–3 (2013)
7. Landis, J.R., Koch, G.G.: The measurement of observer agreement for categorical data. Biometrics 159–174 (1977)
8. Lee, J., Fung, K.P.: Confidence interval of the kappa coefficient by bootstrap resampling [letter]. Psychiatry Res. **49**(97), 98 (1993)
9. Marchesini, G.: MorphoKinematicFST, um banco de dados unificado de dados categóricos, cinemáticos e morfológicos de ratos submetidos ao Teste do Nado Forçado (FST), validado por procedimentos metrológicos. M.Sc. Dissertation, Federal University of Santa Catarina, Technological Center, Postgraduate Program in Electrical Engineering, Florianópolis. https://repositorio.ufsc.br/handle/123456789/215435 (2019)

10. Marchesini, G., Lino-de-Oliveira, C., Marino-Neto, J.: The use of reliability metrics for observational studies in rats submitted to the forced swim test (FST): a systematic review. Acta Neuropsychiatr. **31**(S2), 1–54 (2019). https://doi.org/10.1017/neu.2019.38
11. Marcolan, J.: Ferramenta de código aberto para análise de comportamento e aquisição de vídeo em "tempo real" usando técnicas de visão computacional e processamento paralelo. M.Sc. Dissertation, Federal University of Santa Catarina, Technological Center, Postgraduate Program in Electrical Engineering, Florianópolis. https://repositorio.ufsc.br/handle/123456789/189477 (2017)
12. Marcolan, J., Marino-Neto, J.: ETHOWATCHER OS - Brazilian National Intellectual Property Institute license on protocol number BR512019002103-7 (2019)
13. McKenzie, D.P., Mackinnon, A.J., Péladeau, N., Onghena, P., Bruce, P.C., Clarke, D.M., McGorry, P.D.: Comparing correlated kappas by resampling: is one level of agreement significantly different from another? J. Psychiatr. Res. **30**(6), 483–492 (1996)
14. Rossi, F., Van Beek, P., Walsh, T. (Eds.): Handbook of Constraint Programming. Elsevier (2006)
15. Sim, J., Wright, C.C.: The kappa statistic in reliability studies: use, interpretation, and sample size requirements. Phys. Ther. **85**(3), 257–268 (2005)
16. Smalheiser, N.R., Graetz, E.E., Yu, Z., Wang, J.: Effect size, sample size and power of forced swim test assays in mice: guidelines for investigators to optimize reproducibility. PLoS ONE **16**(2), e0243668 (2021)
17. Spruijt, B.M., Peters, S.M., de Heer, R.C., Pothuizen, H.H., van der Harst, J.E.: Reproducibility and relevance of future behavioral sciences should benefit from a cross fertilization of past recommendations and today's technology: "Back to the future." J. Neurosci. Methods **234**, 2–12 (2014)
18. Tibshirani, R.J., Efron, B.: An Introduction to the Bootstrap. Monographs on Statistics and Applied Probability, vol. 57, pp. 1–436 (1993)
19. Willner, P., Belzung, C.: Treatment-resistant depression: are animal models of depression fit for purpose? Psychopharmacology **232**(19), 3473–3495 (2015)
20. Wright, D.B., London, K., Field, A.P.: Using bootstrap estimation and the plug-in principle for clinical psychology data. J. Exp. Psychopathol. **2**(2), 252–270 (2011)