Effects of prevalence on observer agreement: a subjective assessment of working donkey behavior as an example

Charlotte C. Burn*[1], Joy C. Pritchard[1,2], and Helen R. Whay[1]
[1]*University of Bristol, Department of Clinical Veterinary Science, Langford, Bristol BS40 5DU, UK*
[2]*The Brooke, Broadmead House, 21 Panton Street, London SW1Y 4DR, UK*
**Corresponding author: charlotte.burn@bris.ac.uk*

Kappa statistics are often used to assess the extent of observer agreement over subjective measures of behavior. They determine the degree to which agreement occurs compared with that expected by chance, so they are more conservative than raw percentage agreement scores. When used on measures with skewed prevalences, however, they become unreliable [1, 2]; with good agreement becoming almost impossible because the probability of observers agreeing purely by chance becomes so high. An alternative kappa calculation, 'PABAK', has been proposed to adjust for prevalence and observer bias [3], but this has been criticised for readjusting for the same factors that kappa is designed to control for [1]. There is no easy solution, so we suggest presenting prevalence indices and the raw percentage agreements alongside the kappa values, making kappa reliability more transparent [2, 4].

We have calculated the prevalence indices as the mean proportion of the most common classification relative to each alternative category, as described by a gold standard (in this case, the person who trained the observers). Thus, even for variables with many categories, the prevalence would be approximately 50% if categories were evenly distributed, but if the distribution was asymmetrical for any category, the prevalence index would increase. To aid interpretation of the prevalence indices we have divided them as follows: 50-59% = Well-balanced; 60-69% = Moderately balanced; 70-79% = Moderately skewed; 80-89% = Skewed; 90-100% = Highly skewed. It should be noted that these categories are only a guide, and their influence on agreement statistics will depend on the sample sizes used (even a slight skew could cause problems with small sample sizes).

We illustrate the above approach using the example of donkeys working in India. These animals have a high prevalence of welfare problems, and can appear unresponsive to the external environment and often demonstrate avoidance or aggressive behavior towards humans [5]. Five observers and their trainer (the gold standard) assessed the demeanour, lameness, and responses to humans of 80 donkeys.

The results are shown in Table 1. The percentage agreements for heat stress and gait were ≥ 98%, yet the overall ratings were Poor. This may mean that subjective assessment of these behaviors were indeed poor. Alternatively, however, the prevalence index showed that the gold standard scored 100% of the donkeys as showing no heat stress behavior and 98% as having abnormal gaits, so good observer agreement would have been almost impossible to detect. These measures can be contrasted against the response to observer approach; here the prevalence index was moderately balanced, so although the percentage agreements are much lower than for heat stress (≥ 74%), the overall agreement is Moderate. Demeanour showed Poor agreement, and this is likely to be a fairly reliable assessment of inter-observer performance because the prevalence index was only moderately skewed. For the other behaviors, agreement was moderate or substantial, despite skewed prevalences.

To conclude, kappa values are reliable when prevalence indices are moderately well balanced, and also when good agreement is obtained despite skewed prevalences. However, when prevalences are skewed, it remains unclear whether poor agreement ratings are due to the high probability of agreeing purely by chance, or due to genuinely poor agreement. This uncertainty should be acknowledged and, as is illustrated here, one approach is to provide prevalence indices alongside agreement ratings and percentage agreements.

References

1. Hoehler, F. K. (2000). Bias and prevalence effects on kappa viewed in terms of sensitivity and specificity. Journal of Clinical Epidemiology 53, 499-503.
2. Vach, W. (2005). The dependence of Cohen's kappa on the prevalence does not matter. Journal of Clinical Epidemiology 58, 655-661.
3. Byrt, T., Bishop, J., Carlin, J. B. (1993). Bias, prevalence and kappa. Journal of Clinical Epidemiology 46, 423-429.
4. Sim, J., Wright, C. C. (2005). The kappa statistic in reliability studies: use, interpretation, and sample size requirements. Physical Therapy 85, 257-268.
5. Pritchard, J. C., Lindberg, A. C., Main, D. C. J., Whay, H. R. (2005). Assessment of the welfare of working horses, mules and donkeys, using health and behaviour parameters. Preventive Veterinary Medicine 69, 265-283.
6. Landis, J. R., Koch, G. G. (1977). The measurement of observer agreement for categorical data. Biometrics 33, 159-174.

Table 1. *Inter- observer agreement results for a working donkey behavior assessment in India. Variables are arranged in ascending order according to the number of observers reaching criterion and their overall agreement ratings. Prevalence indices closer to 50% give more reliable kappa values. Kappa (k) values closer to 1.0 indicate better agreement, adjusting for that expected by chance. The rating scale is adapted from Landis and Koch [6] and Sim and Wright [4].*

Variable	Prevalence index (%)	Majority categories (%, if different to Prevalence index)	Minimum agreement (%)	Overall agreement (*k*)	Rating	Number of observers ≥ Moderate (total=5)
Heat stress (present / absent)	100	Absent	98	0	Poor	0
Gait (normal / abnormal)	98	Abnormal	98	0	Poor	0
Demeanour (alert / apathetic / depressed)	77	Apathetic (56%) or Alert (43%)	49	0.14	Poor	1
Response to observer walking down side (no interest / sign of interest)	73	Signs of interest	74	0.47	Moderate	3
Response to observer Approach (moves away / turns head away / no response / turns head towards / aggressive)	69	Turned head away (39%) or No response (38%)	65	0.58	Moderate	5
Chin contact (accepts/avoids	83	Accepted	86	0.67	Substantial	5
Tail tuck (no response to observer walking past rear / clamps tail down)	96	No tail-tuck	96	0.68	Substantial	5