Test-retest reliability of the Welfare Quality® animal welfare assessment protocol for growing pigs

I Czycholl*, C Kniese†, K Büttner‡, E Grosse Beilage§, L Schrader* and J Krieter†

Abstract

The aim of this study was to assess the feasibility and test-retest reliability of the Welfare Quality® Animal Welfare Assessment Protocol for Growing Pigs. Twenty-three German pig farms were visited repeatedly by the same trained observers; each farm being visited six times during two fattening periods. The entire protocol assessment was carried out during each farm visit, i.e. a Qualitative Behaviour Assessment (QBA), behavioural observations (BO), a Human Animal Relationship test (HAR) and different individual parameters (IPs), e.g. bursitis and tail-biting. Test-retest reliability was evaluated by a Wilcoxon signed rank test (W) and by calculation of the Smallest Detectable Change (SDC) and Limits of Agreement (LoA). The QBA presented non-satisfactory agreement between farm visits. However, good agreement, in general, was found for the BO. For the HAR, no reliability could be detected. Most IPs were of acceptable agreement, with the exception of bursitis and manure on the body. Bursitis showed great differences, which can be explained by difficulties in the assessment when the animals moved around or their legs were dirty. The disagreement in the parameter manure on the body can be explained by seasonal effects. Disagreement was further found concerning the parameters coughing, sneezing, pleuritis, pneumonia and milkspots. Feasibility was good; both observers could be well-trained to fulfil the protocol. Furthermore, the time needed for an assessment did not exceed 6 h. The parts of the protocol that proved to be insufficiently reliable need to be addressed in the future in order to enhance and improve the objective measurement of animal welfare.

Keywords: animal-based measure, animal welfare assessment, farm, pig, test-retest reliability, Welfare Quality®

Introduction

The Welfare Quality® Animal Welfare Assessment protocols are considered feasible, valid and reliable measurement methods to determine animal welfare (Velarde 2007). These protocols are based on four main principles — Good feeding, Good housing, Good health and Appropriate behaviour, which is also the constitutional definition of animal welfare. In terms of a top-down process, these principles are divided into twelve subcriteria, which are then measured by a set of approximately 30 predominantly animal-based parameters to be estimated on-farm. After assessment of the parameters on-farm, the measures are usually expressed as percentages of affected animals. A dimensionless number between 0 and 100 is calculated from these percentages utilising different mathematical methods, e.g. decision trees (Magerman 1995) as well as I-Spline functions (Curry & Schoenberg 1966) and Choquet Integrals (Grabisch & Roubens 2000). This is carried out first at the sub-criteria and then at the principle level. Depending on the numbers obtained (the closer to 100 the better) the farms are scored and labelled as excellent, enhanced, acceptable or not classified (Welfare Quality® 2009).

The basic requirements of all such measurement tools are validity, feasibility and reliability. Validity characterises the amount a given measurement method actually assesses what it is supposed to measure and the relevance of that method. Feasibility describes good cost-effectiveness and applicability (Velarde 2007). Reliability refers to the repeatability of measures under consistent conditions (Carlson et al 2009). Reliability consists of inter-observer reliability, intra-observer reliability and test-retest reliability. Inter- and intra-observer reliability can be influenced by the training of the assessors, whereas test-retest reliability refers solely to the method used (Velarde 2007). Inter-observer reliability describes the need to be independent of the results produced by different observers (Wirtz & Caspar 2002). Intra-observer reliability is defined as the extent of agreement that the same observer reaches from the repeated assessment of video clips or pictures, i.e. the same objects under exactly the same conditions (Martin & Bateson 2007). Due to minor changes, this tends not to be the case when assessing farms or animals repeatedly as, typically, there are other individuals on the farms. However, if the same individuals are being assessed, there may well be individual