Validity of 4 categories with text and pictures for scoring of faecal consistency in pigs

Ken S. Pedersen [a,] *

[a] Department of Large Animal Sciences, University of Copenhagen, Groennegaardsvej 8, DK-1870 Frederiksberg C.

***** Corresponding author. Tel.: +45 3533 3017; fax: +45 3533 3022; E-mail address: ken@life.ku.dk.

Abstract

The objectives of the current study were to evaluate validity of a score system with 4 categories, text and pictures for assessment of consistency in faecal samples from pigs 2 to 10 weeks post weaning. Validity was evaluated in terms of repeatability (intra- and inter-observer agreement) and accuracy (in relation to faecal dry matter content). Finally it was determine whether 4 categories, text and pictures could increase inter-observer agreement compared to a simple faecal consistency score system with 3 categories, no text and no pictures.

The 4 consistency categories were score 1 = firm and shaped, score 2 = soft and shaped, score 3 = loose and score 4 = watery. Faeces samples with consistency score 3 or 4 were considered diarrhoeic in the statistical analysis.

Five observers from the same veterinary practice examined 100 faeces samples using the 4 category system. Four of the observers examined the 100 faeces samples twice within the same day. Faecal dry matter content was determined for the 100 samples using microwaves. Another 99 samples were examined by the same 5 observers using the simple score system. No calibration was allowed between observers before or during any of the examinations.

Mean intra-observer agreement was 0.82 (Cohen’s Kappa = 0.76) for consistency scores and 0.93 (Cohen’s Kappa = 0.86) for diarrhoea. Mean pair wise inter-observer agreement was 0.73 (Cohen’s Kappa = 0.64) for consistency scores and 0.89 (Cohen’s Kappa = 0.78) for diarrhoea.

Mean faecal dry matter content was significant different (p-value < 0.05) between all consistency scores. Faecal dry matter cut-off values for each consistency score were calculated (score 1: dry matter content > 22.6%, score 2: dry matter content > 18.8%, score 3: dry matter content > 13.1%). The faeces samples were classified into the 4 consistency categories according to these cut-off values. On average the observers classified 80% of the samples correct in relation to consistency score and 92% correct in relation to diarrhoea.

The proportion of faeces samples where all observers agreed was lower for the system with 4 categories, text and pictures compared to a simple system with 3 categories, no text and no pictures.

In conclusion, the 4 descriptive categories with text and pictures did not eliminate problems of low repeatability within and between observers. An unacceptable accuracy for consistency score classification of faeces samples was observed. Accuracy was considered acceptable after dichotomization of consistency scores.

More objective measures of faecal consistency may be more appropriate in research studies.

Keywords: Intra-observer, Inter-observer, Agreement, Kappa, Faecal consistency, Diarrhoea, Faecal dry matter content, Pigs.

1. Introduction

Diseases of the gastrointestinal tract can affect all ages of pigs and continue to be one of the most important factors that limit efficiency and profitability in the swine industry around the world (Thomson, 2006). Both infectious and non infectious diseases exist. Diarrhoea in pigs post-weaning accounted for most of the antimicrobial treatments after termination of the antimicrobial growth promoters in Denmark (WHO report, 2003). Enteric infections in weaners, growers and finishers continue to be among the most important diseases in Denmark and are generally believed to be most prevalent in pigs between 6 and 14 weeks of age.

Enteric infections are characterized by increased mortality, decreased feed conversion rate, reduced weight gain and increased variation at slaughter. Increased antimicrobial use, decreased welfare of the pigs and economic losses for the individual farmer and the swine industry are the consequences of enteric infections.

The most common causes of enteric infections in pigs post-weaning include enterotoxinogenic *Escherichia coli, Lawsonia intracellularis, Brachyspira hyodysentery, Brachyspira pilosicoli* and *Salmonella spp.* (Thomson, 2006).

Several clinical manifestations of enteric infections have been described in pigs. Clinical signs may be absent (subclinical infections) or include sudden death, anorexia, wasting, ill-thrift, dehydration, vomiting, ataxia, paleness, weakness, perineal irritation of the skin and various forms of diarrhoea (Thomson, 2006). Diarrhoea may be the only sign of an enteric infection (Straw et al. 2006). Diarrhoea may also be observed during outbreaks of (non-enteric) systemic diseases such as septicaemic salmonellosis and classical swine fever (Straw et al., 2006). Non-infectious causes of diarrhoea include soybean meal hypersensitivity and gastric ulceration (Straw et al., 2006).

Diarrhoea may be defined as a change in faecal consistency from normal to more fluid (Straw et al., 2006). Diarrhoea may be characterized by the consistency, colour and pH of faeces, and by the presence of blood, mucus or necrotic material (Straw et al., 2006).

There is no standardized method for characterizing faeces and/or diarrhoea, making comparison between various diarrhoea studies difficult. Examples of different faecal scoring systems can be seen in a number of studies (Guedes et al., 2002; Boesen et al., 2004; Starek and Bilkei, 2004). Further, consistency of faeces may vary according to the diet fed (Straw et al., 2006).

Evaluating the consistency of faeces and hence classification of pigs with diarrhoea are important when undertaking clinical examinations of diarrhoea at the individual and herd level. Standardized protocols are important in research and veterinary practice to obtain valid data and a high level of repeatability.

Assessment of whether a pig has diarrhoea and consistency of faeces are partly subjective. Variation of inter-observer agreement in detection of diarrhoea has been reported (Baadsgaard and Joergensen, 2003; Petersen et al., 2004; Pedersen et al., 2008a). A high level of agreement between observers in assessment of faecal consistency is reported within the field of human medicine (Allen et al., 1994; Bliss et al., 2001; Whelan et al., 2004). To our knowledge, agreement between observers in the assessment of faecal consistency has only been evaluated in one study in veterinary medicine. A large variation of agreement was reported between the participating observers (Pedersen et al., 2008a).

The objectives of the current study were to evaluate validity of a score system with 4 categories, text and pictures for assessment of consistency in faecal samples from pigs 2 to 10 weeks post weaning. Validity was evaluated in terms of repeatability (precision/random error) and accuracy (systematic error). Repeatability was evaluated through assessment of intra- and inter-observer agreement. Accuracy was evaluated through assessment of faecal dry matter content for faeces with different consistency scores.

Finally it was determine whether 4 categories, text and pictures could increase repeatability in terms of inter-observer agreement compared to a simple faecal consistency score system with 3 categories, no text and no pictures.

2. Materials and methods

2.1 Consistency score systems

A consistency score system with 4 descriptive categories and explanations in text and pictures, was developed, table 1.

To test the effect of the 4 categories, text and pictures this system was compared to a simple score. The simple score system consisted of 3 categories, no text and no pictures. The 3 categories were normal, loose or watery. No further definitions of these categories were given to the observers.

2.2 Design

All faeces samples were examined post collection to allow for assessment of both intra- and inter-observer agreement. Former studies have reported a high agreement between examinations of faeces samples pig-side versus post collection (Pedersen et al., 2008a).

One set of faeces samples was used to assess intra-, inter-observer agreement and accuracy for the score system with 4 categories, text and pictures.

In order to test the effect of the system with 4 categories, text and pictures, another set of faeces samples were collected and examined using the score system with 3 categories, no text and no pictures. The two sets of faeces samples were examined approximately 1 month apart. The first set was examined by 5 observers using the simple score system with 3 categories, no text and no pictures. The second set was examined by the same 5 observers using the system with 4 categories, text and pictures. On both occasions examination of the samples was done post collection.

2.2.1 Examination using 3 categories, no text and no pictures

Observer 1 collected a total of 99 faeces samples. The samples consisted of 33 normal, 33 loose and 33 fluid samples (assessed by observer 1 at collection).

The following day, the 5 observers examined the faeces samples. The observers were informed (by observer 1) about the consistency score system immediately before the start of the examination.

2.2.2 Examination using 4 categories, text and pictures

A diagram with explanations in text and pictures of faeces representing each of the 4 consistency scores was made, table 1.

The diagram was send by e-mail to the observers 4 days prior to examination. The observers were told to read the diagram to be familiar with the 4 categories, text and pictures prior to the examination.

At day 1, observer 1 collected 100 faeces samples. The samples consisted of 25 samples from each of the 4 consistency categories.

At day 2, the 5 observers examined the faeces samples two times in order to assess both intra- and inter-observer agreement. The observers were informed (by observer 1) about the consistency score system immediately before the start of the examination. The observers were told to examine the samples comparing to a large diagram (1 x 0.75m) with explanations in text and pictures of the consistency categories, table 1. The diagram was placed in front of the observers during the examination.

In order to avoid fatigue by the observers it was decided to space the two examinations min. 2 hours apart. The observers were physically separated during the study to avoid calibration.

At day 3 the faecal samples were transported to the laboratory and faecal dry matter content determined.

2.3 Clinical procedure

No calibration was allowed between observers prior to the study. During the study the observers were not allowed to discuss the examination of the faeces samples and the individual observers did the examinations one by one.

The individual observers examined the samples in the containers (post collection) and assessed the consistency scores. They were allowed to manipulate the faecal containers and touch the faeces with a spoon.

The observers examined the samples in random order. The identification number of the samples was not blinded to the observers.

2.4 The observers

It was intended to mimic a best case scenario in the study and try to obtain a high level of inter-observer agreement. Geographic differences in assessment of faecal consistency have been proposed by others (Pedersen et al., 2008a). Therefore 5 observers were selected at convenience from the same specialized swine practice. The 5 observers were participating in all parts of the study. They were all experienced swine veterinarians. Observer 1 (corresponding author) was also a researcher at University of Copenhagen.

2.5 Faeces samples

All faeces samples in the study were collected in the same herd by observer 1. A 4500 head weaner/grower herd was selected by convenience. The herd had a history of *Lawsonia intracellularis* associated diarrhoea and represented a modern Danish weaner/grower facility in relation to feeding, diet and housing. Feed was purchased pelleted and fed as restrictive wet feed. The diet was based on whet, barley and soybeans.

Faeces were collected in pens containing pigs between 2 and 10 weeks post weaning. The samples were collected with a clean spoon from fresh deposited faeces at the pen floor or directly from the pigs. Each sample consisted of approximately 10 gram of faeces. The samples were collected in plastic containers with a size of 5x5 cm to allow the faeces to retain its normal shape without adhering to the sides of the container. The containers were closed with a lid to avoid evaporation.

Composition of the samples and prevalence of the individual scores are considered to be important in design of agreement studies (Hoehler, 2000). Therefore the faeces samples were not selected at random by observer 1. It was intended to include all faecal consistency scores with equal representation in the study. Further, during selection of the samples it was intended to include different shades of faecal consistency with-in the same consistency score in order to obtain a complex composition of the sample population.

2.6 Faecal dry matter content:

Faecal dry matter content was determined for approximately 5 gram of each faeces sample. The microwave method reported by Pedersen et al., (2008b) was used.

2.7 Sample size considerations

The study had 3 different objectives.

Preliminary samples size considerations were based on expectations for intra-observer and pair wise inter-observer agreement. Results from the study reported by Pedersen et al., (2008a) were used. Approximately 100 faeces samples would provide an acceptable 95% confidence interval (allowable error = 0.075) for the intra- and inter-observer agreement estimates.

Similar, preliminary sample size for assessment of accuracy in relation to faecal dry matter content was considered using formulae for comparison between two groups. The mean faecal dry matter content of individual consistency scores may differ by 5 to 10% with a standard deviation of 10 to 15% (Carstensen, 2003; Kenworthy and Allen, 1966). A difference in means of 11%, SD=15%, power 80% and confidence 95% would require 23 samples in each group if a one sided test was used.

Preliminary sample size for assessment of the effect of 4 categories, text and pictures compared to 3 categories, no text and no pictures was considered. Expectations of difference between the two systems in the proportion of samples where all observers agreed were used. Using a one-sided test, power 80% and a 95% confidence, it would require two groups with 84 samples each to detect an improvement in overall agreement from 0.65 to 0.80 between the two score systems. An improvement of this magnitude was considered biological relevant.

Based on these preliminary sample size considerations approximately 100 samples for each of the two sets of faeces samples were considered to be acceptable. Also, it was considered that more than

100 faeces samples in each set would lead to fatigue for the observers in performing the examinations. This could potential bias the results of the study.

The number of included observers would preferably be 25 or more to obtain accurate estimates of the study objectives. However, it would be impractical to get a large number of observers to examine 100 faeces samples twice within the same day. Therefore it was decided to include 5 observers which made it possible to conduct the examinations within one day.

2.8 Data analysis

2.8.1 Data management

The dataset was checked for missing values, extreme values and misclassifications. Faeces samples with missing values, extreme values or misclassification would be deleted from the dataset.

In order to analyse the data to fulfil the study objectives a set of new variables were defined for each faeces sample.

2.8.1.1 Definitions of new variables

Intra-observer agreement:

A dichotomous variable was defined. For each observer the two examinations of the same faeces sample using the system with 4 categories, text and pictures were grouped into one variable. If the individual observer had the same consistency score for a sample in the two examinations the variable was classified as “yes”. If the observer had a different consistency score for a sample the variable was classified as “no”.

Merge of score 1 and 2:

An ordinal variable was defined. The consistency scores of the system with 4 categories, text and pictures were grouped into 3 categories matching the system with 3 categories, no text and no pictures. Samples with consistency scores 1 or 2 were considered to be normal and were classified as “normal”. Samples with consistency score 3 were considered to be loose and were classified as “loose”. Samples with consistency score 4 were considered to be watery and were classified as “watery”.

For the samples examined with the system with 3 categories, no text and no pictures the variable was classified according to the original consistency score as normal, loose or watery.

<u>Diarrhoea</u>:

A dichotomous variable was defined grouping the consistency scores into two categories. For the system with 4 categories, text and pictures the samples with consistency scores 1 or 2 was considered not to be diarrhoeic and were classified as “no”. Samples with consistency scores 3 or 4 was considered to be diarrhoeic and were classified as “yes”.

For the system with 3 categories, no text and no pictures the samples scored as normal were classified as “no”. A sample scored as loose or watery was classified as “yes”.

<u>Agreement between all 5 observers, original score system</u>:

A dichotomous variable was defined grouping all observers into one variable. If all 5 observers agreed on the consistency score for a sample the variable was classified as “yes”. If on or more observers had a different consistency score for a sample the variable was classified as “no”. For the system with 4 categories, text and pictures the results of the first examination for each observer was used.

Agreement between all 5 observers, merge of score 1 and 2:

A dichotomous variable was defined grouping all observers into one variable. If all 5 observers had the same outcome when consistency scores 1 and 2 were merged for a sample the variable was classified as “yes”. If on or more observers had different outcomes for a sample the variable was classified as “no”.

Agreement between all 5 observers, diarrhoea:

A dichotomous variable was defined grouping all observers into one variable. If all 5 observers had the same outcome for diarrhoea for a sample the variable was classified as “yes”. If on or more observers had different outcomes for a sample the variable was classified as “no”.

2.8.2 Descriptive analysis

2.8.2.1 Intra-observer agreement

Descriptive analysis of intra-observer agreement was performed for the two examinations using the system with 4 categories, text and pictures. Prevalence for each consistency score and diarrhoea was calculated for the two examinations for each observer. Intra-observer agreement for each observer was calculated for consistency scores and diarrhoea. Intra-observer agreement between the two examinations was defined as the number of samples where the individual observer had the same score at the two examinations divided with the total number of samples.

2.8.2.2 Inter-observer agreement

Descriptive analysis of inter-observer agreement was performed for the first examination using the system with 4 categories, text and pictures. Prevalence for each consistency score and diarrhoea were calculated for each observer. Inter-observer agreement for each pair of observers was calculated for consistency scores and diarrhoea. Inter-observer agreement between two observers was defined as the number of samples where the two observers had the same score divided with the total number of samples.

2.8.2.3 Accuracy

Descriptive analysis of accuracy for the system with 4 categories, text and pictures was evaluated in relation to faecal dry matter content. For consistency scores and diarrhoea different plots and descriptive measures were computed stratified by observer.

2.8.2.4 Effect of 4 categories, text and pictures

A descriptive analysis of the effect of 4 categories, text and pictures compared to the simple system with 3 categories, no text and pictures was performed. For the system with 4 categories, text and pictures the results of the first examination for each observer was used for the analysis. A series of 2 by 2 tables was constructed with one factor being the score system (4 categories, text and pictures or 3 categories without text and pictures) and the other factor being agreement between all observers for the original score system, merge of score 1 and 2 or diarrhoea.

2.8.3 Statistical analysis

2.8.3.1 Intra-observer agreement

Intra-observer agreement was evaluated for the system with 4 categories, text and pictures for both consistency score and diarrhoea. Cohen’s kappa for each observer was calculated using the freq procedure in SAS version 9.1.

The effect of faecal dry matter content on intra-observer agreement was evaluated for each observer. This was performed by evaluating the relation between the intra-observer agreement (dependent variable) and faecal dry matter content (independent variable) by logistic regression using the genmod procedure in SAS version 9.1.

2.8.3.2 Inter-observer agreement

Inter-observer agreement was evaluated for the system with 4 categories, text and pictures using the first examination of the samples. Both consistency score and diarrhoea was evaluated. Cohen's kappa for each pair of observers was calculated using the freq procedure in SAS version 9.1.

The effect of faecal dry matter content on inter-observer agreement was evaluated for agreement between all observers. This was performed by evaluating the relation of agreement for all 5 observers using original score system (dependent variable) and faecal dry matter content (independent variable) by logistic regression using the genmod procedure in SAS version 9.1.

2.8.3.3 Accuracy

Accuracy of the system with 4 categories, text and pictures was evaluated in relation to faecal dry matter content. An equal contribution to the analysis from each observer was intended. Observer 5 only had one examination of the samples. Therefore it was decided to do the analysis only with the results from the observers' first examination of the samples.

For each observer the mean faecal dry matter content for each consistency level was determined using analysis of variance. The mixed procedure in SAS version 9.1 was used.

The results of the analysis of variance were used to calculate an overall faecal dry matter mean for each consistency score by taking the average of all observers. For each consistency score a faecal dry matter cut-off value was determined. The midpoint between the mean faecal dry matter content of two consistency scores was used to define the cut-off values. The cut-off values were used to determine the true consistency score (4 categories) for each faeces sample. For the individual observers the proportion of correctly classified samples was calculated. An overall mean for the proportion of correctly classified samples was calculated by taking the average of all observers. The same analysis was performed for the diarrhoea in relation to faecal dry matter content.

2.8.3.4 Effect of 4 categories, text and pictures

The 4 categories, text and pictures were compared to the simple system with 3 categories, no text and pictures. Each score system was applied to a different set of faeces samples as described. For the system with 4 categories, text and pictures the results of the first examination for each observer was used for the analysis. A logistic analysis was applied to test the association between the score system as the independent variable (4 categories, text and pictures or 3 categories without text and pictures) and either agreement between all 5 observers, original score system; agreement between all 5 observers, merge of score 1 and 2 or agreement between all 5 observers, diarrhoea as the dependent variable. The genmod procedure in SAS version 9.1 was used.

3. Results

Observer 5 did not perform the last of the two examinations using the system with 4 categories, text and pictures.

3.1 Intra-observer agreement

Intra-observer agreement was evaluated for two examinations by observer 1-4 using the system with 4 categories, text and pictures. The results are displayed in table 2. Only minor differences in prevalence of consistency scores and diarrhoea were observed for each observer from one examination to the next, figure 1 and 2. The larges observed difference in consistency score prevalence between two examinations was 0.10 and the smallest was 0. For diarrhoea the larges observed difference in prevalence was 0.10 and the smallest was 0.01.

Ranking the observers according to the intra-observer agreement gave the same order for both the consistency score and the diarrhoea.

Effect of faecal dry matter content on intra-observer agreement was evaluated for the consistency scores. A logistic regression model showed that a decrease in faecal dry matter content was associated with a significant increase in intra-observer agreement for observer 4 and a tendency for observer 1, table 3. Observer 1 and 4 had the highest level of intra-observer agreement among the 4 observers. Assumptions for logistic regression were evaluated and fulfilled.

3.2 Inter-observer agreement

Inter-observer agreement was evaluated for 5 observers using the first examination of the faeces samples with the system having 4 categories, text and pictures. The results are displayed in table 4. The larges observed difference in consistency score prevalence between two observers was 0.17 and the smallest was 0. For diarrhoea the larges observed difference in prevalence was 0.19 and the smallest was 0.01, figure 3 and 4.

Ranking the pairs of observers according to decreasing inter-observer agreement for both the consistency score and diarrhoea gave the same order for the first two and the last observer pair.

For the consistency score all observers agreed on only 48% of the samples. After dichotomization into diarrhoea the observers agreed on 78% of the samples.

Effect of faecal dry matter content on inter-observer agreement was evaluated for the consistency scores. A logistic regression model showed no significant association between faecal dry matter content and inter-observer agreement, table 3. Assumptions for logistic regression were evaluated and fulfilled.

3.3 Accuracy

Accuracy of the system with 4 categories, text and pictures was evaluated in relation to faecal dry matter content. The faecal dry matter content for the samples was between 6.2% and 28% with a mean of 18.0%. The relation between faecal consistency score and dry matter content are displayed in figure 5 for each of the 5 observers. The analysis of variance showed that for each observer there was a significant difference (p-value < 0.05) in the mean faecal dry matter content for each consistency score. Assumptions for analysis of variance were evaluated and fulfilled for all but observer 2. Mean faecal dry matter content for each consistency score are displayed in table 5.

Faecal dry matter cut-off values for each consistency score were determined. Twenty seven percent of the faeces samples were classified as consistency score 1 (dry matter content > 22.6%), 25% classified as score 2 (dry matter content > 18.8%), 22% as score 3 (dry matter content > 13.1%) and 26% as score 4.

The mean proportion of correctly classified samples for all observers was 0.80 (min. = 0.69, max. = 0.89). The proportion of correctly classified samples for each consistency score are displayed in figure 6. The proportions were highest for faeces samples classified as 3 (mean = 0.91) followed by samples classified as 1 (mean = 0.85), 4 (mean = 0.84) and 2 (mean = 0.59).

The mean proportion of correctly classified samples for diarrhoea was 0.92 (min. = 0.85, max. = 0.94). The proportions were highest for faeces samples classified as diarrhoeic (mean = 0.96, min. = 0.90, max. = 1.0) compared to non diarrhoeic samples (mean = 0.87, min. = 0.71, max. = 0.98). Observers with a high proportion of correctly classified diarrhoeic samples had a lower proportion of correctly classified non diarrhoeic samples and vice versa.

3.4 Effect of 4 categories, text and pictures

A total of 98 samples examined with the system having 3 categories, no text and no pictures were included in the analysis. A total of 100 samples examined using the system with 4 categories, text and pictures were included in the analysis.

The proportion of samples where all observers agreed for each system are displayed in table 6. The results of the logistic analysis are displayed in table 7. Assumptions for logistic analysis were evaluated and fulfilled.

The system with 3 categories, no text and no pictures gave a significant higher proportion of samples where all the observers agreed. Except for the situation where score 1 and 2 in the 4 category system were merged to obtain a matching 3 category system. In that situation no effect of score system existed.

4. Discussion

4.1 Study design

This study probably represents a best case scenario when it comes to intra- and inter-observer agreement. All observers were experienced swine veterinarians and were used to examine faeces as part of their job. They were working in the same veterinary practice so any geographic differences in faecal consistency in relation to feeding, medication and diseases should be eliminated. Further, one would expect that text and pictures would increase both intra- and inter-observer agreement, because of the possibility to compare the faeces samples with the diagram during the examination. On the other hand the observers were more used to examine faeces lying on the pen floor and the intra- and inter-observer agreement might have been higher if the examination had been possible to do in the pens.

In relation to intra-observer agreement the two examinations were spaced 3.5 to 10 hours apart for the individual observers. The identification number of the samples was not blinded to the observers.

The observers would potentially be able to remember some of the individual faeces samples, because of the short time between the two examinations. However, the order of the samples was random and it seemed unlikely that an observer would be able to remember a specific faeces sample among 100 samples. On the other hand, the faeces samples could potential change appearance between the two examinations leading to a reduced intra-observer agreement. In fact all observers tended to score more samples as 3 or 4 on the second examination. However, no association between intra-observer agreement and number of hours between the two examinations were observed (data not shown).

This study represents intra-observer agreement within the same day. It is not possible to conclude that the same level of intra-observer agreement would be observed if two examinations were spaced more than a day, a month or even more apart. In fact it seems reasonably to believe that a score system with text and pictures would be beneficial if two examinations in a study are executed on separate days, months or years.

Fatigue may have been a problem during examination though the number of faeces samples was restricted to 100. This could potentially bias the study. Unfortunate the design did not allow for investigation of this aspect.

Composition of study population has been reported to be important in agreement studies (Hoehler, 2000) making comparisons between studies difficult. Similar, others have reported a higher tendency to rule disease out than in (Baadsgaard and Jørgensen, 2003), which could influence results of agreement studies. We investigated this aspect by evaluating intra- and inter-observer agreement for consistency scores in relation to the true state of the faeces samples in terms of faecal dry matter.

Under the conditions of this study one observer had an increasing intra-observer agreement with decreasing faecal dry matter content for the faeces samples. Another observer had a tendency for the same association. This implies that for some observers the proportion of samples getting the same classification in two examinations tends to increase for more fluid faeces samples.

No association was observed between faecal dry matter content and the proportion of faeces samples where all observers had the same consistency score. This implies that for the current score system agreement between all observers was independent of faecal consistency.

4.2 Repeatability and accuracy

For the score system with 4 categories, text and pictures repeatability was evaluated in terms of intra- and inter-observer agreement. Accuracy was evaluated in terms of faecal dry matter content for each consistency score. Variation in the observed accuracy between observers also contributes to the interpretation of the systems repeatability. For that reason both repeatability and accuracy are discussed together.

Within observers the difference in prevalence for the individual consistency scores and diarrhoea between two examinations was on average 0.04 and 0.05 respectively. The larges difference was 0.10 for both consistency scores and diarrhoea. Between observers the difference in prevalence for the individual consistency scores and diarrhoea was on average 0.08 and 0.09 respectively. The larges observed difference was 0.17 (consistency score) and 0.19 (diarrhoea). For comparison, a 95% confidence interval for prevalence estimates would be in the range of 0.10 to 0.20 with a sample size of 100.

Using the current score system it seems that variation within the same observer may be ignored when estimating prevalence of consistency scores and diarrhoea. In relation to variation between observers the large difference in prevalence estimates would be a problem when estimating consistency scores or diarrhoea prevalence in research studies. This implies that the score system with 4 categories, text and pictures could not avoid variation between observers.

Agreement and Cohen’s Kappa were higher for diarrhoea than the 4 consistency categories both within and between observers. This was expected since more categories places more samples on the boundaries between two categories.

The current study shows that 4 descriptive categories with text and pictures do not eliminate problems of intra- and inter-observer agreement. Both intra- and inter-observer agreement must be taken into consideration during research situations where classification of individual samples is important. This implies especially to situations where the 4 categories are not dichotomised during analysis.

In this study we used faecal dry matter as an objective measure of the true state of the faeces samples. Faecal consistency changes according to diet feed (Straw et al., 2006) and faecal dry matter content may not be the only determinant of faecal consistency. This aspect should be taken into consideration in interpretation of accuracy in the current study. Accuracy of the score system was evaluated in relation to faecal dry matter content and not necessary faecal consistency.

The mean faecal dry matter content was significant different between the individual consistency scores for all observers. The small difference between faecal dry matter content for samples scored as 1 and 2 indicate that these two categories may be merged without los of information in designing consistency categories.

Faecal dry matter cut-off values were determined and used to classify the faeces samples. We observed on average 80% accuracy in classification of faeces samples can be expected with the score system having 4 categories, text and pictures. An accuracy of this magnitude may not be considered acceptable. Further, a large variation in accuracy between observers and consistency scores was observed adding to the lack of repeatability for the score system.

For assessment of diarrhoea the observed accuracy may be considered acceptable. Further, variation between observers was observed but to less extends than for consistency scores. Considering faecal dry matter content as the gold standard the diagnostic sensitivity and specificity for the observers in assessment of diarrhoea can be calculated. Mean diagnostic sensitivity and specificity in the current study were 0.96 and 0.87 respectively which may be considered acceptable in most situations. However, a large variation in diagnostic sensitivity and specificity was observed between the observers adding to the lack of repeatability for the score system.

4.3 Effect of 4 categories, text and pictures

Under the conditions of the current study a simple system with 3 categories, no text and no pictures performed better than a system with 4 categories, text and pictures. This was expected when comparing 3 versus 4 categories, since more categories places more samples on the boundaries between two categories. It was not expected that the system with 3 categories, no text and no pictures would be able to match or even perform better than the 4 categories with text and pictures when the number of categories were equalized in the analysis. One explanation could be that the current study represents the best case scenario. This may remove any effect of pictures and text. The observers explained after the study that they found it more difficult to do the examinations comparing to text and pictures. This may be another explanation. More intensive training in use of the text and pictures prior to examination may give a different result.

For both score systems in the current study we observed a higher inter-observer agreement and Cohen’s Kappa value for assessment of diarrhoea than reported by Pedersen et al., (2008a). The study by Pedersen et al., (2008a) and the current study have similar designs except for the applied score systems. It seems that 3 or 4 categories in score systems can increase the inter-observer agreement for assessment of diarrhoea.

5. Conclusion

Validity of 4 categories with text and pictures for scoring of faecal consistency in pigs was assessed in a best case scenario without calibration between observers.

The current study shows that 4 descriptive categories with text and pictures do not eliminate problems of low repeatability within and between observers.

An unacceptable accuracy for consistency score classification of faeces samples was observed.

Accuracy was considered acceptable after dichotomization of consistency scores. Variation in accuracy between observers contributed to lack of repeatability for the score system.

A decreased repeatability was observed for the system with 4 categories, text and pictures compared to a simple system with 3 categories, no text and no pictures.

More objective measures of faecal consistency may be more appropriate in research studies.

6. Acknowledgements

This research was funded by the University of Copenhagen. I thank the swine practice Oe-Vet for their assistance with examination of the faeces samples. I thank the people at the Veterinary Institute for providing laboratory facilities. I thank Klaus H. Pedersen for his assistance with assessment of faecal dry matter.

7. Literature

1. Allen, U.D., Deadman, L., Wang, E.E., 1994. Standardizing the assessment of diarrhea in clinical trails: results of an interobserver agreement study. Acta Paediatrica. 83 (2), 179-182.

2. Baadsgaard, N.P, Jørgensen, E., 2003. A Bayesian approach to the accuracy of clinical observations. Prev. Vet. Med. 59, 189-206.
3. Bliss, D.Z., Larson, S.J., Burr, J.K., Savik, K., 2001. Reliability of a stool consistency classification system. JWOCN. 28 (6), 305-313.
4. Boesen, H.T., Jensen, T.K., Schmidt, A.S., Jensen, B.B., Jensen, S.M., Moller, K., 2004. The influence of diet on Lawsonia intracellularis colonization in pigs upon experimental challenge. Veterinary microbiology. 103 (1/2), 35-45.
5. Carstensen, L. 2003. Post weaning diarrhoea susceptibility in piglets in relation to supplementary creep feeding during the suckling period and selected innate immune factores. Ph.D thesis. The Royal Veterinary and Agricultural University. Denmark.
6. Guedes, R.M.C., Gebhart, C.J., Winkelman, N.L., Mackie Nuss, R.A.C., Marsteller, T.A., Deen, J., 2002. Comparison of different methods for diagnosis of porcine proliferative enteropathy. Canadian Journal of Veterinary Research. 66 (2), 99-107.
7. Hoehler, F.K., 2000. Bias and prevalence effects on kappa viewed in terms of sensitivity and specificity. J. Clin. Epidemiol. 53, 499-503.
8. Kenworthy, R., Allen, W.D., 1966. The significance of *Escherichia coli* to the young pig. J. Comparative Pathology. 76, 31-44.
9. Pedersen, K.S., Holyoake, P., Stege, H., Nielsen, J.P., 2008a. Inter-observer variation in assessment of diarrhoea in pigs and predictive value of selected clinical signs in assessment of pigs with diarrhoea. In Press.
10. Pedersen, K.S., Pedersen, K.H., Stege, H., Nielsen, J.P., 2008b. Assessment of faecal dry matter in pig faeces using microwaves. In Press.
11. Petersen, H.H., Enøe, C., Nielsen, E.O., 2004. Observer agreement on pen level prevalence of clinical signs in finishing pigs. Prev. Vet. Med. 64, 147-156.

12. Starek, M., Bilkei, G., 2004. Sows seropositive to Lawsonia intracellularis (LI) influence performance and LI seropositivity of their offspring. Acta Veterinaria Brno. 73 (3), 341-345.

13. Straw, B.E., Dewey, C.E., Wilson, M.R., 2006. Differential diagnosis of disease. In: Straw, B.E., Zimmerman, J.J., D'Allaire, S., Taylor, D.J. (Eds.), Diseases of swine. 9th ed. Blackwell Publishing, Iowa, pp. 241-283.

14. Thomson, J.R., 2006. Diseases of the digestive system. In: Straw, B.E., Zimmerman, J.J., D'Allaire, S., Taylor, D.J. (Eds.), Diseases of swine. 9th ed. Blackwell Publishing, Iowa, pp. 37-55.

15. Whelan, K., Judd, P.A., Taylor, M.A., 2003. Defining and reporting diarrhoea during enteral tube feeding: do health professionals agree? J. of Hum. Nutr. Dietet. 16, 21-26.

16. WHO report, 2003. Impacts of antimicrobial growth promoter termination in Denmark. WHO/CDS/CPE/ZFK/2003.1, p.32.

8. Appendix

Table 1. Consistency score with 4 categories, text and pictures

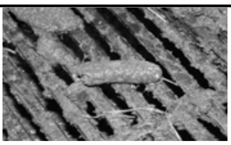
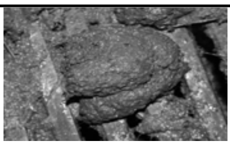
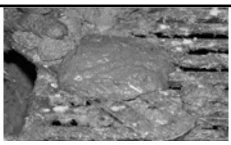
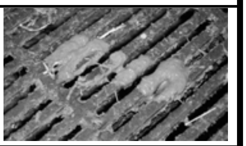
Score	1 Firm and shaped	2 Soft and shaped	3 Loose	4 Watery
Picture				
Texture	Firm. Vary in hardness.	Vary in softness. Like peanut butter	Mush. Often shining surface	Vary form gruel to water.
Shape	Sausage	Vary form sausage shape to small piles	Tends to level with surface. Does not flow through or flows slowly through slatted floors.	Levels with surface. Flows through slatted floors.
In container	Preserves original shape.	Does not flow when container is rotated. Preserves original shape.	Inert when container is rotated. Merges and cover up button of container in most cases.	Flows easy when container is rotated. Merges and cover up button of container.

Table 2
Intra-observer agreement (2 examinations)
for 4 observers using consistency score with 4 categories, text and pictures

Consistency score	Mean	Min	Max
Intra-observer agreement	0.82	0.72	0.91
Cohen's Kappa	0.76	0.61	0.88
Diarrhoea (score 3+4)			
Intra-observer agreement	0.93	0.90	0.95
Cohen's Kappa	0.86	0.80	0.90

Table 3
Logistic regression of association between
faecal dry matter content and intra- or inter-observer agreement for consistency score

Dependent variable	Independent variable	Estimate	OR*	95% Cl	p-value
Intra-observer agreement (observer 1)	Dry matter content	-0.15	2.10	0.95-4.50	0.07
Intra-observer agreement (observer 2)	Dry matter content	0.02	0.89	0.61-1.30	0.56
Intra-observer agreement (observer 3)	Dry matter content	-0.06	1.40	0.86-2.20	0.18
Intra-observer agreement (observer 4)	Dry matter content	-0.21	2.80	1.30-6.00	0.01
Agreement between all 5 observers	Dry matter content	-0.04	1.20	0.86-1.70	0.27

* OR for outcome "yes" at a 5% decrease in faecal dry matter content

Table 4
Inter-observer agreement for 5 observers using consistency score with 4 categories, text and pictures

Consistency score	Mean*	Min	Max
Pair wise inter-observer agreement	0.73	0.61	0.90
Cohen's Kappa	0.64	0.48	0.87
Diarrhoea (score 3+4)			
Pair wise inter-observer agreement	0.89	0.81	0.95
Cohen's Kappa	0.78	0.63	0.90

* Mean of 10 pair wise comparisons

Table 5
Percent mean dry matter content i relation to consistency score for 5 observers

Consistency score	Mean	Min	Max
1	24.00	23.30	24.40
2	21.20	20.40	22.60
3	16.40	14.30	18.00
4	9.70	8.90	10.20

Table 6
Proportion of samples with agreement between all 5 observers

Variable*	4 categories, text and pictures	3 categories, no text and no pictures
Agreement all observers, original score	0.48	0.66
Agreement all observers, merge score 1+2	0.65	0.66
Agreement all observers, diarrhoea	0.78	0.89

* See section on data management for explanation

Table 7

Logistic analysis of effect on inter-observer agreement of 4 categories, text and pictures compared to 3 categories, no text and no pictures

Dependent variable*	Independent variable	Estimate	OR**	95% Cl	p-value
Agreement all observers, original score	Score system	-0.76	0.47	0.26-0.83	0.01
Agreement all observers, merge score 1+2	Score system	-0.06	0.94	0.52-1.70	0.84
Agreement all observers, diarrhoea	Score system	-0.80	0.45	0.20-0.98	0.05

* See section on data management for explanation
** OR for outcome "yes" at 4 versus 3 categories

Figure 1. Consistency prevalence for two examinations

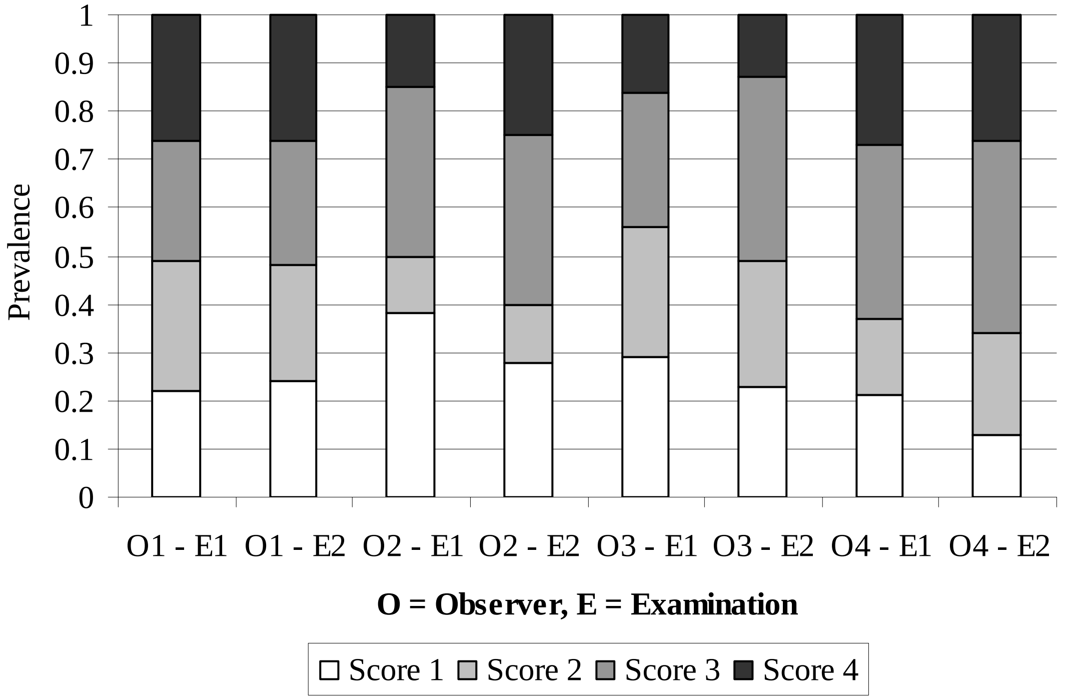


Figure 2. Prevalence of diarrhoea for two examinations

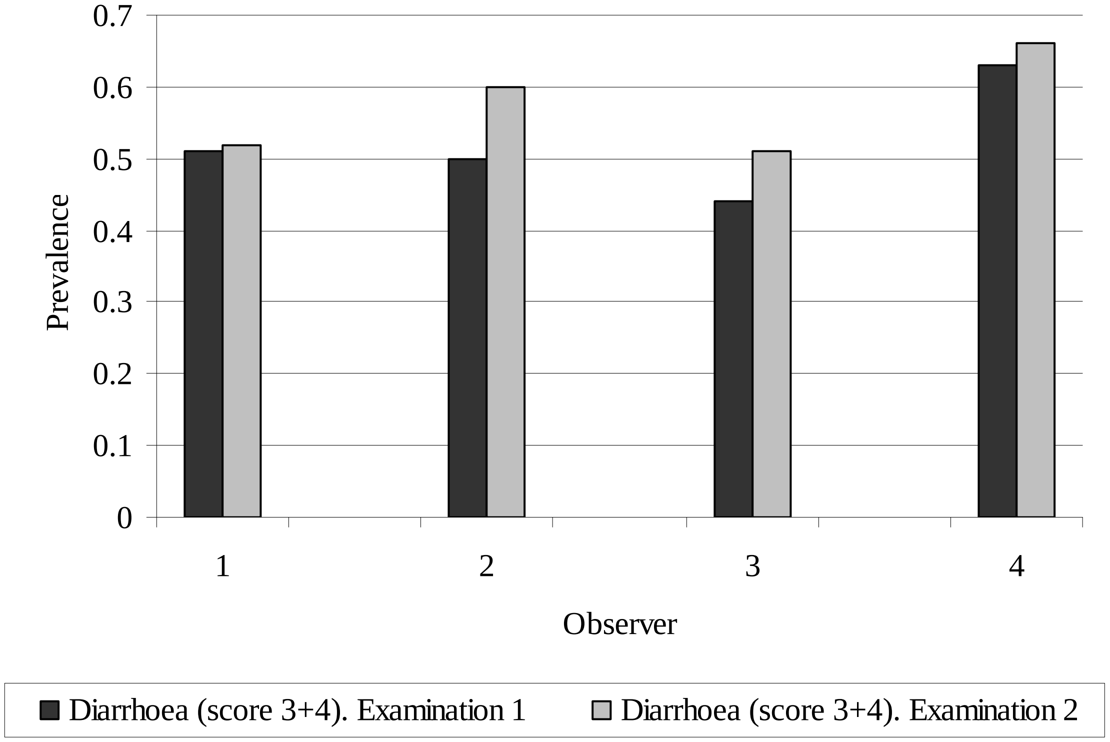


Figure 3. Consistency prevalence for each observer

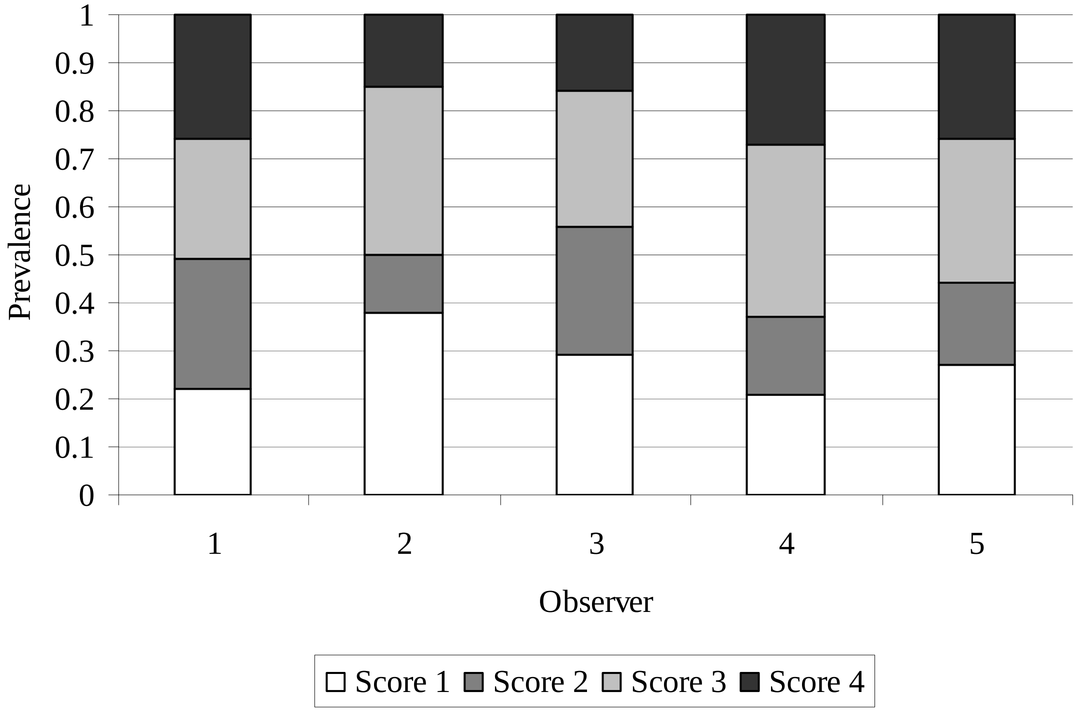


Figure 4. Prevalence of diarrhoea for each observer

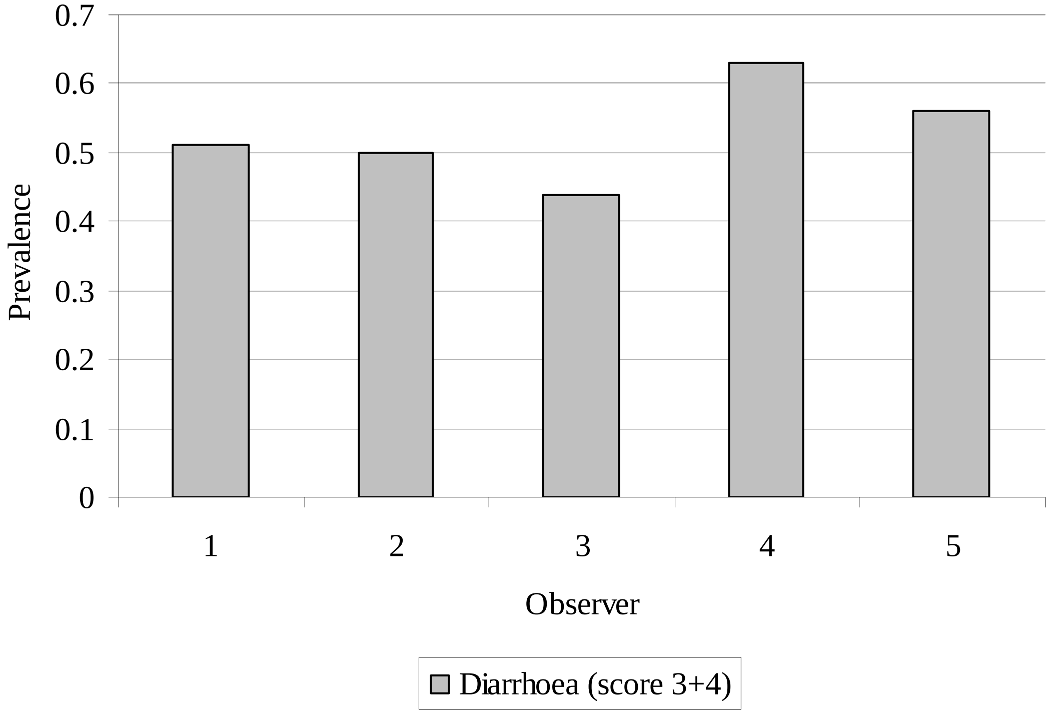
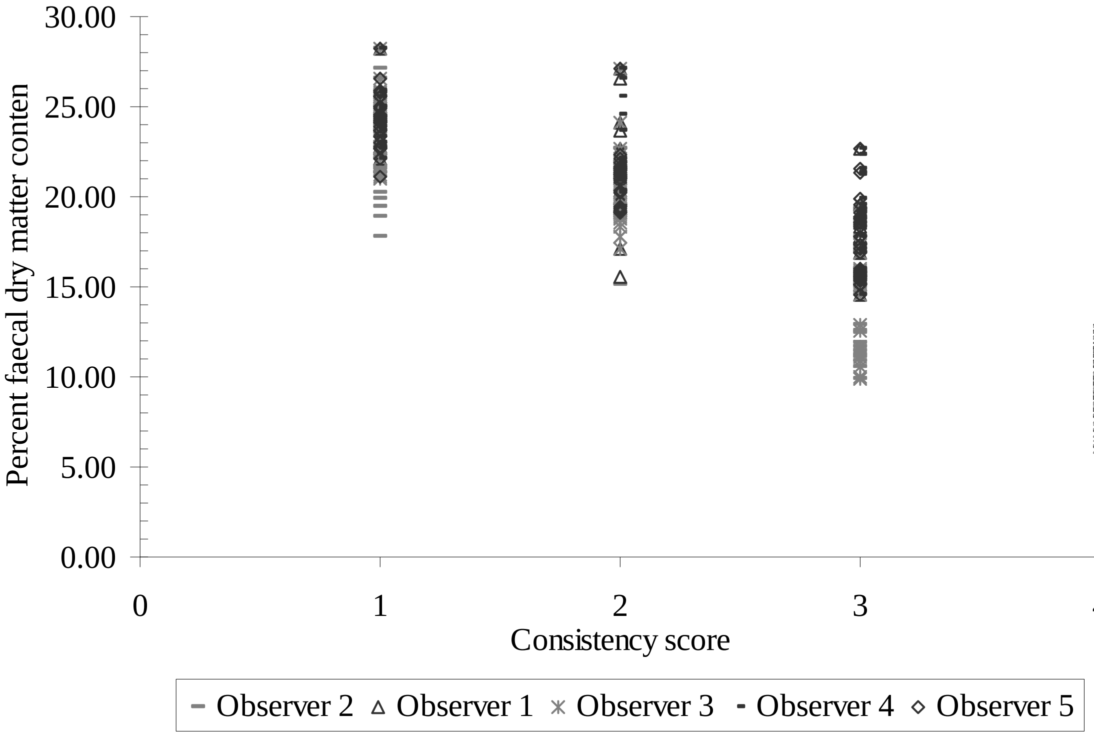
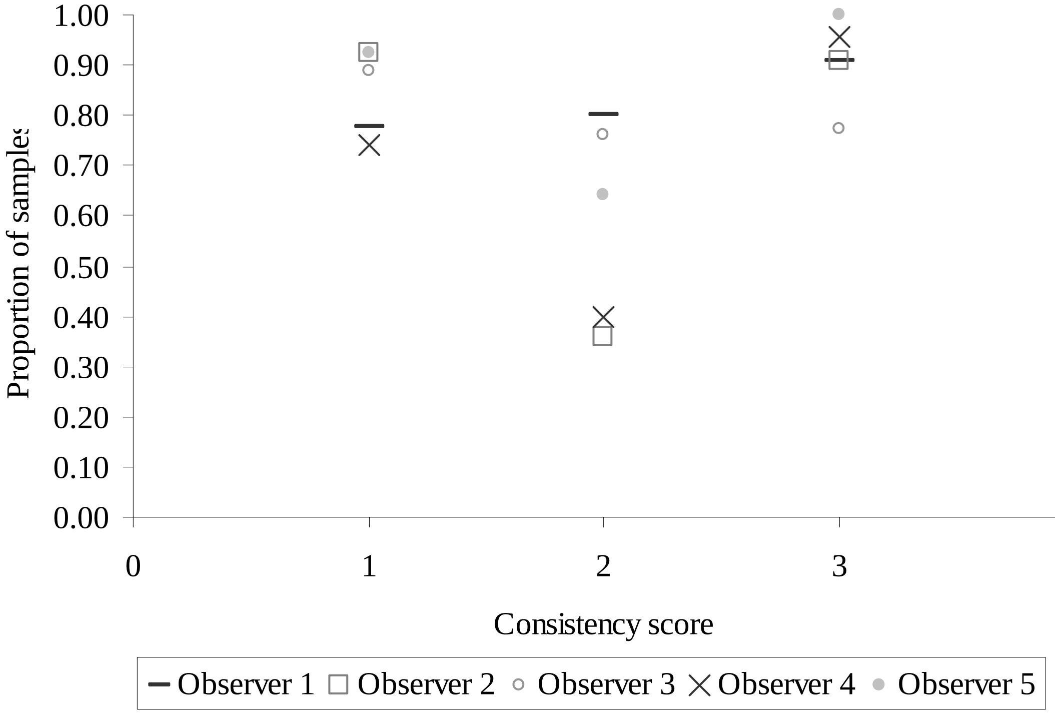


Figure 5. Consistency score and faecal dry matter content



599

Figure 6. Proportion of correctly classified samples



600