ON-FARM ANIMAL WELFARE ASSESSMENTS - HANDLING CHALLENGES OF IMPLEMENTING THE WELFUR-MINK PROTOCOL ON A LARGE SCALE

ANNA FELDBERG MARSBØLL
PhD THESIS · SCIENCE AND TECHNOLOGY · 2018

AARHUS UNIVERSITY
ON-FARM ANIMAL WELFARE ASSESSMENTS - HANDLING CHALLENGES OF IMPLEMENTING THE WELFUR-MINK PROTOCOL ON A LARGE SCALE

ANNA FELDBERG MARSBØLL
PhD THESIS · SCIENCE AND TECHNOLOGY · 2018

AARHUS UNIVERSITY

Department of Animal Science
Aarhus University
Blichers Allé 20
8830 Tjele
Denmark
Supervisors

**Steen H. Møller** (main supervisor)
Senior Researcher, Department of Animal Science - Epidemiology and Management, Aarhus University, Denmark.

**Tine Rousing** (co-supervisor)
Senior Researcher, Department of Animal Science - Epidemiology and Management, Aarhus University, Denmark.

Assessment committee

**Mette S. Herskin** (chairman of the committee)
Senior Researcher, Department of Animal Science - Behaviour and Stressbiology, Aarhus University, Denmark.

**Jaakko Mononen**
Principal Scientist, Production systems, Natural Resources Institute Finland (Luke), Finland.

**Antonio Velarde Calvo**
Head of the Animal Welfare programme, Institute of Agrifood Research and Technology (IRTA), Spain.

Cover photos: Jens Bach
Preface

The present PhD thesis entitled ‘On-farm animal welfare assessments – handling challenges of implementing the WelFur-Mink protocol on a large scale’ is submitted to fulfil the requirements for the degree of Doctor of Philosophy at the Faculty of Science and Technology, Aarhus University, Denmark. In accordance with the rules of the Graduate School of Science and Technology, this thesis includes parts of the progress report which was prepared for the qualifying examination.

The research presented in this thesis is based on a PhD study carried out during the period from April 2015 to August 2018 in the research group Epidemiology and Management, Department of Animal Science, Faculty of Science and Technology, Aarhus University, Denmark. The research was part of the WelFur III research agreement with Fur Europe (an umbrella organisation for the European fur sector) which is a project supporting the implementation of the animal welfare assessment system WelFur-Mink in Europe through different research, teaching and development tasks. One-third of the PhD study was funded by a PhD grant from Aarhus University and two-thirds by the WelFur III research agreement. The PhD study was on hold for five months due to other tasks related to the WelFur III research agreement.

The PhD study is centred around the development, tests and implementation of the WelFur-Mink assessment system. As a part of the PhD study, I have participated in several WelFur related activities in Denmark and abroad. These included attending WelFur science group meetings, updating the WelFur-Mink assessment guidelines, testing the WelFur-Mink welfare assessment protocol in Latvia and Greece, testing the WelFur-Mink mobile application for data collection and educating WelFur-Mink assessors. I have also visited researchers in Finland who work with WelFur-Fox to get inspiration for the training of WelFur-Mink assessors as well as for my own research and Fur Europe in Brussels to experience the political side of the WelFur project. In addition, I have been responsible for teaching bachelor and master students at Aarhus University, co-supervision of students and presentations for visitors at AU-Foulum. I have also completed activities corresponding to 29 ECTS.

The PhD study resulted in manuscripts for three papers. At present, all three manuscripts are under revision (Note: Since this thesis was submitted in September 2018, revised
versions of the manuscripts for the first and second paper have been accepted for publication. However, according to the rules of the Graduate School of Science and Technology, the manuscripts are included in this thesis in the stage they were in when the thesis was submitted) and are included in this thesis with permission from the journal editors.

The first paper is entitled ‘The representativeness of a semi-random sampling method for animal welfare assessments on mink farms’. I was responsible for drafting the manuscript with contributions from the co-authors. I was also responsible for study design, data collection, data management, developing a program to simulate the sampling method and for generating the results. All authors were responsible for the development and revision of the sampling method.

The second paper is entitled ‘Changes in the welfare of mink (Neovison vison) with date of assessment in the winter and growth periods have limited effects on the overall WelFur categorisation’. I was responsible for drafting the manuscript with contributions from the co-authors. I was also responsible for data management and data analysis - the latter with help from Leslie Foldager, Aarhus University. Study design and data collection were conducted before the PhD study was initiated.

The third paper is entitled ‘If stereotypic behaviour cannot be observed before feeding in WelFur-Mink, a reliable conversion factor needs to be developed’. I was responsible for drafting the manuscript with contributions from the co-authors. I was also responsible for study design, data collection, data management and data analysis. The manuscript also included results from welfare assessments in Europe which were made available by Fur Europe.

Anna Feldberg Marsbøll
September 2018
Acknowledgements

First, I want to express my gratitude to my main supervisor, Steen Henrik Møller. Thanks for believing in me throughout the process and for insisting on having fun along the way. I also thank my co-supervisor, Tine Rousing, for bringing a broader perspective and support at all the right times.

Thanks to the farmers who participated in the project. Without your contributions, this study would not have been possible. I also thank Kopenhagen Fur for allowing me to use data from the Danish test of WelFur. A special thanks to Bente Krogh Hansen, Kopenhagen Fur, for introducing me to mink production and for providing the farmers point of view. Also thanks to Elna Louise Krogh Nielsen and Maria Blæsbjerg-Obitsø, both Kopenhagen Fur, for sharing your knowledge on mink production and for challenging my thoughts and ideas. I also thank Mette Lykke Nielsen and Hanna Leiponen-Syyrakki, both Fur Europe, for welcoming me in Brussels last year and for providing data from the WelFur-Mink assessments in Europe. Also thanks to Tarja Koistinen and Jaakko Mononen, both Luke, for welcoming me in Finland last year. I enjoyed our many discussions on calculations and aggregations and appreciated you introducing me to farmed foxes and finnraccoon.

Thanks to all my colleges in the Epidemiology and Management research group at Aarhus University for an inspiring research environment. A special thanks to Britt Henriksen for your support in addition to our professional collaboration. Also thanks to the many other colleagues at Aarhus University who have supported me and given indispensable help. Especially to Lena Karina Hinrichsen and Jens Malmkvist for the many valuable discussions, to Leslie Foldager for statistical assistance and to Tina Albertsen and Lotte Hansen for proofreading.

Finally, I thank my friends and family for being both patient and supportive. I appreciate that you tried to understand why I find the field of welfare assessment and mink production so interesting.

Tak, Thanks, Kiitos, Grazie, Takk.
# Table of contents

Summary ................................................................................................................................. 7  
Sammendrag (Danish summary) ......................................................................................... 10  
List of included papers ......................................................................................................... 13  
1 Introduction .............................................................................................................................. 14  
   1.1 General introduction ........................................................................................................... 14  
   1.2 Aim and hypotheses .......................................................................................................... 18  
   1.3 Outline of the thesis ......................................................................................................... 20  
2 State of the art ......................................................................................................................... 21  
   2.1 Mink production in Europe ............................................................................................... 21  
      2.1.1 Mink production .......................................................................................................... 21  
      2.1.2 Welfare of farmed mink ............................................................................................ 25  
   2.2 On-farm animal welfare assessment ............................................................................... 28  
      2.2.1 Assessing animal welfare at farm level ..................................................................... 28  
      2.2.2 Welfare Quality® ..................................................................................................... 32  
      2.2.3 Assessing animal welfare in all phases of production ................................................. 35  
      2.2.4 Changes in welfare with date of assessment .............................................................. 36  
      2.2.5 Changes in welfare within the day of the assessment ............................................... 37  
      2.2.6 Sampling of animals for on-farm animal welfare assessments .................................. 38  
      2.2.7 Validity, reliability and feasibility ............................................................................ 40  
3 Methodology .......................................................................................................................... 44  
   3.1 The WelFur-Mink welfare assessment protocol .............................................................. 44  
      3.1.1 The overall structure of the assessment protocol ....................................................... 44  
      3.1.2 WelFur-Mink measurements ..................................................................................... 46  
      3.1.3 From measurement results to measurement scores ................................................... 47  
      3.1.4 From measurement scores to principle scores ......................................................... 48  
      3.1.5 Overall categorisation of farms ................................................................................. 48  
   3.2 The methods used in each part of the PhD study .......................................................... 49  
      3.2.1 Part 1 - Sampling method .......................................................................................... 50  
      3.2.2 Part 2 - Changes in welfare with date of assessment ................................................ 53
3.2.3 Part 3 - Stereotypic behaviour in the winter assessment period ..........56

4 Results .......................................................................................................................... 60
  4.1 Paper 1 .................................................................................................................... 61
  4.2 Paper 2 .................................................................................................................... 85
  4.3 Paper 3 ................................................................................................................... 126

5 Discussion .................................................................................................................. 140
  5.1 Discussion of hypotheses ....................................................................................... 140
    5.1.1 Hypothesis 1 - Sampling method .................................................................... 140
    5.1.2 Hypothesis 2 - Changes in welfare with date of assessment .................... 142
    5.1.3 Hypothesis 3 - Stereotypic behaviour in the winter assessment period .... 143
  5.2 General discussion ............................................................................................... 144
    5.2.1 Animal welfare improvements ....................................................................... 145
    5.2.2 Consequences of the aggregations ................................................................ 147
    5.2.3 Conversion factors in animal welfare assessments ....................................... 150
    5.2.4 Balancing validity, reliability and feasibility .................................................. 152
    5.2.5 Generalisation of results ............................................................................... 153
    5.2.6 Methodological considerations ...................................................................... 155

6 Conclusions ................................................................................................................ 158

7 Perspectives .............................................................................................................. 160

8 Additional publications from the PhD study ............................................................ 164

9 Activities resulting in ECTS course credits ............................................................... 167

10 Other dissemination activities ................................................................................ 168

11 References .............................................................................................................. 170
Summary

WelFur-Mink is an on-farm animal welfare assessment protocol for farmed mink. This protocol, and a similar for foxes (WelFur-Fox), is included in an animal welfare assurance scheme for mink and fox pelts produced in Europe. In this scheme, only pelts from farms with an assessed welfare above a defined minimum, which also live up to the industry standards regarding cage size and number of animals per cage and the national legislation, will be sold at the European and North American auction houses.

The WelFur protocols are based on the concept of the animal welfare assessment system Welfare Quality®. In WelFur, the welfare is assessed at farm level based on a range of measurements taken on the farm. In order to take the seasonality of mink and fox reproduction into account, the overall assessment of animal welfare at farm level is based on three assessments, one in each of the three assessment periods. The assessment periods are 1) Winter, 2) Reproduction and nursing and 3) Growth. The measurements are both animal- and resource-based, and most measurements are taken on a sample of the cages with animals on the farm. Each measurement is relevant to some aspect of the welfare according to the four welfare principles with the 12 underlying welfare criteria that were defined in Welfare Quality®. The results of the measurements taken in each assessment period are transformed into scores. These scores are aggregated across the three assessment periods into 12 criteria scores, which are further aggregated into four principle scores. The overall assessment of welfare at farm level is then categorised based on the four principle scores.

The overall aim of this PhD study was to address central challenges regarding the validity, reliability and feasibility of the WelFur-Mink assessment protocol, which became apparent when the protocol was developed and tested. The challenges were addressed in three different parts of this study. The first part had to do with the sampling of cages with mink for the on-farm assessment, and the hypothesis was: A feasible and unbiased method for taking a representative sample of mink for welfare assessments according to WelFur-Mink can be developed. We expect that samples selected with this method will reflect, but not necessarily be an exact representation of, the farm on which they are taken. The second part had to do with the possible change in the welfare assessment with date of assessment in the winter and growth assessment periods, and the hypothesis was: Date of assessment in the winter and growth assessment periods affect the result of some...
measurements in WelFur-Mink. These changes do not necessarily affect the welfare scores or the overall categorisation. The third part had to do with the timing within the day of the observation of stereotypic behaviour in the winter assessment period, and the hypothesis was: The prevalence of stereotypic behaviour in the winter assessment period in WelFur-Mink depends on the time of observation, but it is possible to derive a conversion factor to adjust for the differences.

In the first part, a new sampling method was developed and tested by simulated sampling on a model farm which mimicked a quite large and complicated commercial Danish mink farm. This new method is considered semi-random as it is based on a systematic distribution between sheds followed by a random selection of cages within the respective sheds. This method was developed to replace the original sampling method in WelFur-Mink which was not feasible when used in practice. In addition, there was a risk of assessor bias. The new sampling method was found to be sufficiently feasible. Moreover, assessor bias was avoided as the selection of cages is independent of the individual assessor. The results of the simulated sampling, where the representativeness of the simulated samples was evaluated in regard to factors related to the minks’ characteristics and housing environment, showed that the method has no systematic skewness. The results also showed that individual samples selected with this new method cannot be expected to be representative according to all factors. However, the selected samples can be expected to be representative according to some or most factors. Therefore, this method is suggested to balance feasible, unbiased and representative sampling in WelFur-Mink.

In the second part, the possible changes in the WelFur-Mink assessment with date of assessment during the winter and growth assessment periods were investigated based on repeated welfare assessments on eight Danish mink farms. There was a change with date of assessment in the results of some measurements and the related welfare scores, which may in some cases affect the overall categorisation of welfare at farm level. Some changes were positive in regard to the animal welfare while others were negative. Conversion factors may be derived to adjust for these changes at measurement level to ensure the most correct assessment independent of date of assessment. However, the possible variation at European level needs to be taken into account to ensure that the conversion factors can be applied all across Europe.
In the third part, the difference in the prevalence of stereotypic behaviour observed before feeding and before sunset was investigated based on behavioural observations on five Danish mink farms, and the results were compared to the results from all European WelFur-Mink assessments in the winter assessment period 2018. The results showed a large variation in the prevalence of stereotypic behaviour during winter on mink farms in Europe. The prevalence of stereotypic behaviour was found to vary with the timing of the observation and was higher when the observations were carried out before feeding than before sunset. The results also indicated that the relationship between the prevalence of stereotypic behaviour observed before feeding and before sunset varies between farms, thereby making it challenging to derive a reliable conversion factor. Restricting observations of stereotypic behaviour in WelFur-Mink to be carried out before feeding only (or when feeding is postponed up to 1.5 hours) may be the most feasible and reliable solution until more knowledge about the variation within the day in the prevalence of stereotypic behaviour in the winter assessment period is available.

This PhD study provides examples of how validity, reliability and feasibility may have to be re-evaluated and balanced in order to facilitate a large-scale implementation of an animal welfare assessment system. The challenges addressed in this study can be considered examples of more general challenges of implementing on-farm animal welfare assessment systems on a large scale as there may be similar challenges when assessing animal welfare in other animal production systems.
**Sammendrag (Danish summary)**

WelFur-Mink er en protokol til vurdering af dyrevelfærd på minkfarme. Denne protokol og en lignende for ræv (WelFur-Fox) bruges som en del af en dyrevelfærdsstandard for mink- og ræveskind produceret i Europa. I denne standard vil kun skind fra farme med en vurderet velfærd over et defineret minimum, som også lever op til industristandarden vedrørende burstørrelse og antal dyr per bur samt den nationale lovgivning, blive solgt i de europæiske og nordamerikanske auktionshuse.


Resultaterne fra registreringerne i de tre perioder omregnes til point, som sammenvejes på tværs af de tre perioder til én pointværdi for hvert af de 12 velfærdskriterier, som herefter sammenvejes til én pointværdi for hver af de fire velfærdsprincipper. Den overordnede vurdering af velfærden på besætningsniveau kategoriseres på baggrund af pointværdierne for de fire velfærdsprincipper.


I den anden del blev den mulige ændring i velfærdsvurderingen med observationsdato i vinter- og vækstperioderne undersøgt på baggrund af gentagne velfærdsvurderinger på otte danske minkfärme. Resultaterne viste, at der var en ændring i resultatet af nogle registreringer og de tilhørende velfærdspoint, som i nogle tilfælde kan påvirke den overordnede kategorisering af velfærden på besætningsniveau. Nogle ændringer var positive med hensyn til dyrevelfærd, mens andre var negative. Det kan være muligt at udlede korrektionsfaktorer til at korrigere for denne ændring for at sikre, at vurderingen er så korrekt så mulig uafhængigt af observationstidspunkt. Det er dog vigtigt at tage højde for den mulige variation i Europa for at sikre, at korrektionsfaktorerne kan bruges på tværs af lande i Europa.
I den tredje del blev forskellen i forekomsten af stereotyp adfærd ved observation før fodring og før solnedgang undersøgt på baggrund af adfærdsobservationer på fem danske minkfarmer, og resultaterne blev sammenlignet med resultaterne fra alle europæiske WelFur-Mink-vurderinger i vinterperioden 2018. Resultaterne viste en stor variation i forekomsten af stereotyp adfærd i vinterperioden på minkfarmer i Europa. Forekomsten af stereotyp adfærd varierede med timingen af observationen og var højere ved observation før fodring end før solnedgang. Resultaterne indikerede også, at forholdet mellem forekomsten af stereotyp adfærd observeret før fodring og før solnedgang kan variere mellem farme, hvilket gør det udfordrende at udlede en pålidelig korrektionsfaktor. Den mest brugbare og pålidelige løsning kan være at begrænse observationerne af stereotyp adfærd i WelFur-Mink til kun at blive fortaget før fodring (eller når fodringen udsættes i op til 1,5 time), indtil der er mere viden om variation i forekomsten af stereotyp adfærd i løbet af dagen i vinterperioden.

Dette ph.d.-studie er et eksempel på hvordan det kan være nødvendigt at revurdere og balancere validitet, troværdighed og praktisk anvendelighed for at muliggøre implementeringen af et velfærdsvurderingssystem i stor skala. Udfordringerne, som blev adresseret i dette studie, kan ses som eksempler på generelle udfordringer ved implementeringen af velfærdsvurderingssystemer i stor skala, da der kan være lignende udføringer ved vurderingen af velfærd i andre husdyrproduktionssystemer.
List of included papers

**Paper 1**
The representativeness of a semi-random sampling method for animal welfare assessments on mink farms
*Anna Feldberg Marsbøll, Britt Ingeborg Foseide Henriksen and Steen Henrik Møller*

Submitted for publication in Animal Welfare
(Note: Included in the stage it was in when the thesis was submitted in September 2018. Since then, a revised version has been accepted for publication)

**Paper 2**
Changes in the welfare of mink (*Neovison vison*) with date of assessment in the winter and growth periods have limited effects on the overall WelFur categorisation
*Anna Feldberg Marsbøll, Britt Ingeborg Foseide Henriksen, Bente Krogh Hansen and Steen Henrik Møller*

Submitted for publication in Animal Welfare
(Note: Included in the stage it was in when the thesis was submitted in September 2018. Since then, a revised version has been accepted for publication)

**Paper 3**
If stereotypic behaviour cannot be observed before feeding in WelFur-Mink, a reliable conversion factor needs to be developed
*Anna Feldberg Marsbøll, Tine Rousing and Steen Henrik Møller*

Submitted for publication in Animal Welfare
1 Introduction

1.1 General introduction

In recent years, animal welfare assurance schemes have become increasingly common (Lundmark et al. 2018, More et al. 2017). Such assurance schemes often include some kind of marketing of the animal-based products based on the originating farms’ certified degree of animal welfare (More et al. 2017, Veissier et al. 2008). In many cases the products are labelled. This may be as a part of a more general label, including aspects like product quality, food safety or production philosophy, or as a specific animal welfare label (Veissier et al. 2008). Such labelling is a way of informing the consumer about farm animal welfare which helps some consumers decide which product to choose (Verbeke 2009). Animal welfare assurance schemes may be driven by governments or private companies and organisations and may include animal welfare certification of farms producing meat, eggs and milk and marketing of both fresh and processed products originating from the farms (Lundmark et al. 2014, Veissier et al. 2008). The certification can be described as a formal procedure whereby the compliance with established requirements is evaluated. This means that the farm must follow a defined checklist or that the assessed welfare must be above a given level in order to be certified (Main et al. 2014). The reason for the development of such animal welfare assurance schemes may differ depending on who is the initiator and may include an aim to accommodate consumer demands, attract particular consumer groups, create or maintain a certain image or facilitate animal welfare improvements (Lundmark et al. 2014, Verbeke 2009). In most cases, it is voluntary to join animal welfare assurance schemes, however, in some cases it may be necessary in order to market or even sell the products (Heath et al. 2014b). Farmers may join animal welfare assurance schemes for many different reasons. Bock and van Huik (2007) found that European pig farmers in general are motivated by the potential improved market access and increased prices, but also ethical concerns as well as the possibility to improve animal welfare may be important.

In Denmark alone, there are several examples of such animal welfare assurance schemes. Recently, the Danish government launched a voluntary assurance scheme where pork and processed pork products are labelled on a scale with three levels according to the originating farm’s degree of animal welfare (Danish Veterinary and Food Administration 2017). Another example is driven by the Danish retail cooperative COOP. In this voluntary assurance scheme, meat, eggs and dairy products are labelled on a scale with four levels.
Introduction

according to the originating farm’s degree of animal welfare (COOP n.d.). The Danish animal welfare organisation Animal Protection Denmark also run a voluntary assurance scheme. In this scheme, meat, dairy products, eggs and different processed products originating from farms with a defined level of animal welfare are labelled ‘Approved by Animal Protection Denmark’ (Dyrenes beskyttelse n.d.). Other European examples include the British RSPCA assurance scheme where products originating from farms with a defined level of animal welfare are labelled ‘RSPCA Assured’. This voluntary assurance scheme is driven by the animal welfare organisation Royal Society for the Prevention of Cruelty to Animals (RSPCA) (Pickett et al. 2014, RSPCA n.d.-b).

In many European countries, the debate on whether fur animal production should be allowed is opened with regular intervals. The included arguments is related to both ethics and animal welfare, and in some European countries, the production of fur animals is banned or in the process of being phased out. Thus, documenting and improving animal welfare on fur farms is important for the fur industry. Assurance of fur animal welfare has also been increasingly requested by the fashion industry which is an important purchaser of the pelts. Recently, animal welfare certification of mink and fox farms in Europe have been initiated. This certification is driven by Fur Europe as a part of an animal welfare assurance scheme, where the plan is that only mink and fox pelts coming from welfare certified farms will be sold in the European and North American auction houses from 2020 (Fur Europe 2017). This certification can also be used to label the final products made by the pelts and, hence, be visible for the consumers (M.L. Nielsen, Fur Europe, personal communication, 16 July 2018). The WelFur on-farm animal welfare assessment protocols are the basis of this certification. In order to be certified, the animal welfare on the farm must be above a defined level according to the WelFur assessment of the farm. The farm must also live up to the industry standard regarding cage size and number of animals per cage as well as the requirements given in the national legislations (M.L. Nielsen, Fur Europe, personal communication, 16 July 2018).

In WelFur, the welfare of farmed mink and foxes are assessed according to species specific welfare assessment protocols where the welfare is categorised at farm level on a scale with four levels (Best, Good, Acceptable or Unacceptable current practice) (Mononen et al. 2012). Mink and foxes are seasonal breeders and the female gives birth to one litter per year. All parts of the fur production cycle, from mating to pelting, typically take place on the same farm. In order to take the seasonality of the production into account and cover
Introduction

the entire production cycle, the overall WelFur assessment of the welfare at farm level includes three assessments of the same farm, one in each of the three assessment periods. The assessment periods are 1) Winter, 2) Reproduction and nursing and 3) Growth. Each assessment period corresponds to one of the three seasons of the annual cycle of mink and fox production. Each assessment is based on a visual inspection of a sample of the animals and their housing environment as well as farmer interviews and mortality records for the assessment period. The on-farm assessment takes place during a shorter period towards the end of each assessment period where the welfare problems often are most evident (Mononen et al. 2012).

The WelFur assessment protocols were developed as a part of the WelFur project. This project was initiated by the European Fur Breeders’ Association (now a part of Fur Europe) in 2009 and involved researchers from seven universities and research institutions (Mononen et al. 2012). The first WelFur assessment protocols were published in 2013 (mink) and 2014 (foxes) and updated protocols for both mink and foxes were published in 2015 (EFBA 2013, Møller et al. 2015, WelFur 2014, 2015). The development process and resulting protocols were based on the concept of the EU project Welfare Quality® where animal welfare assessment protocols for pigs, cattle and poultry were developed. A group of scientists which included the Welfare Quality® project manager supervised the development of the WelFur assessment protocols in order to ensure that WelFur lived up to the same standards as Welfare Quality® (Mononen et al. 2012).

The large scale implementation of the fur animal welfare certification, including the WelFur on-farm assessments, was initiated in January 2017 (Fur Europe 2017). An independent third party, a commercial certification company, was engaged to assess the welfare on the farms signing up for the assessment. Therefore, the WelFur system was expanded to also include an educational programme for WelFur assessors which was developed by the WelFur scientists. The first assessors were trained in December 2016 as described for WelFur-Mink by Møller et al. (2017b). It is voluntary for the fur farmers to join the certification scheme. At present, 94 percent of the approximately 4,000 mink and fox farms in Europe have signed up, and about 50 percent of the on-farm assessments have been carried out (M.L. Nielsen, Fur Europe, personal communication, 16 July 2018). As a part of the WelFur system, the results are available for the farmers directly after each assessment; hence, the farmers can adjust their daily management according to the results (Møller et al. 2017a). The WelFur system also includes an educational programme for
advisors which was also developed by the WelFur scientists. Through this programme, advisors have been trained to advice the farmers on what to change in order to improve animal welfare based on the results of the WelFur assessments as described for WelFur-Mink by Møller et al. (2017b).

It is important that any animal welfare assessment system is both valid and reliable. In this regard, validity refers to whether the assessment system actually measures the animal welfare on the farm in regard to the chosen aspects of animal welfare, while reliability has to do with robustness and repeatability of the assessment. The system must also be feasible which means that all procedures must be useful in practice and the costs reasonable (Knierim and Winckler 2009, Veissier et al. 2013). This often requires some compromises as it is not always possible to achieve the highest validity, reliability and feasibility for all aspects of an animal welfare assessment system. As the WelFur assessments are used for animal welfare assurance of pelts, it is of course important that all stakeholders, including the fur farmers, the customers at the auctions houses, the furriers, the fashion houses, the buyers of the pelt garments and products as well as the general public, politicians and animal welfare organisations can trust the assessment. For the individual farmer, a wrong assessment may have a direct financial impact, which further underlines the importance of this.

During the development of the WelFur assessment protocols, the validity, reliability and feasibility of all possible measurements were evaluated, and only measurements that were found sufficiently valid, reliable and feasible were included in the protocols (Mononen et al. 2012). When developing an animal welfare assessment system, some challenges regarding the validity, reliability and feasibility of the system can be foreseen and thus taken into account. Other challenges do not become apparent until the system starts being used on a larger scale. This was the case for the WelFur assessment protocols which were tested in practice during the development and refinement of the protocols. The WelFur-Fox and WelFur-Mink protocols follow the same overall structure, but as there are several differences in the biology of mink and foxes, the content of the protocols in terms of the included measurements and their assessment differ. In the following, only WelFur-Mink is considered, but most is also applicable for WelFur-Fox.

Testing the WelFur-Mink assessment protocol during the development and refinement of the protocol on 19 Danish mink farms in 2011-2013 (Henriksen 2015), 39 Finnish mink
farms in 2012-2014 (Mononen et al. 2014), 18 Danish mink farms in 2014-2015 (see Section 3.2.2) and 27 mink farms in 10 European countries in 2015-2016 (Fur Europe n.d.) revealed challenges regarding the feasibility that needed to be taken into account. Some necessary information could not be retrieved from the farmers, and the daily management routines made it challenging or even impossible to conduct some of the registrations. Representative sampling of animals is necessary in order to ensure a valid assessment. However, this was sometimes challenging to achieve in practice. Furthermore, as the WelFur assessment protocols today are used across farms, in different countries and by different assessors, there are several potential challenges regarding the reliability of the assessment. Due to the required feasibility of the system, each on-farm assessment period is minimum seven weeks long. Even if the reliability was taken into account at measurement level when developing the assessment protocols, it was still expected that the reliability of some measurements and perhaps the overall WelFur-Mink assessment could be challenged by date of assessment in each on-farm assessment period. The robustness in regard to date of assessment was therefore investigated for the on-farm assessment period of the reproduction and nursing period during the initial test of WelFur-Mink, and it became apparent that date of assessment in this period could affect the result of the individual measurements as well as the overall assessment (Henriksen and Møller 2015).

1.2 Aim and hypotheses

The overall aim of this PhD study was to address central challenges regarding the validity, reliability and feasibility of the WelFur-Mink assessment protocol, which became apparent when the protocol was developed and tested. Three different challenges were addressed, one challenge in each part of the study. Similar challenges may exist when assessing the welfare in other animal productions, and thus, they can be considered examples of more general issues regarding the validity, reliability and feasibility of on-farm animal welfare assessments.

The first part had to do with the sampling of cages with mink for the on-farm assessment of welfare in each assessment period. The original stratified sampling method was not always feasible when used in practice. It was often not possible to retrieve the necessary information to select a stratified sample. In addition, it was difficult to select the cages in practice and there was a risk of assessor bias. Hence, it was important to develop a new method to ensure both the reliability and feasibility of the assessment. As when using the
original sampling method, the selected samples must be representative for the farm on which they are selected in order to ensure the validity of the assessment.

The second part had to do with the possible change in the assessed welfare with date of assessment during the on-farm assessment periods of the winter and growth periods. As previous research had shown an effect of date of assessment on the assessed welfare during the on-farm assessment period of the reproduction and nursing period, it was important to investigate if this was also the case for the on-farm assessment periods of the winter and growth periods. Depending on the magnitude, changes in the assessed welfare with date of assessment must be adjusted for in order to ensure the reliability of the assessment.

The third part had to do with the timing within the day of the assessment of the measurement ‘Stereotypic behaviour’ in the on-farm assessment period during winter. In order to standardise the assessment between farms, observations of stereotypic behaviour should be carried out within 1.5 hours before the expected feeding time. In practice, this was not always feasible, and it was decided to allow behavioural observations of stereotypic behaviour to be carried out within 1.5 hours before sunset if it was not possible to do the observations before feeding. However, if the prevalence of stereotypic behaviour in the winter is dependent on the timing of the observation, this must be adjusted for in order to ensure the reliability of the measurement.

The hypotheses of the PhD thesis were:

I. A feasible and unbiased method for taking a representative sample of mink for welfare assessments according to WelFur-Mink can be developed. We expect that samples selected with this method will reflect, but not necessarily be an exact representation of, the farm on which they are taken.

II. Date of assessment in the winter and growth assessment periods affect the result of some measurements in WelFur-Mink. These changes do not necessarily affect the welfare scores or the overall categorisation.

III. The prevalence of stereotypic behaviour in the winter assessment period in WelFur-Mink depends on the time of observation, but it is possible to derive a conversion factor to adjust for the differences.
Three studies were carried out in order to address the three challenges and test the hypotheses. Firstly, a more feasible sampling method was developed and tested by simulated sampling on a model farm. Secondly, the changes in the WelFur-Mink assessment with date of assessment during the on-farm assessment periods of the growth and winter periods were investigated based on repeated assessments on eight mink farms. Thirdly, the difference in the prevalence of stereotypic behaviour before feeding and before sunset was investigated by behavioural observations on five mink farms in the on-farm assessment period of the winter period, and the results were compared to results from all European WelFur-Mink assessments during the winter period 2018.

1.3 Outline of the thesis
The content of this thesis is centred around the three studies that were carried out in order to test the three hypotheses presented in this chapter (Chapter 1). A state-of-the-art gives an overview of mink production in Europe and a thorough introduction to the concept of on-farm animal welfare assessments and the terms validity, feasibility and reliability (Chapter 2). The WelFur-Mink assessment protocol is presented in detail in the following chapter (Chapter 3). In this chapter, the background for the methods used in each of the three studies are also presented. The results are presented as three papers, each related to one of the hypotheses (Chapter 4). The results are discussed in relation to the hypotheses and the overall aim of the thesis (Chapter 5). The methods are also discussed. Finally, the overall conclusion of the thesis is presented in regard to the hypotheses (Chapter 6) followed by a discussion of the perspectives (Chapter 7).
2 State of the art

This chapter is divided into two parts. In the first part, mink production in Europe with emphasis on the seasonality of the production system, the housing conditions and the main welfare challenges are described. In the second part, the concept of on-farm animal welfare assessments is introduced, including the terms validity, reliability and feasibility.

2.1 Mink production in Europe

2.1.1 Mink production
Mink are farmed with the purpose of producing pelts to be used for making garments and other products. In 2017, there were approximately 3,000 mink farms in Europe which produced about 39,000,000 mink pelts. This accounted for approximately 60 percent of the world's production of mink pelts (V. Gospodinova, Fur Europe, personal communication, 16 July 2018). The European farms differ significantly in size, and based on information from WelFur-Mink assessments conducted in the winter assessment period 2018, where the welfare on more than one third of the European mink farms was assessed, it seems that farm size in Europe ranges from less than 100 to more than 40,000 breeding females per farm with an average of 4,500 breeding females per farm (V. Gospodinova, Fur Europe, personal communication, 21 August 2018). Mink are farmed in most of Europe and the main producing countries are Denmark, Poland and the Netherlands. Mink farming is forbidden in some European countries, including Austria, Bosnia-Herzegovina (phasing out until 2028), Croatia, the Czech Republic (phasing out until 2019), Luxemburg, the Netherlands (phasing out until 2024), the Republic of Macedonia, Serbia (phasing out until 2019), Slovenia and the United Kingdom (V. Gospodinova, Fur Europe, personal communication, 16 July 2018). As for other animal productions, the legal framework for mink farming in Europe in regard to transport, housing, killing of animals, etc. is stipulated by the European Union and the Council of Europe. These requirements must be incorporated into national legislations which can also include additional requirements, and each government is responsible for the farms' compliance with the legislation (Veissier et al. 2008).

The farmed mink is the American mink (Neovison vision) which was first farmed in North America (USA (Smith 1981) and Canada (Bowness 1980, cited in Hamre et al. (2014))) in the 1860s and exported to Europe (Norway) in the 1920s (Storsul 2001, cited in Hamre et
al. (2014). The wild American mink is brown with white markings under the chin, down the neck and on the belly (Nes et al. 1988). Mink are strictly carnivores and in nature, they eat a large variety of prey such as small mammals, fish, crustaceans, eggs and birds. Their natural habitat is often nearby water, and they may hunt both on ground and in water, thus they are considered semiaquatic (Dunstone 1993). Mink are seasonal short-day breeders, and the females gives birth to one litter per year. The photoperiod is the main regulator of the reproductive cycle (Sundqvist et al. 1989) as well as the furring cycle (Valtonen et al. 1995). The wild adult mink are solitary and territorial. The males may have overlapping territories with several females, while there is no overlap between the territories of mink of the same sex. The territories are often alongside a river, lake or coastline, and the distance they roam over has been found to range from 1.5-5 km for male mink and 1-3 km for female mink. The adult males and females come together for mating in the early spring and the kits are born in the late spring (Dunstone 1993). The mink kits are altricial at birth, without the ability to see, hear or thermoregulate (Brandt et al. 2013, Foss and Flottorp 1974, Harjunpaa and Rouvinen-Watt 2004) and they depend on maternal care for survival and growth the first four to six weeks of their life (Hovland et al. 2017). In nature, the mink kits disperse as juveniles during autumn. This have been reported to start from the beginning of July which is at about ten weeks of age (Gerell 1970).

As a consequence of the seasonality of mink reproduction, the production system is strictly seasonal, which means that it is synchronised between and within farms (Møller et al. 2003). It can be divided into three main production seasons defined by some of the main events taking place during the production year, as illustrated in Figure 1. These are 1) Winter (breeding animals from pelting to mating), 2) Reproduction and nursing (adults and offspring from mating to weaning, including the separation of the kits) and 3) Growth (adults and juveniles from separation to pelting). The timing and duration of weaning, separation and pelting may vary between farms due to differences in management, while the timing and duration of mating is less variable due to the biological control of mink reproduction. Traditionally, all parts of the production cycle take place on each farm. Due to the seasonal production system, the number of mink on the farm differs drastically between the production seasons. Therefore, some farms may have separate farm units where they keep a part of the mink during the growth season, where the number of needed cages is the highest.
The solitary and social periods of the mink’s nature are to some degree reflected in the housing of the mink. In the winter season, there are only adult mink selected for breeding housed individually on the farms. During mating, the male and female mink are given one or more opportunities to mate naturally – most often by moving the female to the male’s cage. Typically, the farmers keep one male for every five females. After mating, all or some of the males and unmated females are often pelted, thus reducing the number of mink on the farms. When the females give birth, the number of mink on the farms is increased by a factor around six, depending on the breeding result on the farms. The litter is kept with the mother for six to ten weeks before they are physically separated, which is referred to as weaning. The litter is most often split up into smaller groups or pairs, which may be done at the same time as weaning or a few weeks later. This is referred to as separation. During the growth season and until pelting, the juvenile kits are kept in pairs or groups. The adult females are typically housed individually or with one or more male juveniles. The mink selected for pelting are killed on the farm, usually in a box containing carbon monoxide or dioxide (Christiansen et al. 2012).

Across Europe, mink are housed under quite similar conditions. Each farm typically consists of several sheds. As the annual variations in the length of day and night is important for the regulation of mink reproduction and furring cycles, mink are usually kept under natural lighting conditions. This is typically achieved by keeping the sheds open on the sides or by having transparent plates in the roof as shown in Figure 2. In each shed,

![Figure 1. The three seasons of the annual cycle of mink production (Revised from Henriksen (2015) and Møller et al. (2015))](image-url)
the mink are kept in wire mesh cages elevated off the ground (Jørgensen 1985). The Council of Europe (1999) recommends cages with an area of 2,550 cm² and a height of 45 cm for single adult mink, for females with kits and for pair-housed juvenile mink. If more than two mink are housed in the same cage, the space should be increased with 850 cm² for each additional mink. The cages are adjacent to each other, thereby forming long rows of cages as shown in Figure 2 (Jørgensen 1985). Typically, the mink have access to a nest box and some kind of bedding material such as straw or wood shavings, which is also recommended by the Council of Europe (1999). Figure 2 show an example of a nest box can be entered from the cage. Environmental enrichments such as objects that can be moved around or manipulated and platforms for observing surroundings may also be available. The Council of Europe (1999) also recommends that mink have access to ‘suitable occupational material’, which have been specified in the national legislation in several countries. For example, Danish legislation requires that the mink have access to straw, a platform and a plastic tube (Fødevarestyrelsen 2015). Natural occurring mutations in genes controlling fur coat colour have led to several colour types, of which some have been included as a selection criteria in the farmed mink in order to produce pelts with distinct colours and markings (Lohi et al. 2015, Nes et al. 1988). Most farms are specialised in producing mink in one or a few colour types, but on some farms, several different colour types may be present.

Figure 2. Housing of mink in rows of cages in a shed where the transparent roofing plates and windows lets the daylight in (left). From each cage, the mink have access to a nest box (right).
2.1.2 Welfare of farmed mink

One general prerequisite for farming is that the animals are kept in more restricted areas than their wild relatives. Since the early years of mink farming, this has been achieved by keeping mink in cages. Many people consider the housing environment one of the main welfare challenges in farmed mink as the mink’s natural territory is much larger than the space available in the cages, and as the wild mink travel long distances in their search for food. Moreover, the relatively barren environment and the general lack of access to swimming water are often considered problematic as wild mink’s territory is naturally more variable and often includes a water resource (Pickett and Harris 2015). Mink kept in cages of the recommended size with access to a nest box display the most natural behavioural patterns such as walking, resting, grooming, playing, mating and nursing (e.g. Brink and Jeppesen 2005, Brink et al. 2004, Hansen et al. 2007). However, abnormal behaviour associated with a reduced welfare such as stereotypic behaviour and fur chewing may be displayed by some mink (e.g. Malmkvist et al. 2013, Svendsen et al. 2013). In addition, behavioural and physiological fear responses associated with a reduced welfare may be elicited by procedures including capturing and handling of the mink (Korhonen et al. 2000, Malmkvist et al. 2011). This could for example be in relation to blood sampling, vaccination or fur grading. Increasing cage size has been found to have only a limited or no effect on the welfare. Instead, the complexity and enrichment of the cages seem more important (Hansen 1988, Hansen et al. 1994, Hansen et al. 2007). Cage enrichments like ropes, tubes and chains have been found to improve the welfare of mink by reducing the stress hormone level, improving reproduction and reducing the occurrence of abnormal behaviour (Díez-León and Mason 2016, Hansen et al. 2007, Malmkvist et al. 2013, Meagher et al. 2014). Access to swimming water may also improve the welfare, but a similar effect may be achieved by other enrichments (Hansen and Jensen 2006a, Hansen and Jensen 2006b, Vinke et al. 2008). It has also been shown that access to a nest box and bedding material which may be used for several purposes, including resting, nesting, thermoregulation and nursing, are important for the welfare of the mink. For example, mink are willing to work for access to a nest box, and the provision of bedding material may increase kit survival and reduce the maternal stress hormone level (Hansen 1988, Hansen et al. 1994, Malmkvist and Palme 2008, Schou et al. 2018). For most mink, catching and handling are rare procedures which makes habituation to these fear-eliciting situations unlikely. However, it has been shown that it is possible to reduce fearfulness in mink by selecting for confident mink, which has been found to reduce the fear response in
potential fear-eliciting situations both with and without human handling involved (Korhonen et al. 2000, Malmkvist and Hansen 2001, Malmkvist and Hansen 2002).

Many welfare challenges in farmed mink differ between the production seasons and are related to the different animal groups present on the farm in the different seasons. The risk of reduced welfare may also vary within seasons due to changes in age (which is linked to body size, hormonal development etc.), management or the onset of events. One of the main welfare challenges is related to feeding management. Most mink are fed close to ad libitum during the growth season in order to increase the size of the pelt. Thus, the body condition of the mink is increased throughout the season, often to fat or obese, which may have consequences for the health of the mink. For example, the risk of fatty liver is increased (Hunter and Barker 1996). When the following year’s breeding animals have been selected in late autumn, the farmers often aim to condition the selected mink to an average or below average body condition before mating. This is achieved by feeding a more energy restricted diet during this period. The aim of this procedure is to improve reproduction, that is to ensure that the females respond to flushing (a short period of restricted feeding followed by ad libitum feeding), which has been found to increase the number of released eggs (Tauson 1993). The males have also been found to be more willing to mate and the number of barren females are reduced (Hansen et al. 2013). Mink are naturally somewhat adapted to body weight changes during winter. However, the welfare may still be compromised if this procedure is not managed carefully as there is a risk that the mink will experience hunger and become too thin. The restricted feeding may also result in more animals performing stereotypic behaviour (Damgaard et al. 2004) and other abnormal behaviour such as fur chewing (Malmkvist et al. 2013). Another welfare challenge is related to the transition of the kits from solely depending on their mother’s milk to eating solid food and drinking from the watering system. Firstly, there is a risk that the female mink gets too thin, as the requirements of milk production may exceed the possible energy consumption (Hansen 1999, Hansen and Berg 1998). Secondly, there is a risk of aggressions resulting in injuries among the kits. The kits start to eat solid food from around three weeks of age but until they learn to drink enough water from the watering system, which has been reported to start at around six weeks of age when they have access to a standard watering system, they still depend on their mother’s milk (Brink and Jeppesen 2005, Brink et al. 2004). The risk of aggressions and injuries among the kits has been found to increase from four to around seven weeks of age, which may be due to increased competition for fluids (Brink et al. 2004).
or groups in the growth season may also be a welfare challenge due to the increased risk of aggression and injuries when housing solitary and territorial animals in a limited area. This risk increases after equinox in late September when the juvenile mink starts to become territorial. Individual housing of juveniles is in general not considered beneficial due to the need of conspecifics to develop reproductive skills (Hansen et al. 1997). Mink housed in male-female pairs seem to tolerate their overlapping territories as they do in the wild, while the risk of aggression and injuries are higher in other group compositions and with increasing numbers of juveniles (Hansen and Møller 2012).

Within the same farm, the risk of welfare problems may vary between mink or groups of mink due to differences in housing conditions, management or biology. Regarding housing conditions, there may be differences in for example exposure to direct sun, which affects the thermal comfort (Wustenberg and Wustenberg 1988). There may also be differences in the access to resources such as nest boxes, bedding material and enrichments, which, as described in the previous section, may affect the welfare of the mink (Díez-León and Mason 2016, Hansen 1988, Hansen et al. 2007, Malmkvist and Palme 2008, Malmkvist et al. 2013, Meagher et al. 2014). Group composition may also differ, which, as described in the previous section, may affect the welfare as mink housed in groups may have more bite marks than mink housed in pairs (Hansen and Møller 2012). Likewise, provision of additional water supply for the kits may differ within the same farm. This may cause the welfare to differ as providing an additional water supply near the nest box opening has been found to facilitate that the kits start drinking water earlier and reduce the number of injured kits (Brink et al. 2004, Malmkvist et al. 2016). Biological differences may include parity, for example older females may lose more weight than first year females during lactation (Hansen and Berg 1998). Colour type may also affect the risk of a reduced welfare. For example, mink with the Aluetian (aa) gene, which include the colour types sapphire and violet, have been found to be more susceptible to disease (Shewen et al. 1996). It has also been shown that black mink are more fearful than pastel mink (Meagher et al. 2011).

There are also areas where the welfare challenges of farmed mink are limited. For example, there are no physical mutilations (e.g. castrations or similar routines) as an integrated part of the production system. In addition, transport in regard to slaughter is avoided as the mink are killed on the farm (Møller 2008). There are also several areas where the mink are allowed to perform their natural species-specific behaviour. This is for example in regard
to mating where only natural mating is used (Christiansen et al. 2012). Furthermore, the female and her litter have the possibility to perform their natural behaviour regarding nursing and other social interactions such as social play behaviour. In several areas, production and animal welfare considerations go hand in hand. For example, avoiding damage to the pelt due to injuries or fur chewing is positive in regard to both production and animal welfare (Møller 2008).

2.2 On-farm animal welfare assessment

2.2.1 Assessing animal welfare at farm level

On-farm animal welfare assessment systems aim to say something about the animal welfare on farms. Different systems have been developed for different purposes, for example to check compliance with legislation, collect information for research, identification of welfare problems, decision support for farmers and certification of farms (Johnsen et al. 2001, Main et al. 2003). The systems differ in how they define animal welfare, and thus, what they value as important for the animal welfare. The definition of animal welfare as well as the purpose are reflected in the structure and content of the welfare assessment system.

Even if it is generally agreed that animal welfare is a multidimensional concept that encompasses many aspects of the animal’s quality of life, there is no definitive definition of the concept of animal welfare. Fraser and colleagues divided the understanding of ‘the quality of life of animals’ into three main approaches which today are widely accepted. The first is ‘Biological functioning’ which emphasises how the animal functions. According to this approach, the welfare of an animal should be evaluated in regard to physiological markers such as disease status and production results as well as the functioning of behavioural systems. The second is ‘Affective state’ which emphasises how the animal feels. This means that the animal’s perception of the situation in terms of its associated positive or negative feelings is central in evaluating animal welfare according to this approach. The third is ‘Natural living’ which emphasises the animal’s opportunity to live a natural life. According to this approach, the welfare of an animal should be evaluated in regard to the naturalness of the animal’s surroundings and the possibility to perform species-specific behaviour (Duncan and Fraser 1997, Fraser 2003, Fraser et al. 1997). The three approaches may overlap, for example a cow kept on pasture may be considered to fare well as it is natural for a cow to be on pasture (Natural living), but it may also be
considered to fare well as it is a positive experience to be on pasture (Affective state). However, a good welfare according to one approach does not ensure a good welfare according to another approach, and the different approaches may in some cases lead to conflicting conclusions about the welfare (Fraser et al. 1997). Thus, it is important to be aware of how the animal welfare is approached and defined as this may influence what is included in an animal welfare assessment system, and how the welfare is assessed. For example, Duncan (1996) argued that only the animals feelings should be considered when evaluating animal welfare. Contrary to this, Broom (1996, 2011) argued that it is not possible to evaluate animal welfare based on the animal’s feelings alone, and that the animal’s biological functioning should be included. Another example is ‘The Five Freedoms’ which were first published in the late 1970s by the Farm Animal Welfare Council in Great Britain (Farm Animal Welfare Council 1979). Since then, they have been updated and specified. ‘The Five Freedoms’ define the ideal state of welfare rather than a standard for acceptable welfare and are described as ‘Freedom from hunger and thirst - by ready access to fresh water and a diet to maintain full health and vigour; Freedom from discomfort - by providing an appropriate environment including shelter and a comfortable resting area; Freedom from pain, injury or disease - by prevention or rapid diagnosis and treatment; Freedom to express normal behaviour - by providing sufficient space, proper facilities and company of the animal’s own kind; Freedom from fear and distress - by ensuring conditions and treatment which avoid mental suffering’ (The Farm Animal Welfare Committee 2013). To some degree, ‘The Five Freedoms’ broad definition of what constitutes good welfare can be argued to include dimensions of all three welfare approaches. For example, ‘Biological functioning’ is included in ‘Freedom from pain, injury or disease’, ‘Natural living’ is included in ‘Freedom to express normal behaviour’ and ‘Affective state’ is included in ‘Freedom from fear and distress’. However, it may also be argued that important ‘freedoms’ are missing according to all three welfare approaches. For example it may be argued that growth and reproduction (‘Biological functioning’) or positive feelings (‘Affective state’) should be included. Besides what is considered important in relation to production animal welfare, it is also important how animals are valued in a broader sense, and what is considered our duties towards animals. Three important theoretical positions in this regard are the ‘Contractarian’, ‘Utilitarian’ and ‘Animal rights’ point of view. Some may consider that any kind of animal use is acceptable as long as it is beneficial for humans (i.e. the contractarian point of view). Others may consider the consequences for the animals, and that any potential negative effect on production animal welfare should be outweighed by the human benefits as the aim is to
maximise the overall good (i.e. the utilitarian point of view). Finally, some may consider
that animals have certain rights that cannot be discussed, meaning that animal pain and
suffering cannot be outweighed by human benefits (i.e. the animal rights point of view)
(Palmer and Sandøe 2011). These different points of view may affect what is considered an
acceptable level in regard to farm animal welfare. The boarder between acceptable and
unacceptable welfare may also be affected by the type of production (i.e. meat, milk, egg or
fur), as some may consider farming animals for the production of food necessary for
humans whereas farming for fur is unnecessary (Sandøe and Christiansen 2009).

Animal welfare cannot be measured directly, but must be assessed based on indicators of
welfare which are often referred to as welfare measurements. Due to the multidimensional
nature of animal welfare, most assessments of welfare are based on several welfare
measurements. While animal welfare is a characteristic of the individual animal, different
measurements may be taken at different levels, for example animal, pen or farm level, and
the results are used to assess the animal welfare on the farm. Information from other
sources, for example data from inspections of animals at slaughter houses, can also be
included in the assessment (Johnsen et al. 2001, Vannier et al. 2014). The measurements
can be animal-based or resource-based. Animal-based measurements include registrations
of the animal’s health and behaviour. They can be taken directly on the animals including
registration of for example the physical appearance and behaviour or indirectly based on
for example treatment recordings or mortality records (EFSA 2012, Johnsen et al. 2001,
Main et al. 2003). The measured response may result from a specific event, for example an
injury, or be the cumulative outcome over a period of time, for example body condition
(EFSA 2012). Resource-based measurements relate to the environment surrounding the
animals and how the animals are managed within the system. They include registration of
the animal’s access to resources such as drinking water and resting areas and the farmers
management, for example in regard to the chosen age of weaning and the use of hospital
pens (Johnsen et al. 2001, Main et al. 2003). The results of some measurements give a
snapshot of the situation at the day of the assessment, for example the prevalence of
animals with a problem, while other measurements give the accumulated number of
incidences over a period, for example the number of dead animals in the past year. Animal-
based measurements are in general believed to provide more direct information about of
the degree of welfare and are considered independent of the production system, whereas
the resource-based measurements provide information about the risk of reduced welfare.
Moreover, the resource-based measurements are often easier to collect than the animal-
based measurements and are typically more robust (Johnsen et al. 2001, Knierim and Winckler 2009, Main et al. 2014, Vannier et al. 2014, Waiblinger et al. 2001). The first animal welfare assessment systems were mainly based on resource-based measurements, for example the Austrian Animal Needs Index developed from 1985 (Bartussek 1999, 2001) and the British RSPCA welfare standards first launched in 1994 (Pickett et al. 2014). The more recent systems put more emphasis on animal-based measurements. The RSPCA welfare standards have followed the development, and today include more animal-based measurements (Pickett et al. 2014, RSPCA n.d.-a). Welfare Quality® and related protocols (e.g. AWIN and WelFur) are examples of on-farm animal welfare assessment systems where animal-based measurements are prioritised as far as possible, which will be further elaborated in Section 2.2.2.

In most on-farm animal welfare assessment systems, the included measurements are selected based on the available scientific knowledge, for example by reviewing the available literature or interviewing animal welfare experts (Johnsen et al. 2001). The individual and marginal relevance of the measurements should be evaluated in order to ensure that the measurements together describe the chosen dimensions of animal welfare (Rousing et al. 2001). In most animal welfare assessment systems, the results of all the measurements are integrated into one overall assessment of the animal welfare at farm level. When integrating measurement results into one overall assessment, many different rules may be applied, as described by Botreau et al. (2007a). One example is the Austrian Animal Needs Index. Here a range of measurements is taken on the farm. Each measurement is given a score, and the animal welfare at farm level is assessed based on the sum of scores given to each of the measurements. Thus, it is possible to compensate a bad score in one area with a better in another area, and the predefined minimum level of welfare at farm level can be reached in several ways (Bartussek 1999, 2001). Animal welfare experts are often used to interpret the results of the measurements taken on the farm into the overall assessment of the animal welfare at farm level and to define the required minimum level of welfare, but different stakeholders may also be included. This includes indirectly answering to moral questions of what is considered good or bad animal welfare, and what is an acceptable level of welfare (Jensen and Sandøe 2013, Veissier et al. 2011). In the Welfare Quality® system, animal welfare experts as well as different stakeholder groups were involved in this process which will be elaborated in the following section. The overall assessment of the welfare at farm level may be used for labelling of products as described for different animal welfare assurance schemes in Section 1.1. However, for the farmer to be able to use results of a
welfare assessment for decision support, more detailed information about the assessment results is needed (Sørensen et al. 2013).

2.2.2 Welfare Quality®

Welfare Quality®, an EU funded research project, aimed to provide a standardised system for the assessment of animal welfare on the farm or at the slaughterhouse, where the assessment results can be used as decision support as well as provide information about the animal welfare for consumers and other stakeholders (Blokhuis et al. 2013b). It involved the collaboration between researchers from 43 institutes and universities mainly in Europe, as well as inclusion of different stakeholders at several stages of the process (Blokhuis et al. 2013a, Miele et al. 2011). As part of the project, on-farm animal welfare assessment protocols for sows and piglets, growing pigs, dairy cows, fattening cattle, veal calves, broilers and laying hens as well as a welfare assessment protocols for finishing pigs at the slaughterhouse were developed (Welfare Quality® 2009a, b, c).

In the Welfare Quality® system, the welfare is assessed as an index based on a range of measurements which are mainly taken during a one-day visit on the farm or at the slaughterhouse. Each measurement covers some aspects of the welfare according to the four principles for animal welfare specified in 12 animal welfare criteria which were defined in the project. The principles are: 1) ‘Good feeding’ specified in the criteria ‘Absence of prolonged hunger’ and ‘Absence of prolonged thirst’, 2) ‘Good housing’ specified in the criteria ‘Comfort around resting’, ‘Thermal comfort’ and ‘Ease of movement’, 3) ‘Good health’ specified in the criteria ‘Absence of injuries’, ‘Absence of disease’ and ‘Absence of pain induced by management procedures’, and 4) ‘Appropriate behaviour’ specified in the criteria ‘Expression of social behaviours’, ‘Expression of other behaviours’, ‘Good human-animal relationship’ and ‘Positive emotional state’ (Botreau et al. 2007b). The welfare principles and criteria were considered as a list of requirements defining good animal welfare. The starting point for the definition was that animal welfare is multidimensional, and that all possible dimensions should be included. It was also decided to emphasise the animal’s point of view (Blokhuis et al. 2013b).

For each assessment protocol, all potential measurements in relation to the defined welfare principles and criteria were evaluated in regard to their validity, reliability and feasibility (Forkman and Keeling 2009a, b, c). This relates to how well each measurement reflects animal welfare and how robust and operational they are, which will be elaborated in
Section 2.2.7. In the final selection of measurements, it was important that they were mutually exclusive, and only measurements that were found sufficiently valid, reliable and feasible were included in the assessment protocols. The emphasis on the animal’s point of view was underlined by the decision to choose direct animal-based measurements when possible. When direct animal-based measurements were not available or considered appropriate, indirect animal-based measurements or resource-based measurements were included instead (Keeling et al. 2013, Veissier et al. 2013). In the final assessment protocols, both animal-based and resource-based measurements are included. The number of measurements are not the same for each criteria and the number of measurements differ between the assessment protocols (Welfare Quality® 2009a, b, c).

In order to assess the welfare at farm level, several aggregation steps are used. Firstly, the result of each measurement, which may be expressed on different scales, is transformed into a measurement score on a scale ranging from 0 to 100. Here 0 corresponds to the worst possible situation, a score of 50 is neither good nor bad, and a score of 100 corresponds to the best possible situation (Welfare Quality® 2009a, b, c). These transformations were based on animal welfare experts’ interpretation of the animal welfare at farm level (Botreau et al. 2013). Secondly, these measurement scores are aggregated into one score for each welfare criteria and further into one score for each welfare principle (Welfare Quality® 2009a, b, c). The scores are aggregated using Choquet integrals which allow a low score in one area to be partly compensated by a higher score in other areas, depending on the weighing of the importance of each score for the animal welfare at farm level. This allowed the aggregations to take into account that each dimension of welfare may not fully compensate for one another. The weights were derived from consultations with both animal welfare expert’s as well as social scientists (Botreau et al. 2013). The four principle scores are the basis for the overall assessment of the welfare at farm level. This means that a farm is assigned to one of four welfare categories (‘Not classified’, ‘Acceptable’, ‘Enhanced’ and ‘Excellent’) depending on predefined threshold values for the principle scores (Welfare Quality® 2009a, b, c). These thresholds were defined based on consultations with a mix of animal welfare experts, social scientists and different stakeholders (Botreau et al. 2013). When assigning the categories, as indifference threshold of 5 is applied, meaning that a score of 75 is not considered to be lower than 80 (Welfare Quality® 2009a, b, c).
The involvement of stakeholders in the process led to adjustments at all levels, for example by increasing the focus on positive emotions as well as naturalness (Miele et al. 2011). It has been emphasised that the entire process was bound by ethical choices, for example the choice of thresholds between acceptable and unacceptable welfare, or whether it should be possible to compensate a poor result in one area with a better result in another area (Veissier et al. 2011). Today, Welfare Quality® is considered one of the most comprehensive animal welfare assessment systems. It is widely used for research in animal welfare and is often considered the ‘gold standard’ or reference value when evaluating the performance of other assessment systems or welfare indicators. For example, in the development of a Danish animal welfare index, Welfare Quality® was used as a reference value (Danish Veterinary and Food Administration n.d.). Welfare Quality® has also been used as a model for the development of on-farm animal welfare assessment protocols for other species. One example is the Animal Welfare Indicators Project where assessment protocols for farms with sheep, goats, horses, donkeys and turkeys were developed (AWIN 2015a, b, c, d, e). In this system, an initial screening determines if a more detailed assessment is needed. Another example is the WelFur project where assessment protocols for farmed mink and foxes (blue fox, silver fox and their crossbreeds) were developed (EFBA 2013, Mononen et al. 2012, Møller et al. 2015, WelFur 2014, 2015). The structure of the WelFur assessment protocol for mink is described in detail in Section 3.1.

There are some examples that Welfare Quality® and related protocols are used in animal welfare assurance schemes. In Finland, a dairy company uses Welfare Quality® for animal welfare certification of farms. In this system, farms with loose housed dairy cows with access to an outdoor area all year round can sell their milk in a concept labelled ‘Free cows’ milk’ if they are categorised as ‘Enhanced’ according to the Welfare Quality® assessment. The products are also labelled with the Welfare Quality® logo (Wallenius 2018). In Spain, Welfare Quality® and related protocols are used (or are in the process of being implemented) for certification of the animal welfare of dairy cows, beef cattle, sows, piglets, growing pigs, turkeys, laying hens and rabbits on the farms and in some cases also at the slaughter houses. This certification is used for labelling of products as conforming to welfare standards based on Welfare Quality® (Dalmau 2017, 2018). Finally, as also mentioned in Section 1.1, the WelFur assessment protocols are used to certify the animal welfare on mink and fox farms as a part of an animal welfare assurance scheme for pelts produced in Europe (Fur Europe 2017).
2.2.3 Assessing animal welfare in all phases of production

There are large differences in the structure of the production systems for different farm animals. Often, specific phases of farm animal production systems are linked to specific animal groups, housing conditions, management procedures and welfare risk factors. Thus, in order to assess the animal welfare in all or some phases of production on a farm, the structure of the welfare assessment systems also differs.

In the production of non-seasonal breeders such as dairy cows and pigs, breeding may take place all year round. This means that the production phases are independent of season, and on the same farm, several phases of production can take place at the same time, for example rearing and production. This also means that the welfare in several production phases can be assessed at the same time (Møller et al. 2003). If the aim is to cover all aspects of the animal welfare at farm level, the welfare of animals in all production phases on the farm must be included. In Welfare Quality®, the welfare of the animals in the different phases of production is assessed independently. For example, there is one assessment protocol for sows and piglets and one for growing pigs (Welfare Quality® 2009b). Hence, on a farm with both sows and piglets and growing pigs, both assessment protocols should be applied in order to cover the welfare of all the production phases present on the farm. However, some production phases may be physically divided between farms as often seen in pig production where some farms are specialised in producing piglets and others in growing pigs for slaughter. In this case, the welfare on each farms can be assessed separately according to the Welfare Quality® system.

When assessing the welfare in farm animal production systems with seasonal breeders such as sheep, foxes or mink, it is not possible to assess the welfare of animals in all production phases at the same time, as each productions phase is linked to a specific time of year (Møller et al. 2003). This is described for mink in Section 2.1.1 where the production system can be divided into three main seasons. In order to cover the welfare in all production phases, it may therefore be relevant to base the overall assessment on more than one assessment per year (Møller et al. 2003). For that reason, the WelFur assessment protocols for foxes and mink assess the overall welfare at farm level based on three assessments, one in each main production season (Mononen et al. 2012). In sheep, variation in the welfare between production phases has been documented (Munoz et al. 2018). However, in the AWIN assessment protocol for sheep, there are no guidelines for the timing of the assessment (AWIN 2015d). In production systems with non-seasonal
breeders kept in seasonal systems, for example dairy cows kept in grazing systems with
calving in the spring, a similar challenge is present. In this case, the different phases of the
production are also linked to a specific time of year.

2.2.4 Changes in welfare with date of assessment

Within each production phase, animals at different stages such as lactation stage or age
group may be related to different welfare risk factors. Thus, the assessment of welfare may
depend on the stage of the assessed animals.

In non-seasonal production systems where there are animals at several stages on the same
farm at the same time, this can be taken into account by including animals at different
stages in the assessment. For example, when assessing the welfare on a farm with sows and
piglets according to Welfare Quality®, sows in early, mid and late gestation, lactating sows
shortly after farrowing and around weaning as well as litters at different ages should be
included in the assessment (Welfare Quality® 2009b). The result of this approach is that
animals at different stages associated with different risks of reduced welfare are the basis
of the assessment. However, there may still be variations in the assessed welfare due to
season. This could be due to changes in housing conditions with season, for example dairy
cow’s access to pasture, which may have a positive effect on the animal welfare (Burow et
al. 2013). However, even if the housing conditions are relatively constant around the year,
changes in season in itself may affect the outcome of a welfare assessment. For example,
calves’ thermal comfort, assessed based on a range of behavioural indicators of welfare,
may be affected by season (Tripon et al. 2014). Kirchner et al. (2014) also found a low
consistency of several measurements of welfare in beef bulls assessed with intervals of one
or six months. They suggested that this may be caused by seasonal effects or short-term
fluctuations as they only included farms without major changes in for example
management routines or access to resources. Thus, depending on the aim of the welfare
assessment, it may be relevant to base the overall assessment of the welfare at farm level
on more than one assessment per year as suggested by Kirchner et al. (2014).

In seasonal production systems, where animals at different stages of production within
each production phase cannot be separated from date of assessment, date of assessment
may affect the outcome of the assessment. Depending on the number of farms to be
assessed, it is often not possible to assess all farms within a very short timeframe. Thus,
this may challenge the reliability of the welfare assessment system (see Section 2.2.7). The
possible change with date of assessment is taken into account in the evaluation of the measurement for the criteria ‘Absence of prolonger hunger’ in the winter assessment period in WelFur-Mink. Due to the energy restricted feeding of the mink during winter, as described in Section 2.1.2, it is expected that mink body condition will decrease from fat or obese (body condition score 4 or 5) at pelting time in November to average or below average (body condition score 3 or lower) before mating in March. As the aim is to evaluate ‘Absence of prolonger hunger’, which in this case is associated with losing too much weight too fast, mink in body condition 2 or less in January are associated with prolonger hunger. In February, there has been a longer period to lose weight, and only mink in body condition 1 are associated with prolonged hunger (Møller et al. 2015). As previously mentioned, the possible change in welfare with date of assessment has been investigated during the on-farm assessment period of the reproduction and nursing period in mink. In this study, Henriksen and Møller (2015) found that there were changes in the results of the WelFur-Mink assessment with date of assessment with more welfare problems by the end of the on-farm assessment period. A similar challenge is present in the production of broilers. All animals in the same flock will have the same age and, will therefore be at the same stage of growth. Thus, on farms with only one or a few flocks, it is not possible to assess the welfare of animals at different stages of growth at the same time. In this case it may therefore be relevant to have more than one assessment per production cycle in order to cover all stages of growth. Another possibility is to limit the length of the on-farm assessment period and only assess the welfare in flocks where the animals are in the end of the growth phase where the risk of welfare problems in general is higher. This is the approach used in the Welfare Quality® assessment protocol for broilers, where the on-farm assessment should take place within the last five days before slaughter (Welfare Quality® 2009c). However, by using this approach, there is a risk that animal welfare problems in the beginning of the growth phase may be overlooked, unless they are related to observed problems in the end of the growth phase.

2.2.5 Changes in welfare within the day of the assessment
When assessing the welfare at farm level based on registrations taken on the farm, the result of some measurements may depend on the time of the day due to the animals daily rhythm or events taking place during the day. This may make it necessary to time the registration of some measurements accordingly.
When including measurements that are based on observations of animal behaviour, the animals’ daily rhythm may affect the outcome of the assessment. For example, Villagrá et al. (2007) demonstrated that the activity level of growing pigs peaks in the morning and in the late evening and with very low activity at any other times. Thus, at some times during the day, it is difficult to quantify the prevalence of any behaviour requiring that the animals are active. In Welfare Quality®, it is ensured that there are active animals to observe by observing exploratory behaviour and social behaviour in growing pigs in the morning, when the animals are the most active (Welfare Quality® 2009b). For some welfare measurements, the animals’ daily rhythms may indirectly affect the outcome of the assessment. For example, the prevalence of dirtiness in growing pigs has been found to be lower when assessed in the morning. This was suggested to be related to the low activity during the night where the pigs spend most of the time laying down, and the faecal soiling dries and rubs off (Mullan et al. 2009). Events such as feeding may have a direct effect on the behaviour of the animals. For example, the prevalence of stereotypic behaviour in farmed mink has been found to peak prior to feeding, while there are only few occurrences of stereotypic behaviour after the feed have been delivered and for several hours thereafter (Malmkvist et al. 2013). Thus, the observation of stereotypic behaviour must be timed according to feeding time.

The result of some resource-based measurements may also be affected by time of the day. For example, if the availability of straw is used in the assessment of welfare, the time since the last provision may affect the assessment. Similarly, in systems with manual water provision, the time since the last provision of water may affect the direct assessment of water availability. In such cases, it may be a solution to base the assessment on an indirect assessment instead. This approach is used in the WelFur-Mink assessment protocol to assess water availability in cases where water is provided manually. Here, the farmer is interviewed about how many times per day water is provided as the mink are in less risk of being without water if water is provided several times per day (Møller et al. 2015). However, when the assessment of welfare is based on information provided by the farmer, the quality of the information must be considered to ensure that the assessment is as correct as possible.

2.2.6 Sampling of animals for on-farm animal welfare assessments

When assessing animal welfare in commercial farm settings, it is often not possible to do an individual assessment of the welfare of all the animals on the farm. Especially when
direct animal-based welfare indicators are used, the assessment is time-consuming and costly (Knierim and Winckler 2009, Sørensen et al. 2007). Thus, it is often necessary to restrict the assessment to a sample of the animals on the farm. In this case, the welfare of the animals in the sample is considered representative for the welfare of all the animals on the farm. As described for mink in Section 2.1.2, the risk of welfare problems may vary between animals or groups of animals on the same farm due to for example housing conditions, management or biological differences. Thus, in order to ensure the validity of the assessment (see Section 2.2.7), it is highly important that the animals in the sample are representative for the animals on the farm. This also requires that the sample size is large enough for the welfare of the animals in the sample to reflect the welfare of the animals on the farm with a sufficiently high accuracy. Moreover, the feasibility of the sampling method itself plays an important role in choice of sampling method as the theoretically best suited method is not always possible to apply in practical farm settings. As stated by Knierim and Winckler (2009), it may also be problematic to identify a representative sample in a large herd.

Sampling is the selection of individuals from a population in order to estimate characteristics of the whole population (Lohr 2010). On-farm animal welfare assessment systems use different sampling strategies depending on species and production system. Some kind of stratification is often included, which means that the farm is divided into homogenous sub-groups from which the animals are sampled proportionate to the size of the sub-group. The original sampling method in WelFur-Mink was based on stratification which is described in more detail in Section 3.2.1. Another example is the sampling of dairy cows for the clinical assessment in the Welfare Quality® assessment protocol. If the cows are kept in different groups, the animals in the sample should be stratified between the groups according to group size (Welfare Quality® 2009a). The sampling of laying hens for the clinical assessment in Welfare Quality® is also based on stratification. Here 100 birds are picked up at different locations in the house, which ideally should reflect the various areas in the house. If the birds are kept in cage systems, different tier levels should also be represented (Welfare Quality® 2009c). Stratification often increases the representativeness of the sample compared to random sampling, and the distribution of the animals in relation to the included sub-grouping factors will in an ideally stratified sample be similar to the distribution on the whole farm. However, it can be time-consuming as the sub-groups have to be identified and the proportions calculated (Lohr 2010). Many assessment systems also try to incorporate some kind of random sampling,
which means that all animals have an equal chance of being selected. For example, when sampling horses for welfare assessment according to the AWIN welfare assessment protocol, it is suggested to do a random sampling of the horses based on the microchip numbers (AWIN 2015c). When using random sampling, the selected samples will on average be representative for the farm they are selected on. However, there is a risk that the individual sample is not representative (Lohr 2010). For example, all the selected animals may by chance have the same age or be housed in the same area of the farm and thereby only represent one age group or area of the farm. In other cases, interval sampling is incorporated. This means that every n\textsuperscript{th} animal is selected. For example, when sampling dairy cows for the clinical assessment in the Welfare Quality\textsuperscript{®} assessment system, it is suggested to select every n\textsuperscript{th} cow in the group (Welfare Quality\textsuperscript{®} 2009a). Interval sampling reduces the risk of all animals having the same age or being housed in the same area of the farm by chance. However, there is a risk that periodic patterns may induce a bias in the sample (Lohr 2010). There are also examples of more practical solutions developed for sampling animals in practice. For example, when selecting a number of loose-housed sows in a large pen, it may be challenging to identify and find each individual sow. Welfare Quality\textsuperscript{®} developed a feasible sampling method to solve this problem: The first sow in sight is the starting sow. The next sow is ‘the sow who’s head is the fourth away (facing) from the ‘starting sow”, and so forth (Welfare Quality\textsuperscript{®} 2009c). As the identification of individual animals as well as obtaining the required information from the farmers may be a challenge, such innovative and practical solutions are often necessary for sampling of farm animals in practice.

2.2.7 Validity, reliability and feasibility
An on-farm animal welfare assessment system should provide information about the state of animal welfare on the assessed farm. It is important that the assessment system is able to detect differences in animal welfare between farms and changes in welfare on individual farms (Rousing et al. 2001). It is of course also important that all relevant stakeholders can trust the correctness of the welfare assessment. Especially when the results may have economic consequences for the farmers, the risk of obtaining a wrong assessment must be low (Sørensen and Fraser 2010). The practicability and costs of the system, especially when used in commercial settings, are also highly important (Main et al. 2003, Sørensen et al. 2007).
When considering individual welfare measurements, validity has to do with the extent to which a measurement is meaningful in providing information on the intended aspects of welfare of an animal or a group of animals (Knierim and Winckler 2009, Veissier et al. 2013). Validity can be evaluated in regard to the judgement by one or more experts (face validity), correlation with other welfare measurements (criterion validity) or the experimental documentation of its welfare relevance (construct validity) (Knierim and Winckler 2009, Scott et al. 2001). As described in Section 2.2.1, animal-based measurements are expected to provide more direct information on the welfare of the animals whereas resource-based measurements provide information about the risk of a reduced welfare or potential of improved welfare. This means that animal-based measurements often are considered more valid than the resource-based measurements (Johnsen et al. 2001, Knierim and Winckler 2009, Main et al. 2014, Vannier et al. 2014, Waiblinger et al. 2001). Reliability of individual welfare measurements has to do with the agreement between measurement results obtained by different assessors (inter-observer reliability) and the same assessor (intra-observer reliability) (Knierim and Winckler 2009, Veissier et al. 2013). The inter- and intra-observer reliability of resource-based measurements based on for example measuring fixed constructions is naturally high. However, some resource-based measurements may be difficult to assess as there are categories that in some cases may be challenging to distinguish such as whether something is clean or not. The same applies for many animal-based measurements which are also assessed on scales where the categories may be difficult to distinguish such as body condition. Thus, clear definitions of categories and the boarders between them as well as training of assessors are essential to ensure a high inter- and intra-observer reliability (Knierim and Winckler 2009). However, cases where for example the position of the animal may affect the assessment are hard to avoid. A mink laying on its tail may hide fur chewing or an injury on the tail tip. Thus, any assessment of welfare is dependent on the problem being visible to the assessor at the time of the assessment. Each measurement must also be robust to external factors as the results should be approximately the same, independent of for example weather conditions, time of day, season and date of assessment (Knierim and Winckler 2009, Veissier et al. 2013). This is relevant for both resource- and animal-based measurements as described in Sections 2.2.4 and 2.2.5. However, as many resource-based measurements are related to fixed constructions, these are often more robust. The ability to represent the long-term situation and the sensitivity to changes in the welfare must also be considered for each measurement. However, when there are no major changes on a farm, the results should be similar at different points in
When considering reliability in regard to consistency over time, it is important to take into account that if a measurement is not taken on the same animals on the same day, consistency over time should be evaluated at farm level as the welfare of the individual may have changed, but the welfare at farm level may be unchanged. However, it is important to remember that the lack of reliability of a measurement over time may also be caused by ‘real’ changes in the welfare or methodological reasons such as different observers or samples (Temple et al. 2013). Finally, when considering individual welfare measurements, feasibility has to do with the usability of the selected measurements in practice. They have to be applicable in commercial settings, which means that it should be possible to take each measurement without disturbing the farm’s routines (too much) (Veissier et al. 2013). Furthermore, the measurements should be applicable across different housing systems, and the needed time and the costs should be reasonable (Knierim and Winckler 2009, Sørensen et al. 2007, Veissier et al. 2013).

As described in Section 2.2.2, all potential measurements in Welfare Quality® and related protocols were evaluated in regard to their validity, reliability and feasibility, and only measurements that were found sufficiently valid, reliable and feasible were included in the assessment protocols. The validity, reliability and feasibility of each measurement were balanced, and there were cases where more feasible measurements were selected regardless of a lower validity. For example, in the Welfare Quality® assessment of dairy cows, the evaluation of measurements for the criteria ‘Good human-animal relationship’ showed that ‘Avoidance distance in the barn’ was more valid than ‘Avoidance distance at the feeding place’ (Windschnurer et al. 2009). But due to feasibility reasons, that latter was chosen to be included in the final assessment protocol (Veissier et al. 2013). However, the validity, reliability and feasibility of the complete welfare assessment system, including all related procedures such as planning and sampling of animals, should also be considered. In Welfare Quality®, it was decided that it should be possible for a single assessor to carry out the assessment of each farm during a one-day visit (Veissier et al. 2013). Thus, even if the individual measurements are feasible on their own, they must also be feasible when taken on the same farm on the same day. Another aspect is the validity of the overall assessment of welfare at farm level. Due to the required validity of the included measurements, it may be expected that the result of each individual measurement is related to the welfare of the animals on the farm. But when the results of individual measurements are combined into a welfare score or index, is the result still valid? The
combination of measurement results often includes some kind of weighing of the importance of the different result as in Welfare Quality® and WelFur (Sections 2.2.2 and 3.1.1). In Welfare Quality® and WelFur, experts weighed the importance of the different measurement results based on their evaluation of the welfare of the animals at farm level. The results of combining measurements can therefore be argued to have 'Face validity', which means that the overall assessment of the welfare of the animals on the farm is judged to be valid by experts (Scott et al. 2001). The reliability of an animal welfare assessment system is also related to the education and subsequent quality of the assessors carrying out the assessment. Some assessors may consistently under- or over-score some measurements or the assessment of some measurements may drift over time. If assessors only assess farms in a specific area, potential under- or over-scoring by individual assessors may have long-term consequences for the farmers or the animals as this may lead to misclassification of farms or welfare problems being underestimated (Heath et al. 2014b, Mullan et al. 2011). This may be taken into account by only allowing the same assessor to carry out assessments in the same area for a restricted period of time. Moreover, continuous calibration of assessors may be used to minimise that assessment results drift over time. In welfare assessment systems where only a sample of the animals on the farm is included in the assessment, the sampling method itself may also affect the reliability and validity of the entire assessment. First, samples selected with the sampling method must be representative of the farm that they are taken from. If the sample is not representative, the results cannot be generalised to the welfare of the animals on the farm (EFSA 2012, Jensen and Sandøe 2013). In that case, the assessment of welfare may be considered valid for animals in the sample as it is meaningful with regard to providing information on the welfare of the animals in the sample but not with regard to the animals on the farm. As mentioned in Section 2.2.6, this also related to the required sample size. This means that the sample size should be large enough for the welfare of the animals in any sample to reflect the welfare of the animals on the farm with a sufficiently high accuracy. Second, the sampling method should be independent of the individual assessor, and thus avoid any potential assessor bias to ensure the reliability of the system.
3 Methodology

In this chapter, the aim is to provide an extended background for the methods used in each part of this PhD study as a supplement to the methods described in each paper. All parts of this study use data collected on-farm. Data were collected using methods and procedures related to the WelFur-Mink assessment protocol but adjusted to the aim of each study. This chapter, therefore, starts with a description of the WelFur-Mink assessment protocol. In some parts of the study, WelFur-Mink scores were calculated. This chapter, therefore, also includes a description of the aggregation system. Following this, the methods used in each part of this PhD study are presented.

3.1 The WelFur-Mink welfare assessment protocol

3.1.1 The overall structure of the assessment protocol

WelFur-Mink follows the same overall structure as the Welfare Quality® animal welfare assessment protocols but is adapted to the seasonality of mink reproduction. As in Welfare Quality®, the welfare is assessed at farm level based on a range of animal- and resource-based measurements taken on the farm (Mononen et al. 2012). The included measurements are described in Section 3.1.2. Unlike Welfare Quality®, the overall assessment of welfare at farm level in WelFur-Mink includes three assessments on the same farm (Botreau et al. 2012, Møller et al. 2015). The aim is to assess the welfare of the mink on the farm in each of three assessment periods, each corresponding to one of the three seasons of the annual cycle of mink production described in Section 2.1.1. The three assessment periods are 1) the winter period from 1 December to 28 February covering post pelting where the selected breeding animals are slimmed and prepared for mating, 2) the reproduction and nursing period from 1 March to 15 July, including mating, gestation, birth, nursing, weaning and separation of the kits and 3) the growth period from 16 July to 30 November, where the juveniles are maturing and the following year’s breeders are selected while the rest are pelted. Within each period, the on-farm assessment takes place within a shorter period. These on-farm assessment periods are towards the end of each period where the risk of welfare problems often are higher. In addition, on-farm assessments during events like mating or birth are avoided. The on-farm assessment periods range from 1 January to 20 February in the winter period, from 5 May to 1 July in the reproduction and nursing period, and from 23 September to 30 November. In practice, the beginning and end of some periods are affected by the onset of events such as flush
feeding, date of birth, weaning and live animal grading prior to pelting. Most measurements are taken on a representative sample of the cages with mink on the farm in each on-farm assessment period. This sample consists of the mink in 120 cages in the winter assessment period and in the reproduction and nursing assessment period and the mink in 90 cages in the growth assessment period (Møller et al. 2015). The sampling of mink in WelFur-Mink is further described in Section 3.2.1 and Paper 1. The aggregations of the measurement results into the overall assessment of welfare in WelFur-Mink use the same rules and methods as the Welfare Quality® system where experts and stakeholders were consulted in order to weigh the relative importance of each measurement, criteria and principle in regard to the animal welfare at farm level. This was described in Section 2.2.2. Due to the three assessment periods, the aggregations in WelFur-Mink include one additional layer, that is the aggregation of the measurement results across the three assessment periods (Botreau et al. 2012, Møller et al. 2015), which is illustrated in Figure 3. This means that the results of each measurement in each assessment period (i.e. the collected data in each assessment period) is first transformed into one score for each measurement in each period. This is elaborated in section 3.1.3. As in Welfare Quality®, the scores range from 0-100 (see Section 2.2.2). The scores for each measurement in each assessment period are then aggregated across the three assessment periods into 22 measurement scores before they are aggregated into the 12 criteria- and four principle scores, which is elaborated in Section 3.1.4. The weighing when transforming the results of each measurements in each assessment period into a score for each measurement in each assessment period, and the weights used when aggregating the measurement scores across

![Diagram](image)

Figure 3. Overview of the aggregations of the results of each measurement taken in each of the three assessment periods (Winter, Reproduction and nursing and Growth) into the four principle scores.
the three assessment periods and further into criteria scores were derived from experts in fur animal welfare. The weights used when going from the 12 criteria scores to the four principle scores were derived by taking the average of the weights used in the Welfare Quality® assessment protocols (Botreau et al. 2012, Mononen et al. 2012). Each farm is assigned to one of four welfare categories based on the four principle scores where the threshold values are the same as in Welfare Quality® (Mononen et al. 2012). This is elaborated in Section 3.1.5.

3.1.2 WelFur-Mink measurements
As in Welfare Quality®, the potential measurements were identified from reviewing the existing scientific literature and from knowledge of experts within this field. All potential measurements were evaluated in regard to their validity, reliability and feasibility in each assessment period. In order to assess the welfare according to the 12 welfare criteria that constitute the four welfare principles that were defined in Welfare Quality®, also the individual and marginal relevance of each measurement were evaluated. As a result of this process, 22 measurements were selected of which most are registered in all assessment periods (Mononen et al. 2012, Møller et al. 2015). An overview of the selected measurements and how they relate to the welfare criteria and principles can be found in Table 1 in Paper 2. Some measurements are only relevant or valid in one or two assessment periods. For example, ‘Social housing’ is only relevant in the growth assessment period. Another example is ‘Fur chewing’ which is only considered valid in the winter and growth assessment periods, as ‘Fur chewing’ in the reproduction and nursing period cannot be separated from the changes in the fur related to moulting in practice. As in Welfare Quality®, animal-based measurements were emphasised while relevant resource-based measurements were included when these were considered more valid, reliable or feasible (Mononen et al. 2012).

In some cases, more than one measurement were included in order to cover all aspects of the welfare relevant to the criterion, for example the criterion ‘Absence of disease’ with four underlying measurements. In other cases, only one measurement was included, for example the criterion ‘Absence of prolonged hunger’ where ‘Body condition score’ is the only measurement. Some of the included measurements are registered on scales with only two categories (e.g. ‘Access to a nest box’ – yes or no) while other are registered on scales with several categories (e.g. ‘Skin lesions or injuries to the body’- categories 0, 1, 2 or 3). Some measurements have several underlying measurements, for example ‘Resting quality
of the nest box/resting area’ which includes registrations of whether the nest box/resting area is wet, dirty, damaged and/or infested with fleas. The result of each measurement is used to assess the welfare at farm level. Most measurements provide a snapshot of the situation at the day of the assessment presented as a prevalence, for example the prevalence of mink with a given problem, in a specific category or situation. Other measurements provide the accumulated number of incidences over a period, for example the mortality in each assessment period. For some direct animal-based measurements, the calculated prevalence is also an accumulated incidence. For example, an injury where a part of the tail is missing is a lasting change, and an area with fur chewing will not disappear within the duration of the on-farm assessment period.

3.1.3 From measurement results to measurement scores

The methods used to transform the results of each measurement taken in each assessment period into a score for each measurement in each assessment period depend on the number of categories and underlying sub-measurements. In this section, the primary methods are described.

When a measurement only has two categories (e.g. stereotypic behaviour or not), an I-spline function is applied to transform the prevalence of the problem directly into a measurement score. This function reflects the expert’s evaluation in terms of the welfare at farm level and allows the transformation to follow the expert’s, often non-linear, reasoning. This reasoning follows the same pattern in many cases; the first mink with a problem decreases the score dramatically until a certain point where the score is very low, and more mink with the same problem only have a limited decreasing effect on the score.

When a measurement have more than two but maximum four categories (e.g. no fur chewing and three different severities of fur chewing), weighted sums are calculated by multiplying the prevalence in each category with a weight and summarising the results. The expert’s evaluation of the welfare consequences is expressed in the weights which increase with the severity of the problem, and thereby taking the severity of the problem into account. The I-spline function described above is then applied to transform the weighted sum into a measurement score (Botreau et al. 2012, Møller et al. 2015).

Decisions trees are used to transform the results of a measurement with more than four categories or where there are more than four possible situations at animal level into a score (e.g. the measurement ‘Cage enrichments’ where the enrichment categories and the
possible number of enrichments in each category results in 27 different combinations). They are also used to transform the results of several sub-measurements into one measurement score (e.g. the measurement ‘Age and procedures at weaning’ which consists of three sub-measurements with 20 different combinations of categories). In a decision tree, the prevalence of mink in each combination is calculated. Each possible combination is considered a situation, and each situation corresponds to a score assigned by the experts. This enables the system to take the interactions between different conditions into account in the assigned scores. Decision trees are combined with a percentage rule saying that the assigned measurement score is the one corresponding to the worst situation found on for example at least 10 percent of the mink in the sample (Botreau et al. 2012, Møller et al. 2015).

3.1.4 From measurement scores to principle scores
In WelFur-Mink, Choquet-integrals are used to 1) aggregate the scores for each measurement in each assessment period across the three periods into one score for each of the 22 measurements, 2) aggregate the 22 measurement scores into 12 criteria scores and 3) aggregate the 12 criteria scores into four principle scores (Møller et al. 2015). As described in Section 2.2.2, Choquet-integrals are mathematical operators which have the capacity of only partly allowing a low score in one area to be compensated or outweighed by higher scores in other areas. In practice, the lowest score is the starting point in each aggregation step. The compensation by higher scores in other areas is assigned by multiplying the difference between the lowest score and the second lowest score with a weight. These weights control the size of the compensation and correspond to the relative importance of each score as evaluated by the experts (Botreau et al. 2013). For example, a low score in the criteria ‘Absence of prolonged thirst’ can only to a low degree be outweighed by a higher score in the criteria ‘Absence of prolonged hunger’ as access to water is considered more important than feeding, but a low score in ‘Absence of prolonged hunger’ can to a higher degree be outweighed by higher scores in ‘Absence of prolonged thirst’ (Møller et al. 2015).

3.1.5 Overall categorisation of farms
Each farm is assigned to one of four welfare categories based on the values of the four principle scores. The welfare categories are ‘Best current practice’, ‘Good current practice’, ‘Acceptable current practice’ and ‘Unacceptable current practice’. The WelFur categories have different names than in Welfare Quality®, but the categorisation depends on the same
threshold values (Mononen et al. 2012). This means that for a farm to reach the category ‘Best current practice’, all principal scores must be above 55 and two above 80, for the category ‘Good current practice’, all principal scores must be above 20 and two above 55, and for the category ‘Acceptable current practice’, all principal scores must be above 10 and three above 20. The category ‘Unacceptable current practice’ includes farms where the requirements for ‘Acceptable current practice’ is not met (Møller et al. 2015). These threshold values are summarised in Table 2 in Paper 2. As in Welfare Quality®, an indifference threshold equal to 5 is applied when assigning the scores.

3.2 The methods used in each part of the PhD study

This PhD study is centred around the challenges of using WelFur-Mink on a large scale. Thus, as far as possible, the collected data should reflect the conditions found on mink farms in Europe. The three parts of this PhD study are all based on data collected on mink farms; otherwise, they differ in the methods used to investigate the hypotheses.

In the following section, the background for the methods used in each part is presented as a supplement to the methods described in each paper. Alongside with this PhD study, the WelFur-Mink assessment protocol has been adjusted according to the practical experiences from the on-farm tests and recent research in the field of welfare and welfare assessments of mink. As it has not been possible to publish new assessment protocols for each adjustment, some more practical guidelines for the WelFur-Mink assessment have been developed by the WelFur scientists. Each study always used the most recent guidelines in the assessments; hence, there may be some differences between the studies. As WelFur-Mink was in the process of being implemented in Europe when most of the work for this PhD study was carried out, all possible solutions and adjustments were developed and evaluated in the light of their practical usefulness in the large scale implementation. The structure and content of the original assessment protocol had to be respected as the reference value as this was the basis for the experts’ evaluation of welfare, which decided the weights for the transformation and aggregation of the results of the measurements taken in each assessment period into the overall assessment at farm level. Thus, any adjustments should avoid changing the structure and content of the assessment protocol to such a degree that new expert evaluations would be required.
3.2.1 Part 1 – Sampling method

This part addresses the challenge of sampling cages with mink for the on-farm assessment of welfare in WelFur-Mink. The test of WelFur-Mink on 18 Danish farms in 2014-2015 revealed that it was not always possible to get the information needed for the original stratified sampling method described in the assessment protocol (Møller et al. 2015). Furthermore, the selection of the cages was difficult and time-consuming, and there was a risk of assessor bias, which is elaborated in Paper 1. Therefore, a new sampling method was needed in order to ensure the validity, reliability and feasibility of the assessment protocol. This part of the PhD study included the development of a new sampling method, testing the method, revising the method and finally retesting the method.

The new sampling method was developed in order to replace the original sampling method where each sample was selected as a stratified sample of cages with mink in regard to a range of sub-grouping factors related to the characteristics of the mink and their housing environment. The included sub-grouping factors and their levels are listed in Table 2 in Paper 1. As in the original sampling method, the cages should be selected in clusters of six adjacent cages, which is referred to as cage sections. The reason for the selection of cages in sections is mainly the feasibility of the observation of stereotypic behaviour where a lot of time is saved by observing six cages at the same time. This also limits the time necessary to identify, mark and find the selected cages. For the new sampling method, it was also prioritised that it should be relatively easy to identify the selected cage sections on the farm within reasonable time, and the risk of assessor bias should be reduced. Finally, samples selected with this new method should represent the farm they are taken on.

The limiting factor for the development of the new sampling method was the amount of information that farmers were expected to be able to provide and could be required to provide when signing up for the assessment. It was decided that the number of cages with mink in each shed on the farm was an appropriate amount of information to require from the farmers. A completely random sampling was not considered an option due to the risk that all cage sections in the sample would be selected in the same shed or area of the farm. Thus, a more systematic approach was needed. A first version of the new method was developed in 2015 (Marsbøll et al. 2016). In short, this method is based on first distributing the cage sections between the sheds according to the number of cages in use in each shed. Second, the cage sections are randomly selected within the chosen sheds. In other words:
If $k$ is an interval decided by the total number of cages in use on the farm divided by the number of cage sections to be selected, a cage section must be randomly selected in the sheds where every $k$th cage is in use, counting from the first shed on the farm.

This method utilises the systematics of mink farms. This systematics means that most farms consist of ‘groups’ of sheds where the housing conditions are similar, and the mink are often gathered in ‘groups’ according to, for example, colour type, age, sex and social housing. Hence, if the mink and their housing conditions are ‘grouped’ as expected, larger ‘groups’ of mink (larger than $k$) and their housing environment will always be represented in the sample, while it will vary whether smaller ‘groups’ (smaller than $k$) are represented. The systematic distribution of the cage sections between the sheds also eliminates the risk that all cage sections in the sample are selected in the same shed or area of the farm.

In the test of WelFur on 27 mink farms in 10 European countries in January and February 2016, five assessors tested this first version of the sampling method in practice using an excel sheet developed for the purpose. This excel sheet calculated how many cage sections to select in each shed followed by a random selection of the sections in the respective sheds. Based on this practical test, the sampling method was found sufficiently feasible to be used as a part of WelFur-Mink as the farmers could provide the necessary information, and it was possible to do the entire assessment, including sampling, within one day on even very large and complex farms. However, as the method is not completely random, we wanted to investigate if there was some systematic bias. The representativeness of samples selected with this method was tested by simulated sampling, where samples consisting of 15 cage sections (i.e. 90 cages – as in the growth period) were selected on a model farm. The distribution of the sampled mink was compared to the distribution of the mink on the whole farm in regard to the factors included in the original sampling method (Marsbøll et al. 2016). The model farm mimicked a commercial Danish mink farm in the growth season. This farm was considered a worst-case scenario in regard to sampling as it was larger and more complicated (i.e. more variation within and between sheds) than most Danish farms. Moreover, sampling in the growth period is more challenging than the other periods due to the social housing of the mink. This pilot study showed that for some factors, the distribution of the sampled mink was biased compared to distribution of the mink on the whole farm. This was found to be due to the systematic distribution of the cage sections between the sheds, where the counting of cages always starts within the same shed. The consequence is that on some farms, certain sheds would never be represented in
the sample (Marsbøll et al. 2016). The new sampling method was therefore refined in order to avoid this. The solution was to randomly select the ‘starting shed’. The revised method can be described as:

‘If $k$ is an interval decided by the total number of cages in use on the farm divided by the number of cage sections to be selected, a cage section must be randomly selected in the sheds where every $k$th cage is in use. The counting of the cages in use must begin from a randomly picked shed.’

The revised version of the new sampling method is described in more detail in Paper 1 which also includes an example of how to use the method in practice (Table 1 in Paper 1). In this paper, the representativeness of samples selected with this revised method is also evaluated.

The representativeness of samples selected with the revised method was investigated by simulated sampling where samples consisting of 15 cage sections (90 cages - as in the growth period) were selected on the same model farm as in the pilot study. The distribution of the sampled mink was evaluated in regard to the factors included in the original sampling method. The overall performance of the method when selecting several samples on the same farm was evaluated based on the trueness (closeness of agreement between the simulated samples and the actual farm value) and precision (closeness of agreement between the simulated samples) of the method. However, as only one sample is selected on each farm in practice, and this sample is the basis for the assessment of welfare and the resulting certification, it is important that the risk of selecting a sample that is not representative for the farm is low. Therefore, the probability of selecting a sample where the prevalence of mink in the sample in regards to the included factors is within different ranges of the farm prevalence was estimated. This was used in the discussion of what is an acceptable representativeness when sampling mink for welfare assessments in practice.

It was not possible to compare the performance of the new sampling method directly to the original stratified sampling method as it is not possible to simulate sampling according to the original method. The reason for this is that for each simulated sample, the original method requires the acquisition of the information needed for the stratification, the calculation of which cage sections to sample in which sheds and the actual selection of the samples on the farm, including the potential human bias in doing so. To be done correctly this would require one trained assessor for each sample. However, in principle, all
stratified samples should be representative in regard to the chosen strata (Lohr 2010). Thus, by using the actual distribution on the farms as the reference value, the performance of the new method is compared to the expected distribution of any stratified sample.

As the new sampling method will perform better (i.e. increased representativeness of the samples) on less complicated farms, we chose only to test it in this worst-case scenario. This model farm with about 98,000 mink in the growth period may not be representative for the average Danish or European mink farm due to the size of the farm. Based on information from WelFur-Mink assessments conducted in the winter assessment period 2018 where more than one third of all European mink farms were assessed, it seems that average farm size is around 2,900 breeding females per farm in Denmark and around 4,500 breeding females per farm in Europe (V. Gospodinova, Fur Europe, personal communication, 21 August 2018). Depending on the breeding result, this corresponds to around 17,400 mink per farm in the growth period in Denmark and around 27,000 mink per farm in the growth period in Europe. It was important that the method performs well also on larger farms in Denmark and across Europe, where there are farms with more than 40,000 breeding females (V. Gospodinova, Fur Europe, personal communication, 21 August 2018), which corresponds to around 240,000 mink in the growth period depending on the breeding result. The representativeness was evaluated in regard to factors related to the characteristics of the sampled mink and their housing environment. These factors were included in the original sampling method in WelFur-Mink and are known or believed to be related to mink welfare. As also discussed in Paper 1, other studies have investigated the performance of sampling strategies by evaluating how the welfare of the animals in the sample reflects the welfare of all the animals on the farm. This was not considered a possibility in this study as we prioritised to use a quite large farm as a model. Assessing the welfare of all 98,000 mink on the model farm would have been unfeasible within the framework of this study.

3.2.2 Part 2 - Changes in welfare with date of assessment

In this part, the potential change in the assessed welfare with date of assessment during the on-farm assessment periods of the winter and growth periods is investigated. If date of assessment affects the results of the welfare assessment, it could be necessary to adjust for this in order to ensure the reliability of the assessment. This study did not include the on-farm assessment period of the reproduction and nursing period as the changes during this period previously have been studied by Henriksen and Møller (2015).
This study was designed as an on-farm study with repeated WelFur-Mink assessments of the same sample of mink on each farm during the on-farm assessment periods of the winter and growth periods with one to four weeks interval between assessments. It could have been possible to select a new sample of cages for each assessment, but in order to eliminate the risk of even small differences in welfare caused by selecting a new sample interfering with the effects of time, it was chosen to assess the same samples repeatedly. This was also the reason why only cages where the number of mink was constant throughout each on-farm assessment period were included in the analysis. By excluding cages where the number of mink changed during the on-farm assessment periods, mink that were sick or injured during each assessment period may have been excluded as the farmers may have moved them to a ‘hospital section’ or put them down during the on-farm assessment period. Hence, we may have decreased the prevalence of welfare problems in the sample, but at the same time ensured that there were no changes in the sample that could interfere with the effects of time. To avoid potential effects of changes in the environment to interfere with the effects of time, cages where the mink were moved to another location were also excluded from the analysis. For the measurement ‘Temperament test’, repeated testing may have resulted in the mink habituating to the test situation, which may have affected the result. This is further discussed in Paper 2.

The data collection for this study was a part of the Danish test of WelFur-Mink in 2014-2015 where the welfare on 18 commercial mink farms were assessed in all three assessment periods. The welfare on eight of the included farms was assessed four times during the on-farm assessment period of the growth period and three times during the on-farm assessment period of the winter period in order to be able to investigate the potential changes in the WelFur-Mink assessment with date of assessment during these on-farm assessment periods. Advisors from Kopenhagen Fur were responsible for selecting the farms included in this study. The criteria for selection were that the farms should differ in size, geographical location, feed supplier, housing conditions and combinations of colour-types.

Four advisors from Kopenhagen Fur, two in each assessment period, performed the assessments on the eight farms, and were in close contact with the WelFur-Mink scientists to ensure that all measurements were assessed as correctly as possible. The undersigned was one of the assessors and was responsible for managing the collected data and hence had in-depth knowledge of the data when the PhD study began. The first WelFur-Mink
Methodology

assessment protocol was published at the time of data collection and was the basis for the assessments in this study (EFBA 2013). However, as the data collection was a part of testing the assessment protocols, some procedures and categories have been adjusted since then. This means that the farms would not necessarily get the same assessment result today. In addition, the generalisation of the results to the assessment system used today may be limited for measurements where the procedures or categories have been adjusted. The largest change is in regard to the assessment of ‘Stereotypic behaviour’ where the procedure for the timing of the observations was adjusted from ‘1 hour before expected feeding’ to ‘1 hour before expected feeding OR as the last observation of the day’ (assessment guideline used in the data collection for this study) and finally to ‘1.5 hours before expected feeding OR before 1.5 hours before sunset’ (present assessment guideline).

For practical reasons, it was not possible to keep the assessor constant at farm level, which means that the same farm may have been assessed by one or two assessors during each on-farm assessment period. For that reason, assessor had to be included as an effect in the analysis of the change in the assessed welfare with assessment date. The distribution of the assessment dates on each farm was not evenly over the on-farm assessment periods. As an example, this means that the first assessment on one farm in one assessment period may be taken after the second assessment on another farm. The main interest was whether the assessed welfare differed between assessments in the beginning, the middle and the end of each on-farm assessment period. It was therefore decided to divide each on-farm assessment period into sub-periods which were used as explanatory variable.

In this study, the main focus was on the possible change with date of assessment in the result of selected measurements during the on-farm assessment periods of the winter and growth periods as well as the related WelFur-scores and the overall categorisation of the welfare at farm level. The selected measurements were ‘Body condition score’, ‘Fur chewing’ and ‘Stereotypic behaviour’ in the winter period and ‘Skin lesions or injuries to the body’, ‘Diarrhoea’ and ‘Temperament test’ in the growth period. This study, from study design to data analysis, largely followed the same framework as the study by Henriksen and Møller (2015) where the potential changes in the WelFur-Mink assessment with date of assessment during the on-farm assessment period of the reproduction and nursing period were investigated. This was intentional in order to increase the comparison between the studies as they were expected to form the basis for any future adjustments of the assessment protocol in regard to the robustness to changes with date of assessment. One
difference between the studies is that the number of farms included was increased in present study. Moreover, in the present study, each farm was assessed in all three assessment periods, hence, the farms’ own data could be used for the calculations of all WelFur-Mink scores. The farms included in the study by Henriksen and Møller (2015) were only assessed in the reproduction and nursing period, and an average over assessments collected on other farms was used in the calculation of the WelFur-Mink scores. This may have decreased the variation between the farms compared to this study.

3.2.3 Part 3 – Stereotypic behaviour in the winter assessment period
This third part addresses the challenges regarding the timing of the assessment of the measurement ‘Stereotypic behaviour’ in the on-farm assessment period during the winter period in WelFur-Mink. If the prevalence of stereotypic behaviour in the winter period is dependent on the timing of the observation, this has to be adjusted for to ensure the reliability of the measurement. The prevalence of stereotypic behaviour is expected to peak at feeding and shortly after sunset (Hansen and Damgaard 2009, Malmkvist et al. 2013). The prevalence of stereotypic behaviour has been shown to be higher before feeding than before sunset, when evaluating the percentage of time spent performing stereotypic behaviour during 15 minutes of video recordings (Malmkvist et al. 2013) or percent of scans where stereotypic behaviour was observed (Hansen and Damgaard 2009). However, stereotypic behaviour in these studies was registered differently than in the winter period in WelFur-Mink where the occurrences within a 2-minute observation period are registered within 1.5 hours before expected feeding time (before feeding) or 1.5 hours before time of sunset (before sunset). In this study, the difference in the prevalence of stereotypic behaviour observed before feeding and before sunset in the winter period was therefore investigated using the same registration method as in WelFur-Mink. This study was based on data from two sources. The first source was an on-farm study where observations of stereotypic behaviour according to WelFur-Mink were carried out before feeding AND before sunset on the same sample of mink on five Danish farms in the winter period 2017. The second source was results from all European WelFur-Mink assessments in the winter period 2018 where stereotypic behaviour was observed before feeding OR before sunset.

The five Danish farms were selected for the study based on a few criteria. For practical reasons, the driving distance from AU-Foulum should be less than 1.5 hours. It should also be possible to sample brown first-year female mink housed in the same shed, due to data
collection for another study. Furthermore, feeding should take place relatively early in the
day, which means that it should be farms where it could have been challenging to observe
stereotypic behaviour before feeding in practice. Four commercial mink farms in addition
to the research farm at AU-Foulum were selected. A sample of 150 brown first-year female
mink were included on each farm. For the other study, the welfare of the same sampled
mink was assessed according to the WelFur-Mink assessment protocol, including
observations of stereotypic behaviour before feeding and collection of faecal samples for
analysis of faecal cortisol metabolites (Marsbøll et al. 2017). For the present study,
observations of stereotypic behaviour before sunset were included. The same mink were
observed twice (before sunset and before feeding) in order to avoid potential effects of
differences between samples. In order to ensure that the clinical assessment of welfare and
collection of faecal samples would not affect the behaviour of the mink, observations of
stereotypic behaviour before sunset were done on day 1, observations of stereotypic
behaviour before feeding on day 2, and the remaining WelFur-Mink registrations and
collection of faecal samples after feeding on day 2. Two assessors each observed the same
half of the sampled mink on each farm on both days in order to avoid potential assessor
effects.

Only brown first-year female mink were included for the study on the five Danish farms.
This was mainly a restriction imposed by the other study where it was important to
standardise age, sex and colour type. Except on Farm A, observations of stereotypic
behaviour were carried out within 1.5 hours before expected feeding time and within 1.5
hours before sunset (Table 1). On farm A, observations before sunset were carried out as
on the other farms. However, on this farm, feeding would normally start just after sunrise
(i.e. at 08:30). Thus, in order to be able to actually see the mink during the observations
before feeding, feeding was postponed one hour (i.e. to 09:30), and the observations were

Table 1. Expected feeding time and time of sunset (according to The Danish Meteorological Institute) on
the observation days in February 2017 on each of the five Danish farms.

<table>
<thead>
<tr>
<th>Farm</th>
<th>Dates in February 2017</th>
<th>Time of sunset day 1</th>
<th>Expected feeding time day 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>2-3</td>
<td>16:54</td>
<td>08:30^a)</td>
</tr>
<tr>
<td>B</td>
<td>6-7</td>
<td>17:04</td>
<td>09:30</td>
</tr>
<tr>
<td>C</td>
<td>9-10</td>
<td>17:11</td>
<td>10:00</td>
</tr>
<tr>
<td>D</td>
<td>13-14</td>
<td>17:20</td>
<td>10:45</td>
</tr>
<tr>
<td>E</td>
<td>16-17</td>
<td>17:27</td>
<td>12:10</td>
</tr>
</tbody>
</table>

^a) The mink was normally fed at 08:30 but on the day of the observation, feeding was postponed one hour.
carried out during this time. This was not expected to affect the result as postponing feeding up to 1.5 hours has been found to have no effect on the frequency of stereotypic behaviour (Møller et al. 2014). At each farm, the observations started and ended at approximately the same time before feeding and before sunset. For practical reasons such as the farmer’s management routines it was not possible to have the exact same timing of the observations on all farms, and the timing (start of the observations before feeding and before sunset) varied with up to 25 minutes between farms. As in WelFur-Mink, expected feeding time was based on information from the farmer, and may therefore be somewhat inaccurate. The individual mink’s perception of expected feeding may also vary, and the time from the feeding begins until all mink are fed may be very long on some farms. Based on previous studies, the prevalence of stereotypic behaviour is expected to peak at feeding and shortly after sunset (Hansen and Damgaard 2009, Malmkvist et al. 2013). Thus, if the timing of the observations before feeding and before sunset is not the same on all farms, observations on some farms may have taken place closer to the peak of the curve than on other farms. Differences in the mink’s perception of feeding time and time of sunset may have a similar effect, which is discussed in regard to the mink’s perception of time of sunset in Paper 3. Furthermore, at an individual level, some mink may have been observed earlier on the curve than others, as it took approximately one hour to observe the mink. Several observations of each individual mink before feeding and before sunset could help identifying this curve. All observations were done over a period of two weeks in February 2017 (Table 1). Results from Paper 2 suggested that there was no reason to expect that the prevalence of stereotypic behaviour would change with assessment date during the on-farm assessment period of the winter period. Possible changes in the results during the assessment period were therefore not taken into account.

The reason for including results from the European WelFur-Mink assessments in the winter period 2018 was to investigate whether the results from the five Danish farms could be generalised to European conditions. In the European WelFur-Mink assessments, stereotypic behaviour was observed either before feeding or before sunset. Thus, it was not possible to investigate the differences in the prevalence of stereotypic behaviour before feeding and before sunset on individual farms. However, due to the high number of assessed farms (assessments from 1,166 farms were found to be complete), the average difference between the prevalence of stereotypic behaviour before feeding and before sunset was expected to provide a reasonable estimate. It was not possible to get information about the individual assessors; thus, the potential effect of assessor could not
be taken into account. This effect is also confounded with country as most assessors only operate in one country. However, assessment of stereotypic behaviour in general has a high inter- and intra-observer reliability (Møller et al. 2012). Moreover, all assessors followed the same WelFur-Mink training programme, which included a one-week course in 2016 or 2017, where assessment of stereotypic behaviour was taught by the undersigned. They also participated in a training and calibration day in the beginning of the on-farm winter assessment period 2018. The potential effect of assessor is therefore expected to be limited.
4 Results

This chapter consists of the manuscripts for three papers which have been submitted for publication during the PhD study.
4.1 Paper 1

The representativeness of a semi-random sampling method for animal welfare assessments on mink farms

Anna Feldberg Marsbøll, Britt Ingeborg Foseide Henriksen and Steen Henrik Møller

Submitted for publication as a technical contribution in Animal Welfare
(Note: Included in the stage it was in when the thesis was submitted in September 2018. Since then, a revised version has been accepted for publication in Animal Welfare)
The representativeness of a semi-random sampling method for animal welfare assessments on mink farms

Anna Feldberg Marsboll, Brit Ingeborg Fosside Henriksen and Steen Henrik Møller

Department of Animal Science, Aarhus University, Blichers Allé 20, 8830 Tjele, Denmark

Corresponding author: Anna Feldberg Marsboll
E-mail: anna.marsboll@anis.au.dk

Running title: A simple method for sampling mink
Abstract

In this study we present a semi-random sampling method developed for the sampling of mink (Neovison vison) for on-farm welfare assessments according to the WelFur-Mink system. The only information needed for using this method is the number of cages in use in each shed on the farm. The representativeness of samples selected with this method was evaluated in relation to the physical characteristics of the farm and the mink characteristics by simulated sampling on a farm with a complicated structure in the growth period. The selection of 10,000 samples was simulated. The trueeness was in general high, i.e., the method has no systematic skewness. The precision was low for some factors due to the high variation within sheds. The sampling in sections of six adjacent cages means that it is often not possible to select a sample which is an exact representation of the mink and their housing environment. If accepting a deviation of ± one cage section, the estimated probability of selecting a representative sample was high for most of the individual factors. However, the estimated probability of selecting a sample that it representative according to all factors was rather low. This deviation from exact representativeness ought to be evaluated in the light of the increased feasibility and repeatability offered by the method. Also, we expect that the representativeness of samples selected with this method will be higher on other less complicated farms. We suggest that this simple method balances feasibility and representative sampling in a way that makes it useful in the WelFur-Mink system.

Keywords

animal welfare, feasibility, mink, on-farm welfare assessment, simulation study, WelFur
Introduction

When assessing animal welfare at farm level, it is usually not possible to include all the animals on the farm because this would be too time consuming and, thus, unfeasible for practical reasons. Hence, welfare assessments are often based on a sample of the animals, and the welfare of the animals in the sample is considered an estimate of the welfare of the animals on the farm. It is, therefore, important that the sample is representative, i.e., that the animals and housing conditions of the animals in the sample reflect those on the farm. Different animal welfare assessment systems sample according to different rules depending on species and production system. In some systems, a random selection of the animals is suggested, as for example random selection from microchip numbers in horses (AWIN 2015). However, when selecting a number of sows in a large pen, the sows cannot easily be identified. Therefore, a more practical approach was developed in Welfare Quality®. The first sow in sight is the starting sow. The next sow is ‘the sow who’s head is the fourth away (facing) from the ‘starting sow’. and so forth (Welfare Quality® 2009a). Many sampling strategies also include some kind of stratification. This means that the animals in the sample should be distributed so that the relevant subgroups on the farm are represented in the sample according to the size of the subgroup, which often increases the representativeness of the sample compared to random sampling (Lohr 2010). One example is dairy cows kept in different groups where the number of cows sampled in each group should be proportionate to the size of the group (Welfare Quality® 2009b).

In the on-farm welfare assessment system Welfur for foxes (Vulpes vulpes/lagopus) and mink (Neovison vison), the assessment is based on a sample of the animals on the farm. This system is largely inspired by the Welfare Quality® project (Mononen et al 2012). Welfur was implemented in all European mink and fox farms on a voluntary basis starting from January 2017, and the resulting
assessments are used for welfare certification of the pelts from these farms (Fur Europe 2017). As
in Welfare Quality®, animal-based measurements are preferred over resource- and management-
based measurements (Mononen et al 2012), and in the assessment protocol for mink (Welfur-Mink)
nine out of 22 measurements are animal-based (Møller et al 2015). Mink are seasonal breeders, and
the strict annual production system can be divided into three seasons, each associated with specific
animal groups: 1) winter season with adult breeders on the farms, 2) nursing season with mostly
adult females and kits on the farms and 3) growth season with mostly adult females and juveniles
on the farms (Møller et al 2003). Due to the seasonal production system, the full welfare assessment
according to the Welfur-Mink protocol includes one assessment in each season. In each season, a
sample of mink is selected and used for the assessment of all animal- and resource-based
measurements (Møller et al 2015). Mink farms typically consist of several sheds in which the mink
are kept in rows of cages elevated off the ground. The cages are generally constructed as battery
cages with approximately 2 metres between the bearing posts. This means that each row can be
physically divided into battery cage sections which, depending on the design, often consist of 5 to 8
cages each (Jørgensen 1985). In order to increase the feasibility of the assessment, and especially
the assessment of stereotypic behaviour where minimum two minutes of observation is required, the
sample is selected in sections of six adjacent cages. The sample consists of 15 sections of six
adjacent cages (90 cages) in the growth period and 20 sections of six adjacent cages (120 cages) in
the winter- and nursing period (Møller et al 2015). Each cage section in the sample thus represents
6.7% of the sample in the growth period (6 out of 90 cages) and 5% in the winter- and nursing
period (6 out of 120 cages). In the winter period, there are only adult breeders on the farms. They
are usually kept individually, hence the sample in this period consists of 120 mink. In the nursing-
and growth period, the sample consists of a varying number of mink. In the nursing period, most
66
67
after the mating period. In the growth period, adult females are typically housed individually or
with one or two juvenile kits, while the rest of the juveniles are typically housed in pairs or groups.

In WelFur-Mink, the original sampling method used stratification in order to ensure a representative
sample. The subgrouping factors were sex, age and colour type in the nursing- and winter period,
and sex, age, social housing conditions and colour type in the growth period. Besides this, different
types of sheds, cages, nest boxes and watering systems should also be considered (Møller et al
2015). Due to the many subgrouping factors in the stratification, a lot of information was needed
beforehand. Practical tests of WelFur-Mink before implementation showed that retrieving this
information could be a challenge. In case the farmers did not have the information, the feasibility of
the assessment would be impaired if the assessors should attempt to collect the information
themselves, as this would be impossible to do within one day on most farms. In case it was possible
to get the information needed, the practical test also showed that it was difficult and time-
consuming to select the cage sections on the farm in a way that made the resulting sample
representative according to all factors. Hence, the original sampling method was not applicable in
practice. Furthermore, the individual assessor was responsible for selecting the cage sections on the
farm based on the stratification. Thus, the resulting samples could be unintentionally biased. In
order to overcome these problems we developed a new sampling method. This new method takes
the systematic structure of mink farms into account and the only information that is needed is the
number of cages in use in each shed on the farm. The sampling of the cage sections in the sample is
based on 1) a systematic distribution between the sheds according to the number of cages in use in
each shed and 2) a random selection within the sheds, hence it can be considered a semi-random
sampling method. A preliminary version of the method was described by Mølbøll and colleagues
(2016). In the present paper we present an improved and revised method.
The purpose of this study was to describe this new semi-random sampling method and to explore the representativeness of samples taken with this method as well as the probability of taking a representative sample. This was done by simulated sampling in the growth period on a farm with a complicated structure. We chose the growth period due to the social housing of the mink and because almost all cages are in use in this period, which makes sampling more challenging than in the other periods.

Materials and methods

Model farm

A model farm was used to test the sampling method. This model farm mimicked the set-up of a private Danish mink farm, thus ensuring a realistic model, and was also chosen due to its complexity. This complexity was made up both by the size of the farm as well as the many different combinations of social housing, colour types and housing conditions on the farm. Such complexity would make it particularly challenging to use the original sampling procedure described in the WelFur-Mink assessment protocol. In October 2015 each battery cage section on the private farm was described according to the mink’s characteristics, the social housing conditions and the housing environment. If there were any differences within the battery cage sections, the most frequent situation was noted. This description of the private farm formed the basis for the model farm.

On this model farm, there were about 42,000 cages divided between 23 sheds and about 98,000 mink. As typical for this period, there were no adult males on the farm. The ratio between the adult females and juveniles was 1:5 and there were several different colour types. The number of cages in use in each shed ranged from 500 to 6800. In some sheds, there were only mink in one stage (eg
Results - Paper 1

145 adult females) or one colour type. In other sheds, there were mink in several stages (eg adult
146 females and juveniles) and colour types housed in different combinations (eg one adult female with
147 one juvenile male, or one juvenile female with one juvenile male). Housing conditions differed
148 between the sheds, and some housing-related factors (eg cage type, nest box insulation and nest box
149 position) also differed within individual sheds.
150
151 Sampling method
152 The only information needed for using the new sampling method on a mink farm is the number of
153 cages in use in each shed on the farm. We expect the farmers can provide this information, but, if
154 not, the assessors can collect it themselves. This will still be time-consuming, but possible within
155 the limitations of one workday for each welfare assessment. As an example, the sampling of 15
156 cage sections in the growth period on a farm with mink in seven sheds is shown in Table 1. First,
157 each shed is given a sequential number. Most mink farms have a systematic layout and the
158 numbering of the sheds should follow this systematics. For example, if there are two row sheds in
159 one area and multi row sheds in another area, all the sheds in one area should be numbered before
160 the sheds in the other area. Next, a randomly picked shed number identifies which shed is the
161 starting one in the calculations. The sample threshold is the total number of cages in use on the farm
162 divided by the number of cage sections (eg 15 or 20) in the sample. The ratio between the sample
163 threshold and the number of cages in use is calculated for the stepwise addition of each shed. This
164 means that first the ratio between the sample threshold and the number of cages in use in the
165 starting shed is calculated. This ratio is rounded to its nearest integer, and this number is the number
166 of cage sections to be randomly selected within the starting shed. Subsequently, the ratio between
167 the sample threshold and the number of cages in use in the starting shed plus the following shed is
168 calculated. This ratio is rounded to its nearest integer, and this number minus the number of cage
sections that have been selected in previous sheds, is the number of cage sections to be randomly
selected in the second shed. This procedure is repeated until the number of cages in use in all sheds
has been added. The random selection in the respective sheds is based on the sheds physical
division into battery cage sections of 5 to 8 cages. If the calculations show that one cage section
should be selected in a shed, which can be divided into eg 50 battery cage sections, one of the 50
battery cage sections is randomly selected and included in the sample. However, as the sample is
based on sections of six adjacent cages, if the battery sections consist of eg five cages, an additional
cage must be included from one of the neighbouring physical sections, and if the battery section
consists of eg seven cages, one cage must be omitted.

Testing the sampling method

A program that was able to simulate the new sampling method was developed in R (R Core Team
2017). The selection of 10,000 samples on the model farm was simulated in order to have a
sufficient number of samples to examine the representativeness of samples selected with the
method. Each sample consisted of 15 sections of six adjacent cages as in the growth period in
WelFur-Mink. The prevalence of mink in relation to the levels of the subgrouping factors included
in the original stratification in WelFur-Mink was calculated for the simulated samples and for the
model farm (Farm prevalence). The subgrouping factors and their levels are listed in Table 2. Only
factors and levels that showed variability within the model farm were included in the analysis. This
means that the following factors and levels were omitted: the factors ‘Watering system’ (all mink
had access to an automatic and frost protected water supply) and ‘Nest box presence’ (all mink had
access to a nest box), and the factor levels ‘Adult males’ (there were no adult males), ‘Adult
females family housed with their litter’ (this housing was not used), ‘Adult females housed with
other adults’ (this housing was not used), ‘Black’ (there were no mink with this colour type),
‘Mahogany’ (there were no mink with this colour type) and ‘Cage wall other than wire mesh or solid’ (all cage walls were wire mesh or solid).

The representativeness of the method was evaluated based on trueness and precision. Trueness refers to the closeness of agreement between the simulated samples and the actual farm value, while precision refers to the closeness of agreement between the simulated samples. Trueness and precision were assessed individually for each of the included factor levels. The trueness was assessed based on the Error, in order to evaluate whether there is a systematic bias. The Error for each factor level was calculated as the median of the sample prevalences minus the farm prevalence. The precision was assessed based on the 95% central range in order to evaluate the range of the sample distribution. The 95% central range for each factor level was calculated as the 97.5 percentile of the sample prevalences minus the 2.5 percentile of the sample prevalences.

The probability of selecting a representative sample was assessed by estimating the probability of selecting an individual sample where the prevalence of mink in the sample is within a range of ± 5, 10, 15, 20 and 25 percentage points of the farm prevalence. This was done by calculating the share of samples where the prevalence of mink in each sample is within a range of ± 5, 10, 15, 20 and 25 percentage points of the farm prevalence for each factor level and for groups of factors and their levels. All calculations were made in R (R Core Team 2017).

Results

Each sample consisted of 15 sections of six adjacent cages, i.e. each sample always consisted of 90 cages. The number of mink in the samples varied from 168 to 246 with a median of 210. All samples included one or more cage sections from the six largest sheds, while it varied how many
cage sections the samples included from the smaller sheds. Cage sections were selected in all sheds.

The prevalence of mink on the model farm and the distribution of the prevalences in the simulated samples for each factor level are shown in Table 3. The Error was low for most of the included factors, ranging from -2.6 to 1.2 percentage points. The lowest Error was for the prevalence of the factor levels ‘Adult females’ and ‘Juveniles’, while the highest was for the prevalence of the colour type ‘Palomino’. The range of the sample distributions varied from a 95% central range of 4.2 percentage points for the prevalence of the factor level ‘Housing of juveniles - individually’ to a 95% central range of 43.0 percentage points for the prevalence of the factor ‘Housing of adults’.

The estimated probability of selecting a sample where the prevalence of mink in the sample is within ± 5, 10, 15, 20 or 25 percentage points of the farm prevalence for each factor level is shown in Table 4. The estimated probabilities increased with the range of the deviation from the farm prevalence. For some factor levels, the estimated probability was low at a deviation range of ± 5 percentage points but increased to a probability above 0.90 already at a ± 10 percentage points deviation range (eg the ‘Palomino’ colour type with an increase from an estimated probability of 0.57 to 0.94). A few factor levels had a smaller increase and did not reach an estimated probability above 0.90 until the deviation range was increased to ± 25 percentage points (eg ‘Housing of adult females – individually’ with an increase from an estimated probability of 0.32 at deviation range of ± 5 percentage points to 0.95 at a deviation range of ± 25 percentage points). For a few factor levels, the estimated probability was above 0.90 for a deviation range of ± 5 percentage points (ie the factors ‘Nest box position’ and ‘Shed type’ and the factor levels ‘Housing of juveniles – individually’ and ‘Nest box insulation – low’).
The estimated probability of selecting a sample where the prevalence of mink in the sample is within ± 5, 10, 15, 20 or 25 percentage points of the farm prevalence for groups of factors is shown in Table 5. The estimated probability of selecting a sample where the prevalence of mink in the sample are within a deviation range of ± 5 percentage points in all factor levels was very low, but increased to 0.52 at a deviation range of ± 15 percentage points and reached an estimated probability above 0.9 at a deviation range of ± 0.25 percentage points. The estimated probability of selecting a sample, where the prevalence of mink in the sample are within a given range for all factor levels in a group, were above 0.9 at a deviation range of ± 15 percentage points for the factors related to housing environment. For the factor levels related to colour types, the probability were above 0.9 at a deviation range of ± 20 percentage points, and for the factor levels related to sex, age and housing conditions, the probability were above 0.9 at a deviation range of ± 25 percentage points.

Discussion

The systematics of the method

With this semi-random sampling method, the cage sections in the sample are selected based on a systematic distribution between the sheds according to the number of cages in use in each shed. As a consequence of this method, cage section is always selected in sheds or a group of sheds where the total number of cages in use exceeds the sample threshold, is larger ‘groups’ of mink and their housing environment will always be included in the sample, while it may vary whether smaller ‘groups’ will be included. For example, if a farmer keep mink with a special colour type in one part of the farm, these will always be represented in the sample if the number of mink in this group is larger than the sample threshold. The individual samples selected on the model farm presented in this study always included one or more cage sections from the largest sheds, while it varied from
which smaller sheds cage sections were included. As more mink were housed in the larger sheds, they were also more representative for the model farm. This systematic approach, therefore, increases the representativeness of the method. Also, the risk of selecting all cage sections in the same shed or groups of sheds, as could have been the case with a completely random sampling, is excluded.

Trueness and precision

Trueness and precision were used to evaluate how well the method performed when selecting several samples on the same farm. The Error was low for most of the included factors. This means that the method has no systematic skewness, i.e. the systematics of the method did not increase nor reduce the probability of selecting a sample with specific characteristics. However, the range of the sample distribution varied between the factors. The farm that we used as a model had a large variation within the sheds in colour type, sex, age and the social housing of the mink, but, surprisingly, also in housing conditions as there could be e.g. several different cage and nest box designs within the same shed. This has decreased the precision of the method as the random selection, in combination with a large variation, results in several different combinations of the different factors and their levels. We, therefore, expect that samples selected on farms with less variation within sheds will have a higher precision than in this study.

The probability of selecting a representative sample

In practice, only one sample is selected on each farm. The sample is selected in sections consisting of six adjacent cages where the housing conditions are similar and the mink most often share the same characteristics. As each cage section represents 6.7% of the total sample in the growth period and 5% in the winter- and nursing period, is it only in rare cases possible to select a sample that is
an exact representation of the mink and their housing environment. For example, if the actual
prevalence of males on a farm in the winter period is 17.5%, the closest representation in the
WelFur sample would be 15% (three cage sections with males) or 20% (four cage sections with
males). And as the number of mink per cage varies in the nursing and growth periods, some cage
sections will also represent more than 5% and 6.7% of the total sample in these periods. Thus, no
matter which sampling method is used, we cannot expect the samples to be an exact representation
of the farm they are selected from. But how large a deviation from the actual farm value can be
considered representative? If no deviation is accepted, the range of the closest representation will be
5 percentage points in the winter period and minimum 5 and 6.7 percentage points in the nursing-
and growth period, respectively. If a deviation of ± one cage section is accepted, the range of an
acceptable representation is increased to 15 (± 7.5) percentage points in the winter period and
minimum 15 (± 7.5) and 20.1 (± 10.5) percentage points in the nursing and growth periods,
respectively. If a deviation of ± 10 percentage points is considered acceptable in the growth period,
the estimated probability of selecting an acceptable representative sample in this study was 0.95 for
age and sex, and between 0.80 and 1.00 for all the housing-related factors, which we consider a
quite high probability. For the factor levels related to colour type, the probability ranged from 0.67
to 0.97, and for the factor levels related to social housing of females and juveniles, the probability
ranged from 0.42 to 1.00. This means that the probability is quite low for some factors. However, if
we increase the acceptable deviation to ± 15 percentage points, the probability of selecting a
representative sample is increased, i.e. for the factor levels related to colour type the estimated
probability ranged from 0.83 to 0.99, and for the factor levels related to social housing of females
and juveniles the estimated probability ranged from 0.82 to 1.00. Thus, for individual factors, the
estimated probability of selecting a representative sample is quite high if ± 15 percentage points is
considered an acceptable deviation. However, when looking at the estimated probability of selecting
a sample, where the prevalence of mink in the sample are within a deviation range of ± 15 percentage points for all factor levels, the estimated probability is only 0.53. When considering groups of factors, the estimated probability of selecting a sample where the prevalence of mink in the sample are within a deviation range of ± 15 percentage points is 0.92 for the factor levels related to housing environment, 0.75 for colour types and 0.74 for sex, age and social housing. This indicates that there is a high probability that the individual sample is representative according to the factor levels related to housing environment, while the probability of a representative sample in regards to age, sex, social housing and colour type is lower. But due to the high probability of selecting a sample that is representative in regards to the individual factor levels, it seems that it can be expected that a sample that is not representative to all factor levels, as a minimum is representative according to some factor levels.

Choice of method

Large efforts have been put into the development of valid and reliable welfare indicators, as this is crucial for a correct assessment of animal welfare at farm level (e.g. Mononen et al. 2012, Veissier et al. 2013). However, if the entire farm is not included in the assessment, a correct assessment requires that the sample is representative in relation to both the physical characteristics of the farm and the animals on the farm. If the sample does not represent the farm, the assessment may be misleading, even though the sample is assessed correctly using both valid and reliable welfare indicators. Also, sampling must be reliable, hence assessor bias must be avoided. In the present study, a feasible procedure for taking a sample of mink at farm level without the risk of assessor bias is presented. To our knowledge, this is the first published study evaluating a sampling method for taking a sample of animals for welfare assessments at farm level by investigating how well the samples represent the farm with regards to the characteristics of the animals and their housing
environment. Other studies have evaluated sampling strategies by assessing how the welfare of the
animals in the sample reflects the welfare at farm level (eg Main et al 2010). In our case, this was
not possible as it would have taken us several months to assess the welfare of the 98,000 mink on
the model farm. On a smaller farm, this would have been possible, but we aimed to test the method
in a worst-case scenario, ie with as many different combinations of housing systems, type of
animals etc as possible. We, therefore, believe that evaluating the performance of the methods in
regards to the welfare of the mink in the sample should be based on simulating different prevalences
of welfare problems on a model farm.

Using the semi-random sampling method in WelFur-Mink

The samples selected in this study were not all acceptable representations of the farm, even if a
deformation of a one cage section was accepted. However, this deviation from exact
representativeness ought to be evaluated in the light of the increased feasibility and reduced risk of
assessor bias, which the method offers. Also, according to our experience from welfare assessments
across Europe, the farm that was used as a model in the study was more complicated (ie more
variation within and between sheds) than most farms. The farm was chosen for that exact reason,
and this study is, therefore, a ‘worst-case scenario’. Hence, we expect that the representativeness of
samples selected with this method will be higher on other less complicated farms. Based on the
results and discussions above we, therefore, suggest this semi-random sampling method as a
reliable way to balance feasibility and representative sampling in WelFur-Mink.

Animal welfare implications

On-farm welfare assessment can be used to determine animal welfare status of farms, eg for
certification of farms. The resulting assessments can also be used for advising farmers on what to
do in order to enhance animal welfare on their farm (Main et al 2003). However, all procedures must be highly feasible for welfare assessments to be useful in practice, and the time constraint is often a challenge, as shown for example in dairy cattle where the estimated time needed for a Welfare Quality® assessment in a herd with 200 cows is 7.7 hours (Kriirink and Windcller 2009).

From an economic point of view, it is also important to reduce the time needed for a full assessment (Sorensen et al 2007). In this study, a sampling method that satisfies these requirements is presented. This semi-random sampling method makes it possible to sample mink for welfare assessments according to the Welfur-Mink protocol. Also, the repeatability of the sampling is enhanced compared to the original method as the individual assessor is no longer responsible for selecting the cage sections. Whether samples selected with the semi-random sampling method have an acceptable representativeness depends on the purpose of sampling; however, we suggest that this method may be used in Welfur-Mink. The implementation of Welfur-Mink all European mink farms started in January 2017. The results of the welfare assessments are to be used for certification of farms and to provide decision support for the farmers. If all farms sign up for the assessment, the welfare of mink on approximately 3000 mink farms will be assessed annually in at least one of the three seasons. This feasible sampling method contributed making this large scale implementation possible, with the potential to improve the welfare of mink in Europe. The method can also be used for selecting representative samples of mink in other countries and for other purposes, as for example health inspection by vets or official inspection of compliance with rules and regulation by authorities. It may also be useful for sampling in other animal productions where animals within distinct units are more similar than between units. One example is fox farms which are often structured similarly to mink farms. Other examples include pigs and cattle where ‘the number of cages in use in each shed’ can be replaced by ‘the number of animals in each pen/section’.
Acknowledgements

We thank the owners of the private farm for kindly allowing us to use their farm as a model and
Tarja Koistinen, Eeva Ojala and Elina Mortensen for testing the first version of sampling method in
practice. All authors were responsible for development and revision of the method and contributed
to the manuscript. AFM was responsible for data collection, data management, simulated sampling
and generating the results. Fur Europe and a PhD grant from Aarhus University, Denmark financed
this project.

References

AWIN 2015 AWIN welfare assessment protocol for horses. DOI: 10.13130/AWIN_horses_2015

Fur Europe 2017 4,000 European fur farms to be Welfur certified by 2020. URL
Accessed 30th January 2018

Jørgensen G 1985 Mink Production. Scientifur

Knierim U and Winckler C 2009 On-farm welfare assessment in cattle: validity, reliability and
feasibility issues and future perspectives with special regard to the Welfare Quality® approach.
Animal Welfare 18: 451-458


Main DCJ, Kent JP, Wemelsfelder F, Ofler E and Tuyttens FAM 2003 Applications for


435
440 Wageningen Academic Publishers
441
442 Welfare Quality* 2009a Welfare Quality* assessment protocol for pigs. Welfare Quality*
443 Consortium: Lelystad, Netherlands
444
445 Welfare Quality* 2009b Welfare Quality* assessment protocol for cattle. Welfare Quality*
446 Consortium: Lelystad, Netherlands
Table 1: An example of the calculations that form the basis of the semi-random sampling method for on-farm welfare assessments according to the WeiFur-Mkn system.

<table>
<thead>
<tr>
<th>Shed order</th>
<th>Cages in use</th>
<th>Ratio</th>
<th>Cage sections to select</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>400</td>
<td>15.4</td>
<td>10</td>
</tr>
<tr>
<td>2</td>
<td>400</td>
<td>11.9</td>
<td>12</td>
</tr>
<tr>
<td>3</td>
<td>400</td>
<td>13.5</td>
<td>14</td>
</tr>
<tr>
<td>4</td>
<td>400</td>
<td>15.0</td>
<td>15</td>
</tr>
</tbody>
</table>

1 The total number of cages in use on the farm divided by the number of cage sections in the sample
2 A randomly picked shed number identifies which shed the starting one in the calculations and the rest of the sheds follow in sequential order
3 The number of cages in use is summarised for the stepwise addition of each shed and divided by the sample threshold
4 The ratio is rounded to its nearest integer and the number of cage sections that have been selected in the previous sheds is subtracted.
Table 2 The subgrouping factors and their levels included in the original stratification in WeFur-Mink (Adjusted from Müller et al. 2015).

<table>
<thead>
<tr>
<th>Factor</th>
<th>Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sex and age</td>
<td>Adult males</td>
</tr>
<tr>
<td></td>
<td>Adult females</td>
</tr>
<tr>
<td></td>
<td>Juveniles</td>
</tr>
<tr>
<td></td>
<td>Individually</td>
</tr>
<tr>
<td>Housing of adult dams</td>
<td>One of two juvenile males</td>
</tr>
<tr>
<td></td>
<td>With their litter</td>
</tr>
<tr>
<td></td>
<td>With other adults</td>
</tr>
<tr>
<td>Housing of juveniles</td>
<td>Male-female pairs</td>
</tr>
<tr>
<td></td>
<td>Other groups</td>
</tr>
<tr>
<td>Colour types</td>
<td>Brown</td>
</tr>
<tr>
<td></td>
<td>Mahogany</td>
</tr>
<tr>
<td></td>
<td>Black</td>
</tr>
<tr>
<td></td>
<td>Cross</td>
</tr>
<tr>
<td></td>
<td>Palomino</td>
</tr>
<tr>
<td></td>
<td>Pearl</td>
</tr>
<tr>
<td></td>
<td>Silver blue</td>
</tr>
<tr>
<td></td>
<td>White</td>
</tr>
<tr>
<td></td>
<td>Other/mixed</td>
</tr>
<tr>
<td>Cage type</td>
<td>Single/bair</td>
</tr>
<tr>
<td></td>
<td>Group</td>
</tr>
<tr>
<td>Cage wall</td>
<td>Solid</td>
</tr>
<tr>
<td></td>
<td>Wire mesh</td>
</tr>
<tr>
<td></td>
<td>Other</td>
</tr>
<tr>
<td>Nest box presence</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>No</td>
</tr>
<tr>
<td>Nest box position</td>
<td>Normal</td>
</tr>
<tr>
<td></td>
<td>Top</td>
</tr>
<tr>
<td>Nest box insulation</td>
<td>High</td>
</tr>
<tr>
<td></td>
<td>Medium</td>
</tr>
<tr>
<td></td>
<td>Low</td>
</tr>
<tr>
<td>Watering system</td>
<td>Automatic with frost protection</td>
</tr>
<tr>
<td></td>
<td>Automatic without frost protection</td>
</tr>
<tr>
<td></td>
<td>Manual</td>
</tr>
<tr>
<td>Shed type</td>
<td>Two row</td>
</tr>
<tr>
<td></td>
<td>Multi row</td>
</tr>
</tbody>
</table>
Table 3 The prevalence of mink on the model farm and the distribution of the prevalences in the 10,000 simulated samples for the factor levels related to sex, age, social housing, colour type and housing conditions.

<table>
<thead>
<tr>
<th>Factor</th>
<th>Level</th>
<th>Farm prevalence (%)</th>
<th>Error $^a$ (percentage points)</th>
<th>95% range $^b$ (percentage points)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sex and age</td>
<td>Adult females</td>
<td>13.9</td>
<td>0.0</td>
<td>19.7</td>
</tr>
<tr>
<td></td>
<td>Juveniles</td>
<td>81.1</td>
<td>0.0</td>
<td>19.7</td>
</tr>
<tr>
<td>Housing of adults</td>
<td>Individually</td>
<td>13.1</td>
<td>1.2</td>
<td>43.0</td>
</tr>
<tr>
<td></td>
<td>One or two juvenile males</td>
<td>86.9</td>
<td>-1.2</td>
<td>43.0</td>
</tr>
<tr>
<td></td>
<td>Individually</td>
<td>0.6</td>
<td>0.6</td>
<td>4.2</td>
</tr>
<tr>
<td>Housing of juveniles</td>
<td>Male-female pair</td>
<td>47.0</td>
<td>-0.6</td>
<td>33.9</td>
</tr>
<tr>
<td></td>
<td>Other groups</td>
<td>52.4</td>
<td>0.7</td>
<td>33.7</td>
</tr>
<tr>
<td>Colour types</td>
<td>Brown</td>
<td>30.6</td>
<td>-0.5</td>
<td>37.9</td>
</tr>
<tr>
<td></td>
<td>Cross</td>
<td>1.6</td>
<td>-1.6</td>
<td>11.8</td>
</tr>
<tr>
<td></td>
<td>Palomino</td>
<td>8.5</td>
<td>2.6</td>
<td>21.2</td>
</tr>
<tr>
<td></td>
<td>Pearl</td>
<td>10.5</td>
<td>0.3</td>
<td>27.8</td>
</tr>
<tr>
<td></td>
<td>Silver blue</td>
<td>15.4</td>
<td>-0.2</td>
<td>26.0</td>
</tr>
<tr>
<td></td>
<td>White</td>
<td>17.0</td>
<td>0.1</td>
<td>22.9</td>
</tr>
<tr>
<td></td>
<td>Other mixed</td>
<td>10.4</td>
<td>-1.8</td>
<td>27.3</td>
</tr>
<tr>
<td>Cage type</td>
<td>Single/pair</td>
<td>54.5</td>
<td>-0.3</td>
<td>27.7</td>
</tr>
<tr>
<td></td>
<td>Group</td>
<td>45.4</td>
<td>0.3</td>
<td>27.7</td>
</tr>
<tr>
<td>Cage well</td>
<td>Solid</td>
<td>48.1</td>
<td>-0.6</td>
<td>15.5</td>
</tr>
<tr>
<td></td>
<td>Wire mesh</td>
<td>50.9</td>
<td>0.6</td>
<td>15.6</td>
</tr>
<tr>
<td>Nest box position</td>
<td>Normal</td>
<td>98.0</td>
<td>-0.7</td>
<td>9.1</td>
</tr>
<tr>
<td></td>
<td>Top</td>
<td>2.0</td>
<td>0.7</td>
<td>9.1</td>
</tr>
<tr>
<td>Nest box insulation</td>
<td>High</td>
<td>60.5</td>
<td>0.1</td>
<td>27.5</td>
</tr>
<tr>
<td></td>
<td>Medium</td>
<td>30.9</td>
<td>-0.2</td>
<td>27.7</td>
</tr>
<tr>
<td></td>
<td>Low</td>
<td>0.6</td>
<td>0.6</td>
<td>5.6</td>
</tr>
<tr>
<td>Shed type</td>
<td>Two row</td>
<td>11.2</td>
<td>0.5</td>
<td>13.7</td>
</tr>
<tr>
<td></td>
<td>Multi row</td>
<td>88.8</td>
<td>-0.6</td>
<td>13.7</td>
</tr>
</tbody>
</table>

$^a$ The difference between the farm prevalence and the median of sample prevalences

$^b$ The difference between the 2.5 percentile and the 97.5 percentile of the sample prevalences
Table 4 The estimated probability of selecting a sample where the prevalence of mink are within ±5, 10, 15, 20 or 25 percentage points of the farm prevalence for the factor levels related to sex, age, social housing, colour types and housing conditions.

<table>
<thead>
<tr>
<th>Factor</th>
<th>Level</th>
<th>±5</th>
<th>±10</th>
<th>±15</th>
<th>±20</th>
<th>±25</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sex and age</td>
<td>Adult females</td>
<td>0.68</td>
<td>0.95</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>Juveniles</td>
<td>0.68</td>
<td>0.95</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Housing of adults</td>
<td>Individually</td>
<td>0.32</td>
<td>0.42</td>
<td>0.92</td>
<td>0.88</td>
<td>0.95</td>
</tr>
<tr>
<td></td>
<td>One or two juvenile males</td>
<td>0.32</td>
<td>0.42</td>
<td>0.92</td>
<td>0.88</td>
<td>0.95</td>
</tr>
<tr>
<td></td>
<td>Individually</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Housing of juveniles</td>
<td>Male-female pairs</td>
<td>0.42</td>
<td>0.74</td>
<td>0.92</td>
<td>0.88</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>Other groups</td>
<td>0.44</td>
<td>0.74</td>
<td>0.92</td>
<td>0.88</td>
<td>1.00</td>
</tr>
<tr>
<td>Colour types</td>
<td>Brown</td>
<td>0.37</td>
<td>0.67</td>
<td>0.88</td>
<td>0.77</td>
<td>0.99</td>
</tr>
<tr>
<td></td>
<td>Cross</td>
<td>0.83</td>
<td>0.97</td>
<td>0.99</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>Paromino</td>
<td>0.57</td>
<td>0.94</td>
<td>0.96</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>Pearl</td>
<td>0.65</td>
<td>0.74</td>
<td>0.96</td>
<td>0.99</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>Silver blue</td>
<td>0.51</td>
<td>0.67</td>
<td>0.96</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>White</td>
<td>0.57</td>
<td>0.89</td>
<td>0.96</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>Other/mixed</td>
<td>0.45</td>
<td>0.74</td>
<td>0.96</td>
<td>0.99</td>
<td>1.00</td>
</tr>
<tr>
<td>Cage type</td>
<td>Single pair</td>
<td>0.49</td>
<td>0.89</td>
<td>0.96</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>Group</td>
<td>0.49</td>
<td>0.89</td>
<td>0.96</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Cage wall</td>
<td>Solid</td>
<td>0.60</td>
<td>0.98</td>
<td>0.99</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>Wire mesh</td>
<td>0.80</td>
<td>0.98</td>
<td>0.99</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Nest box position</td>
<td>Normal</td>
<td>0.95</td>
<td>0.99</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>Top</td>
<td>0.95</td>
<td>0.99</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Nest box insulation</td>
<td>High</td>
<td>0.46</td>
<td>0.80</td>
<td>0.96</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>Medium</td>
<td>0.47</td>
<td>0.82</td>
<td>0.96</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>Low</td>
<td>0.98</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Shed type</td>
<td>Two row</td>
<td>0.96</td>
<td>0.97</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>Multi row</td>
<td>0.96</td>
<td>0.97</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Table 5 The estimated probability of selecting a sample where the prevalence of mink are within ±5, 10, 15, 20 or 25 percentage points of the farm prevalence for all factors in a group.

<table>
<thead>
<tr>
<th>Group of factors</th>
<th>±5</th>
<th>±10</th>
<th>±15</th>
<th>±20</th>
<th>±25</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sex, age and social housing</td>
<td>0.11</td>
<td>0.28</td>
<td>0.74</td>
<td>0.86</td>
<td>0.94</td>
</tr>
<tr>
<td>Colour types</td>
<td>0.02</td>
<td>0.29</td>
<td>0.75</td>
<td>0.93</td>
<td>0.99</td>
</tr>
<tr>
<td>Housing environment</td>
<td>0.16</td>
<td>0.65</td>
<td>0.92</td>
<td>0.99</td>
<td>1.00</td>
</tr>
<tr>
<td>All factors</td>
<td>0.00</td>
<td>0.07</td>
<td>0.53</td>
<td>0.80</td>
<td>0.93</td>
</tr>
</tbody>
</table>
4.2 Paper 2

Changes in the welfare of mink (*Neovison vison*) with date of assessment in the winter and growth periods have limited effects on the overall WelFur categorisation

Anna Feldberg Marsbøll, Britt Ingeborg Foseide Henriksen, Bente Krogh Hansen and Steen Henrik Møller

Submitted for publication as an original article in Animal Welfare
(Note: Included in the stage it was in when the thesis was submitted in September 2018. Since then, a revised version has been accepted for publication in Animal Welfare)
Changes in the welfare of mink (Neovison vison) with date of assessment in the winter and growth periods have limited effects on the overall WelFur categorisation

Anna Feldberg Marsboll1, Britt Ingeborg Foseide Henriksen1, Bente Krogh Hansen2 and Steen Henrik Møller1

1Department of Animal Science, Aarhus University, Blichers Alle 20, 8830 Tjele, Denmark
2Kopenhagen Fur, Agro Food Park 15, 8200 Aarhus N, Denmark

Corresponding author: Anna Feldberg Marsboll
E-mail: anna.marsboll@anis.au.dk

Running title: WelFur-Mink in the winter and growth periods
Abstract

In this study we investigated the robustness of the WelFur welfare assessment system for farmed mink (*Neovison vison*) to date of assessment in the winter and growth assessment periods. The prevalence of occurrences of some measurements was hypothesised to increase with date of assessment (too thin, fur chewing and stereotypic behaviour in the winter period, and injuries, diarrhoea and exploratory mink in the growth period). The welfare was assessed on eight Danish mink farms according to the WelFur-Mink protocol. Each farm was assessed once in the nursing period, four times in the growth period and three times in the winter period. WelFur scores were calculated based on the assessments in the three periods: one calculation for each assessment in the winter and growth periods. As expected, the odds of fur chewing increased with date of assessment in the winter period, and the odds of injuries, diarrhoea and exploratory mink increased with date of assessment in the growth period. The odds of too thin mink in the winter period decreased, i.e. the change was in the opposite direction of what was expected. The effect of these changes on the WelFur scores on the higher levels was limited, but could potentially lead to changes in the overall welfare categorisation if the principle scores were close to a threshold between two categories. A potential way to eliminate the effect of date of assessment could be to develop a correction factor for the measurements that can be expected to change within each assessment period.

Keywords

animal welfare, feasibility, mink production, reliability, welfare assessment, WelFur
Introduction

In any animal welfare assessment system, robustness to external factors, such as time of day and weather conditions, is important in order to ensure that the assessment is reliable and feasible. In this study, the robustness of the WelFur-Mink welfare assessment system for farmed mink (Neovison vison) to date of assessment is investigated.

WelFur-Mink was developed based on the concept of the EU project Welfare Quality® (Mononen et al 2012). An equivalent system was simultaneously developed for farmed foxes (blue fox (Vulpes lagopus) and silver fox (Vulpes vulpes) and their crossbreeds). In WelFur-Mink, the welfare is assessed at farm level based on a range of measurements taken on the farm (Moller et al 2015). All measurements are relevant to some aspect of the 12 animal welfare criteria that constitute the four principles for good animal welfare that were defined within the Welfare Quality® project (Botreau et al 2007). Due to the seasonal production system, a full WelFur-Mink assessment is based on three assessments: one in each of the three main production seasons. The on-farm assessment periods are the winter period from 1 January to 20 February, the nursing period from 5 May to 1 July and the growth period from 23 September to 30 November. In practice, the beginning and end of each assessment period are affected by the onset of events, such as flush feeding in the winter period, average date of birth and onset of weaning in the nursing period, and onset of sorting and pelting in the growth period. A representative sample of the mink on the farm in each assessment period is used for the assessment of all measurements, except a few mainly management-based measurements taken at farm level. After the three assessments, each farm is assigned one of four welfare categories. The categorisation is based on a transformation of each measurement result into a standardised score on a scale from 0-100 followed by a stepwise aggregation of the measurement scores across the three periods into 12 criteria and four principle scores (Moller et al 2015).
The assessments take place at times in each season where the potential welfare problems are most likely to be observed and each assessment period is only six to eight weeks long in order to limit the possible variation in welfare within each period. However, the prevalence of welfare problems may still vary with date of assessment as shown in the nursing period for several measurements with increasing welfare problems closer to weaning (Henriksen and Møller 2015). This variation in welfare with date of assessment can affect the overall categorisation of the farms in case the farms’ principle scores are close to a threshold value for another category (Henriksen 2015). The welfare can also be expected to show some variation with date of assessment in the winter and growth periods. During the winter, breeders are slimmed in order to facilitate reproduction. If the slimming is not managed carefully, there is a risk that the mink will experience hunger as well as become too thin. As a consequence of the unfulfilled feeding motivation, the mink may start performing stereotypic behaviour (Damgaard et al 2004) and potentially also fur chewing (Malmbjörk et al 2013). In the growth period, mink are fed ad libitum or close to ad libitum, and most mink will become heavy or obese. Sustained over-feeding may result in an increased risk of diarrhoea (Hansen 1985). There is also an increasing risk of developing fatty liver (Hunter 1996), which in practice has been associated with abnormal faeces. Also, as the juvenile mink become more territorial as winter approaches, there is an increasing risk of aggression resulting in injuries, especially among group-housed mink (Hansen et al 2014). Juveniles’ response in a temperament test has also been found to change during the growth period with more confident and exploratory mink as peeling time approaches (Hansen 1996, Malmbjörk and Hansen 2001).

The aim of this study was to explore whether there is a change in welfare with date of assessment in the winter and growth assessment periods that needs to be taken into consideration in the WelFur
Results - Paper 2

97 assessment of min. We hypothesised that the prevalence of too thin mink, fur chewing and
98 stereotypic behaviour increases with date of assessment in the winter period and that the prevalence
99 of injuries, diarrhoea and exploratory mink increases with date of assessment in the growth period.
100 The potential changes at measurement level may not necessarily affect the welfare scores at
101 measurement, criteria, principal and overall category level. Firstly, because the transformation into
102 scores is not linear. Secondly, because the stepwise aggregation of the scores may somewhat
103 diminish the effect of changes at measurement level. Thirdly, because the overall categories are
104 broad, and the final categorisation will change only if the principle scores are close to a threshold
105 between two categories.
106
107 Materials and methods
108
109 Study design
110 Eight commercial Danish mink farms, which varied in geographical location, size, feed suppliers,
111 housing conditions and combinations of colour types, were included in the study. Each farm was
112 assessed four times during the growth period in 2014 and three times during the winter period in
113 2015 according to the WelFur-Mink protocol (Moller et al 2015). The farms were also assessed
114 once in the nursing period in 2014 to have data for the calculation of WelFur scores across three
115 periods for each farm. Data collection at farm level was only conducted at one of the assessments in
116 each period, while data collection at animal/cage level was conducted at all assessments. For each
117 farm and assessment period, a stratified sample was taken in order to represent the farm in relation
118 to sex, age, colour type and housing conditions as described in Moller et al 2015. The same sample
119 was assessed repeatedly on each farm in each period. Four different assessors, two in each
120 assessment period, were involved in the study. At each assessment, one assessor did the evaluation,
121 while an assistant recorded the outcome. In the beginning of each assessment period, the two
assessors calibrated by assessing one farm together. Situations with doubt about the assessment were noted, photographed and used for continuous calibration. At some assessments, the assessors were each other's assistants, thus further increasing the calibration between assessors. Both the assessor and assistant observed stereotypic behaviour (each observed approximately half the cages in the sample). Stereotypic behaviour was observed either 1 hour before the farms usual (expected) time of feeding or as the last observation of the day, following the WelFur-Mink assessment protocol at the time. On the individual farms, stereotypic behaviour was observed at the same time within each assessment period.

Sample size

The samples consisted of 120 cages with one mink per cage in the winter period and 90 cages with one to four mink per cage in the growth period. Cages where the number of mink changed during the periods were excluded in order to ensure that the samples consisted of the same mink at all assessments. Cages where the mink were moved to another location were also excluded in order to avoid potential effects of changes in the environment to interfere with the effects of time. In the winter period, 47 cages were excluded. Only one or two cages were excluded from each farm, except two farms where seven and 32 cages were excluded, mainly because the mink were moved to a new location. The resulting sample size ranged from 88 to 120 cages per farm with an average of 114 cages. In the growth period, 35 cages, one to ten cages per farm, were excluded because the number of mink in these cages changed within the period. The resulting sample size ranged from 81 to 89 cages per farm with an average of 86 cages. Missing values for individual measurements did not exclude the respective cage from the study but reduced the sample size for the actual measurement.
variable and invariable measurements

The evaluation of all animal-based measurements may vary within the assessment periods. The only exception is mortality which is evaluated at farm level for each season. The management-based measurements such as ‘Social housing’ and ‘Frequency and duration of handling and transportation’ are also evaluated at farm level for the entire seasons and cannot vary within the assessment periods. As this study was limited to mink kept in the same cages throughout each assessment period, the resource-based measurements ‘Type of watering system’ and ‘Protection from exceptional weather conditions’ and the sub-measurement ‘Nest box insulation capacity’ (a part of the measurement ‘Nest box material and bedding/nesting material’) will not vary. Hence, these measurements were standardised to the first assessment in each assessment period. The evaluation of the remaining resource-based measurements may vary within the assessment periods. For example, some cages consisted of several compartments, and sometimes the number of available compartments was changed thus affecting the evaluation of the measurement ‘Space available for moving (cage area and cage height)’. Also cage enrichments and bedding material can be added, removed or used up by the mink and thereby change the evaluation of the measurements ‘Cage enrichment’ and the sub-measurement ‘Bedding material’ (a part of the measurement ‘Nest box material and bedding/nesting material’). Finally, the evaluation of the sub-measurement ‘Protection from draft’ (a part of the measurement ‘Nest box material and bedding/nesting material’) and the measurement ‘Functioning and cleanliness of the water points’ may vary within the assessment periods.

Scoring and aggregations

Welfur-Mink uses 22 different measurements in the evaluation of animal welfare at farm level (Table 1). Some measurements consist of several sub-measurements, as for example the
measurement ‘Protection from exceptional weather conditions’ which is based on an evaluation of the protection from sun and wind and the possibility of cooling. Most measurements are assessed in all periods (e.g. ‘Obviously sick animals’) while other are period specific (e.g. ‘Age and procedures at weaning’). The result of each measurement taken on the farm in each assessment period (e.g. percentage without access to a nest box) is first transformed into a measurement score for each period on a standardised scale ranging from 0 (worst) to 100 (best) and aggregated across the three assessment periods into 22 measurement scores. The 22 measurement scores are then aggregated into 12 scores at criteria level and further into four scores at principle level (Møller et al 2015).

Table 1 gives an overview of which measurements that are aggregated to which criteria and which criteria that are aggregated to which principles. All aggregations use Choquet integrals (Møller et al 2015). This means that when aggregating a number of scores, the lowest score is the starting point, and this score is partly compensated by the higher scores, depending on the assigned weights. The transformation of the measurements taken on the farm in each period into a measurement score for each period, and the weights used when going from the measurement scores for each period to measurement scores across the three assessment periods to criteria scores, were derived from experts in fur animal welfare. The weights used when going from criteria scores to principle scores were derived from the weights used in Welfare Quality®. Finally, each farm is assigned one of four welfare categories (‘Best current practice’, ‘Good current practice’, ‘Acceptable current practice’, or ‘Unacceptable current practice’) based on the four principle scores, depending on the threshold values described in Table 2. These are the same as in Welfare Quality® (Møller et al 2015).

Data analysis

The aim of our study was to evaluate the effect of date of assessment on the resulting welfare assessment. Date of assessment is defined as ‘Day in assessment period’, i.e. assessments on 9
January are day 9 in the winter period. Observations in the winter period were initiated on 6 January (day 6) and completed on 19 February (day 50; 1 day before the end of the assessment period). Observations in the growth period were initiated on 22 September (day 0; one day before the start of the assessment period) and completed on 11 November (day 50). The end of the growth period varies between farms according to their practices regarding sorting before pelting. However, in practice, there will only be a few assessments after 15 November. For the analysis, the winter period was divided into three sub-periods and the growth period into four sub-periods (Table 3). The difference with sub-period in the assessed welfare was evaluated individually for the winter and growth periods and was investigated both at measurement level and for the calculated WelFur scores.

The outcome variables included in the analysis at measurement level include ‘Too thin’, ‘Fur chewing’ and ‘Stereotypic behaviour’ in the winter period, and ‘Injuries’, ‘Diarrhoea’ and ‘Exploratory’ in the growth period (Table 4). The variables were included as binary variables as in Henriksen and Møller (2015). Regarding ‘Injuries’ and ‘Fur chewing’, it should be noted that the WelFur-Mink system also takes the severity of the injury or fur chewing into account. However, due to the relative low prevalence of mink with injuries and fur chewing in the higher categories, it was not possible to analyse each category separately. The effect of date of assessment on the measurement variables in each period was analysed using binomial mixed models with a logit link function with sub-period as explanatory variable. Assessor was included as a fixed effect, whereas farm, cage nested in farm (due to repeated measurements on cage) and assessment number nested in farm (in order to take variance heterogeneity between assessments into account) were included as random effects. Assessor was removed from the model when the effect was insignificant.
In order to investigate how the potential changes with date of assessment at measurement level in the winter and growth periods affect the measurement scores in each period, criterion scores, principle scores and the overall classification, we calculated WellFur-Mink scores for each assessment on each farm, using the actual values for the measurements that we hypothesised to change. This was done for one assessment period at a time. When investigating the changes in the winter period, the calculations included the actual values for each assessment for the measurements ‘Body condition score’ (percentage too thin), ‘Fur chewing’ (a combination of the percentage in each category) and ‘Stereotypic behaviour’ (percentage stereotypic) in the winter period. When investigating the changes in the growth period, the calculations included the actual value for each assessment for the measurements ‘Skin lesions or injuries to the body’ (a combination of the percentage of mink with injuries in each categories), ‘Diarrhoea’ (percent cages with signs of diarrhoea) and ‘Temperament test’ (a combination of the percentage of mink categorised as ‘Exploratory’, ‘Fearful’, ‘Aggressive’ and ‘Undecided’) in the growth period. The remaining variable measurements in the winter and growth periods were set to an average over the assessments for each farm in each period, and the invariable measurements were included with the actual values for each farm in each period. All measurements taken in the nursing period were included with the actual values for each farm. In order to investigate whether the potential changes in the WellFur-Mink scores are affected by changes in other measurements, we repeated the above calculation but included all variable measurements in the winter and growth periods, respectively, with their actual values. The distribution of the calculated scores was approximately normal in both sets of calculations. The effect of date of assessment on the calculated scores was analysed using mixed models with sub-period as explanatory variable, assessor as a fixed effect and farm as a random effect. Assessor was removed from the model when the effect was insignificant.
The calculations of the scores were done using the Welfur-Mink calculation tool developed by the French National Institute of Agronomic Research (INRA). The statistical analyses were performed with the package ‘lme4’ (Bates et al. 2015) of the statistical software R (R Core Team 2017).

Pairwise comparisons were performed with the package ‘multcomp’, and the p-values were adjusted according to the Bonferroni method (Hothorn et al. 2008). The limit for statistical significance was set to 5%, i.e. only effects where the p-value of the test was less than 0.05 were declared statistically significant. All graphical presentations were made in Microsoft Excel.

Results

Changes with date of assessment in the winter period

The prevalence of occurrences for the outcome variables that we hypothesised would change in the winter period is shown in Figure 1. Figure 1a indicates a decrease in the prevalence of ‘Too thin’ mink with assessment date, at least for some of the farms. The analysis confirmed this, as there was a significant change with sub-period in the odds of ‘Too thin’ mink ($\chi^2 = 31.1, \text{df} = 2, P < 0.001$) where the odds in sub-period 3 were 0.002 (95% CI [0.0001; 0.03]) times lower than in sub-period 1 and 0.01 (95% CI [0.001; 0.2]) times lower than in sub-period 2. Figure 1b indicates a slight increase in the prevalence of ‘Fur chewing’. The analysis confirmed this, as there was a significant change with sub-period in the odds of mink with ‘Fur chewing’ ($\chi^2 = 30.1, \text{df} = 2, P < 0.001$) where the odds in sub-periods 2 and 3 were 4.5 (95% CI [2.3; 8.8]) and 6.4 (95% CI [3.3; 12.6]) times higher than in sub-period 1. Figure 1c indicates that the prevalence of ‘Stereotypic behaviour’ in general does not change with date of assessment, as some farms increase while others decrease. The analysis confirmed this, as there was no change with sub-period in the odds of ‘Stereotypic behaviour’ ($\chi^2 = 6.0, \text{df} = 2, P = 0.05$).
The results from the transformation of the outcome variables that we hypothesised would change in
the winter period into the measurement scores for the winter period are shown in Figure 2. Figure
2a shows a high variation between and within farms for the score for ‘Body condition score’ in the
winter period. Overall, it seems that the scores increase with assessment date. The analysis did not
confirm this, as there was no change with sub-period ($\chi^2 = 5.8, df = 2, P = 0.06$). Figure 2b
indicates a gradual decrease in the score for ‘Fur chewing’ in the winter period with date of
assessment. The analysis confirmed this, as there was a significant change with sub-period ($\chi^2 =
21.2, df = 2, P < 0.001$) where the scores in sub-periods 2 and 3 were 8.0 points (95% CI [-12.7:
3.4]) and 13.0 points (95% CI [-18.0; -8.0]) lower than in sub-period 1. Also, the score in sub-period
3 was 5.0 points (95% CI [-9.6; -0.3]) lower than in sub-period 2. Figure 2c shows a high variation
between and within farms for the score for ‘Stereotypic behaviour’ in the winter period but no
general trend. The analysis showed no change in the score with sub-period ($\chi^2 = 5.6, df = 2, P =
0.06$).

The criteria scores calculated when the measurements ‘Body condition score’, ‘Fur chewing’ and
‘Stereotypic behaviour’ were included with their actual values at different assessment days the in
the winter period are shown in Figure 3. Figure 3a indicates that there are no general or only small
changes in the scores for the criteria ‘Absence of prolonged hunger’ with date of assessment, except
for one farm with an increase of approximately 30 points from the second to the third assessment.
However, the analysis showed no change in the score with sub-period ($\chi^2 = 2.8, df = 2, P = 0.3$).
Figure 3b indicates a small but general decrease in the score for the criteria ‘Expression of other
behaviours’ with assessment date. The analysis confirmed this, as there was a significant change
with sub-period ($\chi^2 = 12.5, df = 2, P = 0.002$) where the scores in sub-periods 2 and 3 were 4.0
points (95% CI [-6.5; -0.9]) and 4.5 points (95% CI [-7.0; -1.1]) lower than in sub-period 1.
The principle scores calculated when the measurements ‘Body condition score’, ‘Fur chewing’ and ‘Stereotypic behaviour’ were included with their actual values at different assessment days the in the winter period are shown in Figure 4. Figure 4a indicates that there are no or only small changes in the scores for the principle of ‘Good feeding’ with date of assessment, except for one farm where there was an increase of approximately 20 points from the second to the third assessment. The analysis showed no change with sub-period ($\chi^2 = 2.6, df = 2, P = 0.3$). Figure 4b indicates that there is no change in the score for the principle of ‘Appropriate behaviour’ with assessment date. The analysis, however, showed that there was a change with sub-period ($\chi^2 = 9.6, df = 2, P = 0.01$) where the score in sub-period 2 was 1.0 point (95% CI [-1.6; -0.3]) lower than in sub-period 1.

The overall categorisation of the farms was not affected by the changes in the principal scores.

Seven farms were classified as ‘Good current practice’ and one farm as ‘Best current practice’ at all assessments.

Analysing the WellFar scores that were calculated with the actual values for all variable measurements in the winter period showed that also the score for the measurement ‘Cage enrichment’ changed with sub-period ($\chi^2 = 10.6, df = 2, P = 0.005$), as the score in sub-period 3 was 9.0 points (95% CI [3.0; 15.0]) higher than in sub-period 2. The aggregated score for the criteria ‘Expression of other behaviours’ still decreased with sub-period ($\chi^2 = 10.6, df = 2, P = 0.005$), and the scores in sub-periods 2 (95% CI [-6.9; -1.1]) and 3 (95% CI [-7.1; -0.9]) were 4.0 points lower than in sub-period 1. There were no changes in any other criteria or principle scores, and the overall categorisation of the farms was the same as described above.
Changes with date of assessment in the growth period

The prevalence of occurrences for the outcome variables that we hypothesised would change in the growth period is shown in Figure 5. Figure 5a indicates a small increase in the prevalence of ‘Injuries’ with date of assessment. The analysis confirmed this, as there was a change with sub-period in the odds of mink with ‘Injuries’ ($\chi^2 = 9.0, \text{df} = 3, P = 0.03$) where the odds were 2.1 (95% CI[1.3, 3.8]) times higher in sub-period 4 compared to sub-period 1. Figure 5b indicates that the prevalence of ‘Diarrhoea’ is higher in the end of the assessment period. This was confirmed by the analysis, as there was a change with sub-period in the odds of mink with ‘Diarrhoea’ ($\chi^2 = 17.6, \text{df} = 3, P < 0.001$) where the odds in sub-period 4 were 11.8 (95% CI[2.2, 62.4]), 5.1 (95% CI[1.2, 21.5]) and 27.9 (95% CI[27.2, 291.9]) times higher than in sub-periods 1, 2 and 3. Figure 5c indicates a gradual increase in the prevalence of ‘Exploratory’ mink with date of assessment. As the prevalence of ‘Exploratory’ mink increased, the prevalence of both ‘Fearful’ and ‘Undecided’ mink decreased, while there were no mink categorised as ‘Aggressive’. The increase in the prevalence of ‘Exploratory’ mink was confirmed by the analysis, as there was a change with sub-period ($\chi^2 = 31.1, \text{df} = 3, P < 0.001$) where the odds in sub-periods 2, 3 and 4 were 1.8 (95% CI[1.3, 2.5]), 3.2 (95% CI[2.1, 4.8]) and 2.2 (95% CI[1.5, 3.1]) times higher than in sub-period 1. There was also a difference between sub-periods 2 and 3, where the odds of ‘Exploratory’ mink in sub-period 3 were 1.8 (95% CI[1.2, 2.6]) times higher than in sub-period 2.

The results from the transformation of the outcome variables that we hypothesised would change in the growth period into the measurement scores for the growth period are shown in Figure 6. Figure 6a indicates that there are no or only a small change with date of assessment in the score for ‘Skin lesions or injuries to the body’ in the growth period. The analysis showed that there was no change with sub-period in the score for ‘Skin lesions or injuries to the body’ in the growth period ($\chi^2 = 5.4$, 95% CI[0.8, 0.9]) times higher than in sub-period 1. There was also a difference between sub-periods 2 and 3, where the odds of ‘Exploratory’ mink in sub-period 3 were 1.8 (95% CI[1.2, 2.6]) times higher than in sub-period 2.
Results - Paper 2

$df = 3, P = 0.1$). Figure 6b shows a high variation between and within farms for the score for

‘Diarrhoea’ in the growth period. The analysis showed that there was a change with sub-period in

the score for ‘Diarrhoea’ in the growth period ($\chi^2 = 15.4$, $df = 3$, $P = 0.002$) where the score in sub-

period 4 was 38.6 points (95% CI [-65.7; -11.4]) lower than in sub-period 1 and 41.5 points (95% CI

[-71.9; -11.2]) lower than in sub-period 3. Figure 6c indicates that the score for ‘Temperament test’

in the growth period is lower in the first part of the assessment period. This was confirmed by the

analysis, as there was a change with sub-period in the score for ‘Temperament test’ ($\chi^2 = 43.8$, $df =

3$, $P < 0.001$) where the scores in sub-periods 2, 3 and 4 were 12.1 points (95% CI [7.9; 16.2]), 16.4

points (95% CI [11.7; 21.2]) and 14.3 points (95% CI [10.9; 18.7]) lower than in sub-period 1.

The criteria scores calculated when the measurements ‘Skin lesions or injuries to the body’,

‘Diarrhoea’ and ‘Temperament test’ were included with their actual values at the different

assessment days in the growth period are shown in Figure 7. Figure 7a indicates that there is no

change in the score for the criteria ‘Absence of injuries’ with date of assessment. This was

confirmed by the analysis as there was no change with sub-period ($\chi^2 = 4.7$, $df = 3$, $P = 0.2$). Figure

7b indicates that the score for the criteria ‘Absence of disease’ may decrease with date of

assessment. This was confirmed by the analysis as there was a change with sub-period ($\chi^2 = 17.1$,

$df = 3$, $P < 0.001$) where the scores in sub-period 4 were 6.5 points (95% CI [-10.9; -2.0]) lower than

in sub-period 1, 4.9 points (95% CI [-9.1; -0.7]) lower than in sub-period 2 and 7.8 points (95% CI [-

12.6; -2.9]) lower than in sub-period 3. Figure 7c indicates that the score for the criteria ‘Good

human-animal relationship’/‘Positive emotional state’ is lower in the beginning of the assessment

period. This was confirmed by the analysis, as there was a change with sub-period ($\chi^2 = 36.3$, $df =

3$, $P < 0.001$) where the scores in sub-periods 2, 3 and 4 were 6.0 points (95% CI [3.4; 7.1]), 7.5

points (95% CI [4.8; 9.0]) and 6.7 points (95% CI [4.2; 8.1]) lower than in sub-period 1.
The principle scores calculated when the measurements ‘Skin lesions or injuries to the body’, ‘Diarrhoea’ and ‘Temperament test’ were included with their actual values at the different assessment days in the growth period are shown in Figure 8. Figure 8a shows a large but quite stable variation between the farms in the score for the principle ‘Good health’. Even so, the analysis showed that there was a change with sub-period (χ² = 16.5, df = 3, P < 0.001) where the scores in sub-period 4 were 4.7 points (95% CI [-7.8, -1.5]) lower than in sub-period 1, 3.4 points (95% CI [-6.4, -0.5]) lower than in sub-period 2 and 5.2 points (95% CI [-8.6, -1.7]) lower than in sub-period 3. Figure 8b indicates a slight increase with date of assessment in the score for the principle ‘Appropriate behaviour’. The analysis confirmed this, as there was a change with sub-period (χ² = 36.1, df = 3, P < 0.001) where the scores in sub-periods 2, 3 and 4 were 3.0 points (95% CI [1.8, 4.3]), 4.1 points (95% CI [2.6, 5.5]) and 3.7 points (95% CI [2.4, 5.0]) higher than in sub-period 1.

For one farm, the overall categorisation was affected when the measurements ‘Skin lesions or injuries to the body’, ‘Diarrhoea’ and ‘Temperament test’ were included with their actual values at the different assessment days in the growth period. The change in category was determined by changes in the principle of ‘Good health’ where the value was either above or below the threshold of 80 points, while all the principle scores for ‘Good housing’ were above 80 and ‘Good feeding’ and ‘Appropriate behaviour’ were between 55 and 80. When the actual values from the first and fourth assessments in the growth period were used (principle scores of 78.6 and 76.1, respectively), the farm was categorised as ‘Good current practice’. However, when the values for the second and third assessments were used (principle scores of 80.1 and 85.2, respectively), the farm was categorised as ‘Best current practice’. The remaining seven farms were in the same category independent of the
changes with assessment date in the growth period. Six farms were categorised as ‘Good current
practice’ and one farm as ‘Best current practice’.

Analysing the WellFar scores that were calculated with the actual values for all variable
measurements in the growth period revealed a change in the measurement score for ‘Nest box
material and bedding/nesting material’ (χ² = 9.5, df = 3, P = 0.02) where the score in sub-period 4
was 19.0 points (95% CI [2.7; 35.2]) higher than in sub period 1. Also, the criterion score for
‘Thermal comfort’ was found to change with sub-period (χ² = 10.3, df = 3, P = 0.02) where the
score in sub-period 4 was 2.8 points (95% CI [0.5; 5.2]) higher than in sub-period 1. There were still
changes with sub-period for the criteria ‘Absence of disease’ (χ² = 16.3, df = 3, P = 0.001) and
‘Good human-animal relationship’/‘Positive emotional state’ (χ² = 36.3, df = 3, P < 0.001) and the
principles ‘Good Health’ (χ² = 15.6, df = 3, P = 0.001) and ‘Appropriate behaviour’ (χ² = 34.0, df =
3, P < 0.001). These changes were almost identical to the changes when only including the actual
values for the measurements that we expected to change. There were no changes in any other
criteria or principle scores, and the overall categorisation of the farms was the same as described
above.

Discussion

The results support the hypothesis that the prevalence of mink with fur chewing increases with date
of assessment in the winter period and that the prevalence of mink with injuries, cages with
diarrhoea and exploratory mink increases with date of assessment in the growth period.

Unexpectedly, there was no change in the prevalence of stereotypic behaviour in the winter period,
and the prevalence of too thin mink in the winter period decreased. As expected, only the largest
changes affected the WellFar scores at measurement, criteria and principle level and can potentially
Results - Paper 2

Winter period

There was a change with sub-period in the odds of too thin mink in the winter period. However, the change was in the opposite direction of what we hypothesised, as the odds in sub-period 3 were lower than in sub-period 1 and 2. The difference in the categorisation of too thin mink in January and February may be the reason for this. In January, mink in body condition 1 and 2 are considered too thin, while in February only mink in body condition 1 are considered too thin (Table 3). The reason for this categorisation is the objective of the criteria, is to identify mink exposed to prolonged hunger. On most farms, all mink are fed ad libitum during the growth season in order to produce large pelts. As a result, the majority of the mink will be in body condition 4 or 5 by the end of the growth season. Based on grading of pelt quality and evaluation of body size in November, the best animals are selected as the following years' breeders. Thereafter, slimming of the selected breeding animals before mating can begin. If slimming is done too drastically, it is associated with prolonged hunger. Hence, in WelFur-Mink, mink slimmed to a body condition 2 or less in January are associated with prolonged hunger. In February, there has been a longer period to slim down the mink, and only mink slimmed to a body condition 1 are associated with prolonged hunger (Moller et al. 2015). This means, that even if the odds of mink in body condition 1 and 2 increase throughout the winter assessment period, the odds of too thin mink may decrease, as only body condition 1 is associated with prolonged hunger in February. A more gradual categorisation of body condition could help reduce the change in the odds of too thin mink with sub-period. The change with sub-period in the odds of too thin mink was not apparent when the prevalence of too thin mink was transformed into a measurement score for the winter period. This may be caused by the increased variation in the data caused by the non-linear transformation where the score for this measurement...
is reduced dramatically for a prevalence of up to approximately 10% and less thereafter (Moller et al. 2015). This highlights the importance of an accurate assessment of the prevalence of too thin mink, as even small changes may have a large impact on the score.

As hypothesised, the odds of far chewing increased with sub-period with higher odds in sub-periods 2 and 3 compared to sub-period 1. This change affected the WelFur scores at all aggregated levels. However, the actual effect on the scores was relatively low at the higher levels. At principle level (‘Appropriate behaviour’), only the score based on the prevalence from sub-period 2 was lower than sub-period 1, and the estimated difference was only 1 point. The assigned weights increase the diminishing effect of the aggregations, as the measurement score for ‘Fur chewing’ in the winter period and its related scores at the higher levels are weighed less than the scores they are aggregated with (Moller et al. 2015). It differs whether the measurement score for ‘Fur chewing’ in the winter period and its related scores at the higher levels have the highest or lowest score in the aggregations, thus, the effect on calculated scores differs.

Contrary to what was hypothesised, there was no change in the odds of stereotypic mink in the winter period. One reason could be that stereotypic behaviour was observed before feeding or, if this was not possible, as the last observation of the day. As ‘the last observation of the day’ can vary considerably, this has now been changed to ‘from 1.5 hours before sunset’ in order to increase the reliability of the assessment. Furthermore, a correction factor to adjust for difference in the prevalence of stereotypic behaviour ‘before feeding’ and ‘before sunset’ is under development.

Thus, the results regarding stereotypic behaviour in this study are, therefore, not necessarily comparable to how the assessment is carried out today. Hansen et al. (2009) found that the prevalence of stereotypic behaviour observed during postponed feeding increased during January.
and February in a study where the mink were fed restrictively from December and the weight gradually decreased until the end of February. The eight farms included in the present study may have used different management strategies when slimming the mink, for example the length of the slimming period and the degree of weight loss, which may have affected the mink’s feeding motivation and, hence, the development and display of stereotypic behaviour. Studies comparing restrictive feeding to ad libitum feeding in the winter period have found that restrictive feeding increases the prevalence of stereotypic behaviour (Damgaard et al. 2004, Houbak and Møller 2000). To our knowledge, there are no studies on how restrictive feeding is managed in practice and how differences in management affect the development and display of stereotypic behaviour.

When including the actual values for all variable measurements in the winter period, the score for ‘Cage enrichment’ in the winter period increased with sub-period. The measurement score for ‘Cage enrichment’ is aggregated with the measurement scores ‘Fur chewing’ and ‘Stereotypic behaviour’ into the criteria score ‘Expression of other behaviours’. However, the score for the criterion ‘Expression of other behaviours’ still decreased with sub-period. This is probably due to the aggregations where the lowest score is the starting point, and a low score can only partly be compensated by higher scores in another area. In this case, the score for ‘Cage enrichment’ was higher than the scores for ‘Fur chewing’ and ‘Stereotypic behaviour’, thus, the score for ‘Cage enrichments’ has the least impact on the aggregated score. The decreasing score for ‘Fur chewing’ with sub-period, therefore, has a larger impact than the increasing score for ‘Cage enrichment’. However, the changes in the principle score ‘Appropriate behaviour’ were no longer apparent. This indicates that variation in other variable measurements masks some of the observed changes.
Results - Paper 2

Growth period

The odds of mink with injuries in sub-period 4 were increased compared to sub-period 1, which supports our hypothesis. However, this change was not apparent when transformed into a measurement score for the growth period. The reason for this may be that this calculation also takes the severity of the injury into account; hence, an increased prevalence of the total number of injuries may not lead to an increase score for injuries if the severity is lower.

As hypothesised, the odds of cages with signs of diarrhoea increased with sub-period and were higher in sub-period 4 than in the first three sub-periods. The variation increased when the prevalence of cages with diarrhoea was transformed into a measurement score for the growth period due to the non-linear transformation where the score of welfare for this measurement is reduced dramatically for a prevalence of up to approximately 8% and less thereafter (Moller et al 2015).

However, the Welfar scores at all aggregated levels were still affected by the change in the odds of cages with signs of diarrhoea, but the effect was relatively low at the higher levels. At principle level (‘Good health’), the estimated score based on the prevalence from sub-period 4 was only decreased by 3 to 5 points compared to sub-periods 1, 2 and 3, while the scores on measurement level were decreased by 38 to 41 points. The assigned weights increase the diminishing effect of the aggregations as the measurement score for ‘Diarrhoea’ is not given the highest weight in the aggregation into the criteria ‘Absence of disease’ (Moller et al 2015). It differs whether the measurement score for ‘Diarrhoea’ in the growth period and its related scores at the higher levels have the highest or lowest value in the aggregations, thus, the effect on calculated scores differs.

The odds of exploratory mink increased with sub-period, thereby supporting the hypothesis. The prevalence of exploratory mink is aggregated with the prevalence of fearful and
aggressive or undecided mink into the measurement score for “Temperament test”. The increased odds
of exploratory mink (and thereby decreased odds of fearful or undecided mink, as there were no
aggressive mink) affected the WelFur scores at all levels but with a decreasing effect at the higher
levels. At principle level (“Appropriate behaviour”), the score based on the prevalence from sub-
periods 2, 3 and 4 was only increased 3 to 4 points compared to sub-period 1, compared to an
increase of 12 to 16 points at measurement level. The assigned weights reduce the diminishing
effect of the aggregations, as the weight given to the measurement score for “Temperament test” in
the growth period and its related scores at the higher levels is quite high, thereby adding to maintain
the effect of the changes throughout the aggregations (Møller et al 2015). It differs whether the
measurement score for “Temperament test” in the growth period and its related scores at the higher
levels have the highest or lowest value in the aggregations, thus, the effect on calculated scores
differs. Due to the repeated testing of the mink throughout the period, there is a risk that the mink
habituated to the test situation, thus increasing the prevalence of exploratory mink. Malmkvist and
Hansson (2002) suggested that the reduced fear response of mink they found when testing juvenile
mink weekly during the growth season in different fear tests was due to habituation to the test
situations, even though the tests were not identical. In the present study, the majority of the tested
mink were juveniles. However, due to the repeated testing it is not possible to separate the effect of
habituation and the possible effect of maturation of the juvenile mink. This difference should be
investigated further, as only an effect of the latter is a reliability issue in the WelFur system.

When including the actual values for all variable measurements in the growth period, the changes
found in the WelFur scores, when only varying the measurements that were expected to change,
were still present. This means that the changes in these measurements are not masked by variation
in other measurements. The measurement “Nest box material and bedding/nesting material” was
also found to change with sub-period with a higher score in sub-period 4 than in sub-period 1.

Management may have caused this measurement score to change, as the mink’s access to straw is
often increased by the end of the growth period. This change also affected the aggregated criteria
score for ‘Thermal comfort’ but not the principle score for ‘Good housing’. The assigned weights
increase the diminishing effect of the aggregations, as the criteria score for ‘Thermal comfort’ is
given the lowest weight in the aggregation into the principle score for ‘Good housing’ (Møller et al
2015). It differs whether the criteria score for ‘Thermal comfort’ has the highest or lowest value in
the aggregation into the principle score for ‘Good housing’, thus, the effect on calculated scores
differs.

Changes in overall evaluation

There were changes in the prevalence of welfare problems with date of assessment that could
potentially lead to changes in the overall welfare categorisation. Some of the changes were positive
in regards to animal welfare (decreased odds of too thin mink and increased odds of exploratory
mink) and others were negative (increased odds of cages with diarrhoea, mink with injuries and fur
chewing). However, the resulting changes in the principle scores were limited. Later date of
assessment in the winter period only led to a small change in the principle ‘Appropriate behaviour’
with a reduction of only 1 point in sub-period 2 compared to sub-period 1. Also, this change was
not apparent when all variable measurements were included with their actual values. Later date of
assessment in the growth period led to a decrease in the principle score for ‘Good health’ where the
score in sub-period 4 was 3 to 5 points lower compared to the other periods and an increase in the
principle score for ‘Appropriate behaviour’ where the score in sub-periods 2, 3 and 4 were 2 to 4
points higher than in sub-period 1. Except for one farm in the growth period, the overall
categorisation of the farms did not change with date of assessment in the winter or growth periods.
This shows that the assessment protocol is quite robust to changes with date of assessment within
the defined time windows. Each measurement taken on the farm in each period uncovers one aspect
of the welfare, but one measurement should not alone explain the evaluation of the welfare at farm
level. However, changes in the evaluation of individual measurements taken on the farm in each
period should, to some degree, affect the evaluation at the higher levels as also shown for several
measurements in this study. The aggregations control how much each score affects the evaluation of
welfare at the higher levels. Higher scores can only partly compensate a lower score, which means
that changes in lower scores will have a larger effect than changes in higher scores. This also means
that the effect of changes in the evaluation of individual measurement is affected by the results of
other measurements. The compensation depends on the assigned weights; hence, changes in scores
with a high weight will have a larger effect than changes in scores with a lower weight. Finally, as
the number of measurements per criteria and number of criteria per principle vary, the more scores
that are aggregated, the lower is the effect of individual scores of the evaluation of welfare at the
higher levels.

As discussed by Henriksen (2015) for changes in welfare in the nursing period, the overall
categorisation of the farms could have changed if the principle scores were closer to the threshold
between two categories. The changes in the principle scores in this study were smaller than the
estimated changes with date of assessment for the principles ‘Good feeding’ and ‘Good housing’ in
the nursing period, which were above 25 points (Henriksen and Møller 2015). In order to avoid any
effect of the changes in the nursing period on the overall evaluation of welfare, Henriksen and
Møller (2015) suggested stratifying the assessments between the assessment periods into the
beginning, the middle and the end of each period in case the picture was similar in the other two
periods. Based on the results in this study, the changes in welfare with date of assessment in the
winter period only have a limited effect of the overall evaluation. Also, the changes in the growth period are pointing in opposite directions, as the changes with date of assessment for the principle ‘Good health’ are negative in regards to animal welfare and the changes for the principle ‘Appropriate behaviour’ are positive. The suggested stratification of the assessments does, therefore, not seem to be an adequate solution in order to avoid that the changes with date of assessment at measurement level affect the overall categorisation. An alternative solution could be to shorten the time window for each assessment period in order to limit the variation. This would, however, seriously challenge the ongoing assessment of the approximately 3000 mink farms in Europe that was initiated in 2017. A more feasible way could, therefore, be to develop a correction factor for the few measurements that can be expected to change within each assessment period. Finally, one can just accept the documented changes with date of assessment, and appreciate the fact that the effects on the principle scores and overall category are as limited, as they have turned out to be.

Conclusion

There was a change in the odds of welfare problems with date of assessment for all of the expected measurement variables, except stereotypic behaviour in the winter period. Also, the change in the odds of too thin mink was in the opposite direction of what was expected, as the odds decreased. Some changes with date of assessment were positive in regards to animal welfare, and others were negative. However, these changes only had a limited effect on the Welfare scores at the higher levels. Except for one farm in the growth period, the final categorisation of the farms did not change with date of assessment. However, if the scores were closer to the threshold between two categories, this could have been the case for more farms.
Animal welfare implications

In a seasonally synchronised production system such as mink production, it is important to include all seasons in the assessment, as the different seasons may be associated with specific animal groups and welfare risk factors (Møller et al. 2003). But also date of assessment within each season has to be considered, as there may be rapid changes within season which may affect the outcome of the assessment as highlighted by this study. As the results of the Welfur assessment are directly available to the farmers to be used in their daily management, the possible changes in welfare with date of assessment may be important for the farmers, even if it does not affect the overall evaluation. Knowing the effect of date of assessment gives the farmer the possibility to interpret the results in regards to the risk factors that were present or absent on the day of the assessment. This must also be considered when assessing the welfare in other production systems with seasonal breeders, as for example sheep production, or in production systems with non-seasonal breeders kept in seasonal breeding systems, as for example dairy cows in grazing systems with calving in spring. In production systems where all animals on the farm are the same age at the day of the assessment, as for example laying hens, this is also important.

Acknowledgements

We thank the farmers for kindly letting us into their farm. We also thank Elna Mortensen and Maria Bjaasbjerg-Olitsa, both Copenhagen Fur, for their help with data collection and valuable comments. Steen Henrik Møller was responsible for study design and Bente Krogh Hansen for data collection. Anna Feldberg Marsbell was responsible for data analysis which was completed with help from Leslie Foldager, Aarhus University. All authors contributed to the manuscript. Fur Europe and Aarhus University financed this project.
References


Scientifur


Hansen SW, Møller SH and Damgaard BM 2014 Bite marks in mink - Induced experimentally and as reflection of aggressive encounters between mink. *Applied Animal Behaviour Science 158*: 76-85

Honriksen BIF and Møller SH 2015 The reliability of welfare assessment according to the Welfur-protocol in the nursing period of mink (Neovison vison) is challenged by increasing welfare problems prior to weaning. *Animal Welfare* 24: 193-201


Table 1 The four principles of welfare with the 12 underlying welfare criteria and 22 measurements in WeFur-Mink (adapted from Mätler et al 2015).

<table>
<thead>
<tr>
<th>Principle</th>
<th>Criterion</th>
<th>Measurement</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Good feeding</td>
<td>1 Absence of prolonged hunger</td>
<td>1 Body condition score</td>
</tr>
<tr>
<td></td>
<td>2 Absence of prolonged thirst</td>
<td>2 Type of watering system</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3 Functioning and cleanliness of the water points</td>
</tr>
<tr>
<td>2 Good housing</td>
<td>3 Comfort around resting</td>
<td>4 Access to a nest box</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5 Resting quality of the nest box/resting area</td>
</tr>
<tr>
<td></td>
<td>4 Thermal comfort</td>
<td>6 Protection from exceptional weather conditions</td>
</tr>
<tr>
<td></td>
<td></td>
<td>7 Nest box material and bedding/nesting material</td>
</tr>
<tr>
<td>3 Good health</td>
<td>6 Absence of injuries</td>
<td>8 Space available for moving (cage area and cage height)</td>
</tr>
<tr>
<td></td>
<td>7 Absence of disease</td>
<td>9 Skin lesions or injuries to the body</td>
</tr>
<tr>
<td></td>
<td></td>
<td>10 Mortality</td>
</tr>
<tr>
<td></td>
<td></td>
<td>11 Diarrhoea</td>
</tr>
<tr>
<td></td>
<td></td>
<td>12 Lameness and impaired movement</td>
</tr>
<tr>
<td></td>
<td></td>
<td>13 Obviously sick animals</td>
</tr>
<tr>
<td></td>
<td>8 Absence of pain induced by management procedures</td>
<td>14 Killing methods for petting of mink</td>
</tr>
<tr>
<td>4 Appropriate behaviour</td>
<td>9 Expression of social behaviours</td>
<td>15 Killing methods for individual mink</td>
</tr>
<tr>
<td></td>
<td>10 Expression of other behaviours</td>
<td>16 Social housing$^2$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>17 Age and procedures at weaning$^3$</td>
</tr>
<tr>
<td></td>
<td>11 Good human-animal relationship$^1$ &amp; 12 Positive emotional state$^1$</td>
<td>18 Stereotypic behaviour</td>
</tr>
<tr>
<td></td>
<td></td>
<td>19 Cage enrichment</td>
</tr>
<tr>
<td></td>
<td></td>
<td>20 Fur chewing$^4$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>21 Frequency and duration of handling and transportation</td>
</tr>
<tr>
<td></td>
<td></td>
<td>22 Temperament test$^4$</td>
</tr>
</tbody>
</table>

$^1$ The two criteria are based on the same measurements
$^2$ Only assessed in the growth period
$^3$ Only assessed in the nursing period
$^4$ Not assessed in the nursing period
Table 2: Welfare-Mink classification of farms based on the four principle scores (adapted from Møller et al 2015).

<table>
<thead>
<tr>
<th>Category</th>
<th>Required principle scores</th>
</tr>
</thead>
<tbody>
<tr>
<td>Best current practice</td>
<td>Two principal scores above 60 and the remaining scores above 55</td>
</tr>
<tr>
<td>Good current practice</td>
<td>Two principal scores above 55 and the remaining scores above 20</td>
</tr>
<tr>
<td>Acceptable current practice</td>
<td>Three principal scores above 20 and the remaining score above 10</td>
</tr>
<tr>
<td>Unacceptable current practice</td>
<td>If the requirements for 'Acceptable current practice' is not met</td>
</tr>
</tbody>
</table>

Table 3: The division of the winter and growth assessment periods into sub-periods.

<table>
<thead>
<tr>
<th>Assessment period</th>
<th>Length of assessment period</th>
<th>Sub-period</th>
</tr>
</thead>
<tbody>
<tr>
<td>Winter</td>
<td>1 January (day 1) to 20 February (day 51) (or when flush feeding begins)</td>
<td>1: day ≤ 17</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2: day ≥ 18 and ≤ 35</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3: day ≥ 36</td>
</tr>
<tr>
<td>Growth</td>
<td>23 September (day 1) until 30 November (day 69) (or when sorting/petting begins)</td>
<td>1: day ≤ 13</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2: day ≥ 14 and ≤ 27</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3: day ≥ 28 and ≤ 40</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4: day ≥ 41</td>
</tr>
</tbody>
</table>
### Table 4 Description of the outcome variables included in the analysis at measurement level

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Too thin</td>
<td>Body condition is assessed on a scale from 1 (very thin) to 5 (obese). Mink in category 1 and 2 in January and category 1 in February are associated with prolonged hunger (too thin). Binary variable (too thin or not).</td>
</tr>
<tr>
<td>Fur chewing</td>
<td>Fur (guard hair or wool) have been chewed off. Assessed on a scale from 0 (no/very little) to 3 (extensive fur chewing). Binary variable (≤1 or 3).</td>
</tr>
<tr>
<td>Stereotypic</td>
<td>A repetitive, invariant behaviour without any obvious function or goal. If three repetitions of the same behaviour is observed in the two-minute observation period the mink is considered stereotypic. Binary variable (stereotypic or not).</td>
</tr>
<tr>
<td>behaviour</td>
<td></td>
</tr>
<tr>
<td>Injuries</td>
<td>Healed or unhealed wounds and injuries. Assessed on a scale from 0 (no injuries) to 3 (major unhealed injuries). Binary variable (≥1 or 0).</td>
</tr>
<tr>
<td>Diarrhoea</td>
<td>Diarrhoea is defined as very mucous, watery, fluent or bloody manure without form or texture. Diarrhoea is registered at cage level as it is not possible to determine how many mink in a cage that are affected. Binary variable (signs of diarrhoea or not).</td>
</tr>
<tr>
<td>Exploratory</td>
<td>The behavioural response of the mink to a wooden tongue spatula inserted into the cage. The mink’s response is assessed as exploratory, fearful, aggressive or undecided. Binary variable (exploratory or not).</td>
</tr>
</tbody>
</table>
Figure 1: Percentage of too thin mink (a), mink with fur chewing (b) and stereotypic mink (c) at the different assessment days in the winter period. Each symbol represents a farm and the vertical dotted lines indicates the shift between sub-periods.
Figure 2 Measurement scores for ‘Body condition score’ (a), ‘Fur chewing’ (b) and ‘Stereotypic behaviour’ (c) at the different assessment days in the winter period. Each symbol represents a farm and the vertical dotted lines indicates the shift between sub-periods. 100 is the best score and 0 is the worst in regards to animal welfare.
Figure 3 Criteria scores for 'Absence of prolonged hunger' (a) and 'Expression of other behaviours' (b) when the measurements 'Body condition score', 'Fur chewing' and 'Stereotypic behaviour' were included with their actual values at different assessment days in the winter period. Each symbol represents a farm and the vertical dotted lines indicate the shift between sub-periods. 100 is the best score and 0 is the worst in regards to animal welfare.
Figure 4 Principle scores for 'Good feeding' (a) and 'Appropriate behaviour' (b) when the measurements 'Body condition score', 'Fur chewing' and 'Stereotypic behaviour' were included with their actual values at the different assessment days in the winter period. Each symbol represents a farm and the vertical dotted lines indicate the shift between sub-periods. 100 is the best score and 0 is the worst in regards to animal welfare.
Figure 5 Percentage of mink with injuries (a), cages with diarrhoea (b) and exploratory mink (c) at the different assessment days in the growth period. Each symbol represents a farm and the vertical dotted lines indicate the shift between sub-periods.
Figure 6: Measurement scores for 'Skin lesions or injuries to the body' (a), 'Diarrhoea' (b) and 'Temperament test' (c) at the different assessment days in the growth period. Each symbol represents a farm and the vertical dotted lines indicate the shift between sub-periods. 100 is the best score and 0 is the worst in regards to animal welfare.
Figure 7 Criteria scores for 'Absence of injuries' (a), 'Absence of disease' (b) and 'Good human-animal relationship/Positive emotional state' (c), when the measurements 'Skin lesions or injuries to the body', 'Diarhoea' and 'Temperament test' were included with their actual values at the different assessment days in the growth period. Each symbol represents a farm and the vertical dotted lines indicates the shift between sub-periods. 100 is the best score and 0 is the worst in regards to animal welfare.
Figure 3 Principle scores for 'Good Health' (a) and 'Appropriate behaviour' (b) when the measurements 'Skin lesions or injuries to the body', 'Diarrhoea' and 'Temperament test' were included with their actual values at the different assessment days in the growth period. Each symbol represents a farm and the vertical dotted lines indicates the shift between sub-periods. 100 is the best score and 0 is the worst in regards to animal welfare.
4.3 Paper 3

If stereotypic behaviour cannot be observed before feeding in WelFur-Mink, a reliable conversion factor needs to be developed

Anna Feldberg Marsbøll, Tine Rousing and Steen Henrik Møller

Submitted for publication as a short communication in Animal Welfare
If stereotypic behaviour cannot be observed before feeding in WelFur-Mink, a reliable conversion factor needs to be developed.

Anna Feldberg Marsboll, Tine Rousing and Steen Henrik Møller

Department of Animal Science, Aarhus University, Blichers Allé 20, 8830 Tjele, Denmark

Corresponding author: Anna Feldberg Marsboll
E-mail: anna.marsboll@anis.au.dk

Running title: WelFur-Mink assessment of stereotypic behaviour
Abstract

Observations of stereotypic behaviour according to the WelFur-Mink assessment protocol should preferably be within 1.5 hours before the expected feeding time (before feeding) - alternatively within 1.5 hours before sunset (before sunset). In this study, we investigated the difference in the prevalence of stereotypic behaviour before feeding and before sunset in the winter period. We hypothesised that the prevalence of stereotypic behaviour is higher before feeding than before sunset and that the prevalence of stereotypic behaviour before sunset is sufficient to predict the prevalence before feeding. Based on repeated observations on five Danish mink farms, we found that the prevalence of stereotypic behaviour on average was 4.1 times higher before feeding than before sunset. The rank of the farms was, however, not the same, and the variation was larger before feeding than before sunset. This makes it problematic to derive a reliable conversion factor, as the relationship between prevalence of stereotypic behaviour before feeding and before sunset varies between farms. WelFur-Mink assessments on 1166 European farms in the winter period, where stereotypic behaviour was observed before feeding or before sunset, showed a large variation in the prevalence of stereotypic behaviour, illustrating that mink can be farmed with a low prevalence of stereotypic behaviour. The difference in the prevalence of stereotypic behaviour before feeding and before sunset was lower than on the five Danish farms. As the prevalence of stereotypic behaviour on the five Danish farms was in the high end of all the assessed European farms, they do not seem representative.

Keywords

animal welfare, feasibility, mink, reliability, welfare assessment, WelFur
Introduction

There are several reasons for why captive animals perform stereotypic behaviour, but it is generally related to present or previous sub-optimal housing conditions (Mason 2006). Stereotypic behaviour is, therefore, used as an indicator of reduced welfare in several on-farm animal welfare assessment schemes, as for example in the welfare assessment of farmed pigs (Sus scrofa domesticus) (Welfare Quality 2009), horses (Equus caballus) (AWIN 2015) and mink (Neovison vison) (Møller et al 2015).

In the WelFur welfare assessment protocol for mink (WelFur-Mink), where the welfare on individual mink farms is evaluated based on one assessment in each of the three annual assessment periods, stereotypic behaviour is defined as a repetitive, invariant behaviour without any obvious function or goal. As criteria for repeatability, three repetitions of the behaviour is used. Mink shoving stereotypic behaviour are registered in all assessment periods during two-minutes observations after habituation to the assessor. All adult and juvenile mink in the sample, which consists of 120 cages in the winter and nursing periods and 90 cages in the growth period, are observed (Møller et al 2015). When using animal behaviour as a measure of welfare, it is important that external factors that may affect the reliability are taken into account. For stereotypic behaviour in farmed mink, time of day is an important factor. Stereotypic behaviour in the winter period is mainly shown at the time of day when the mink are most active, i.e. at dawn, before the expected feeding time and at dusk (Hansen and Damgaard 2009; Malmkvist et al 2013). In order to standardise the assessment of stereotypic behaviour in WelFur-Mink, stereotypic behaviour should as a general rule be observed from 1.5 hours before the usual feeding time (Møller et al 2015). By doing this, the level of stereotypic behaviour is expected to be representative of the farm. However, sometimes it is not possible to observe stereotypic behaviour before feeding, for example due to
very early feeding on the farm. For feasibility reasons it was, therefore, decided to allow
observations of stereotypic behaviour in the winter period in WelFur-Mink to be carried out either
within 1.5 hours before the usual feeding time (before feeding) or within 1.5 hours before sunset
(before sunset). The observations before sunset are then expected to be carried out before the
activity peak at dusk, ie while the activity is increasing.

In this study, we investigate the difference in the prevalence of stereotypic behaviour in the winter
period observed before the expected feeding time and before sunset according to the WelFur-Mink
protocol (Moller et al 2015). Based on findings in previous studies (Hansen and Damgaard 2009;
Malmkvist et al 2013), we hypothesise that the prevalence of stereotypic behaviour in the winter
period is higher before feeding than before sunset. Hence, in order to ensure the reliability of the
assessment, ie that each farm will get the same assessment no matter if the assessment of
stereotypic behaviour is based on observations before feeding or before sunset, a conversion factor
may be needed. We hypothesise that the prevalence of stereotypic behaviour before sunset can be
used to predict the prevalence before feeding and that this is sufficient to derive a conversion factor
that will make the prevalence comparable. This study is based on 1) observations of stereotypic
behaviour before feeding and before sunset on five Danish mink farms in February 2017, and 2)
results from all European WelFur-Mink assessments conducted in the winter period 2018 where
stereotypic behaviour was observed either before feeding or before sunset.

Materials and methods
We observed stereotypic behaviour on five Danish mink farms in February 2017. The same 150
brown first-year female mink on each farm were observed twice; the first day they were observed
before sunset and the second day before feeding. The mink were housed individually and,
depending on cage design, it was possible for the assessors to observe six to twelve mink at the
time. Occurrences of stereotypic behaviour were registered according to the WelFur-Mink.
assessment protocol, i.e., each mink was registered as showing stereotypic behaviour or not based on
two-minute observations after habituation to the assessor (Møller et al. 2015). On each farm, two
calibrated assessors observed the same approximately 75 mink before feeding and before sunset.

Results from observations of stereotypic behaviour from all WelFur assessments on European mink
farms in the winter assessment period 2018 (from 1 January to 20 February) were made available by
Fur Europe. Assessments from 1166 farms from 16 different countries were found to be complete,
i.e., a sample of 120 mink observed on each farm and the timing of the observations registered as
either before feeding or before sunset. On approximately 75% of the farms, stereotypic behaviour
was observed before feeding. On the remaining 25%, stereotypic behaviour was observed before
sunset. In total, 54 assessors conducted the welfare assessments. More than 50% of the assessed
farms were Danish, and the assessments on these farms were conducted by 18 assessors. All
assessors had completed a one-week WelFur-Mink assessor course in 2016 or 2017 which included
theoretical and practical training in the assessment of stereotypic behaviour, taught by the first
author of this paper. In addition, all assessors participated in a WelFur-Mink calibration day before
the start of the winter assessment period 2018.

The statistical analyses were performed with the statistical software R (R Core Team 2017). The
limit for statistical significance was 5%. Based on visualisation and the Shapiro-Wilk normality
test, data from the five Danish farms were found to be approximately normally distributed. The
difference in the prevalence of stereotypic behaviour observed before feeding and before sunset was
analysed using a Paired t-test, and the correlation between prevalence of stereotypic behaviour
Results

On the five Danish farms, the prevalence of stereotypic behaviour observed before feeding ranged from 18% to 41% and before sunset from 5% to 13%. On all farms, the prevalence of stereotypic behaviour was higher before feeding than before sunset (Figure 1). The analysis confirmed this, as the prevalence of stereotypic behaviour observed before feeding was significantly higher than before sunset (df = 4, t = 4.1, P = 0.013). The average prevalence of stereotypic behaviour before feeding was 4.1 times higher than the prevalence before sunset. The rank of the farms regarding the prevalence of stereotypic behaviour was not the same when stereotypic behaviour was observed before feeding and before sunset (Figure 1). This was confirmed by the analysis, as the prevalence of stereotypic behaviour observed before feeding did not correlate with the prevalence observed before sunset (r = -0.41, n = 5, P = 0.49). Based on a visual assessment, one farm (farm C in Figure 1) stood out. This farm had the lowest prevalence of stereotypic behaviour before feeding and the highest before sunset.

On the 1156 European farms, the prevalence of stereotypic behaviour on farms observed before feeding ranged from 0% to 49% and before sunset from 0% to 25%. The majority of the prevalence was below 10%. Also, on 11% of the farms, the prevalence of stereotypic behaviour was 0% (Figure 2a). The analysis showed that the prevalence of stereotypic behaviour observed before feeding was significantly higher than before sunset (W = 149740, P < 0.001). The median of the
prevalence of stereotypic behaviour was 1.3 times higher before feeding compared to before sunset.

147 The prevalence of stereotypic behaviour on the 656 Danish farms had a similar distribution to all
148 the European farms (Figure 2b). The analysis also showed that the prevalence of stereotypic
149 behaviour observed before feeding was higher than before sunset ($W = 55902, p < 0.001$), and the
150 median of the prevalence of stereotypic behaviour was 1.4 times higher before feeding compared to
151 before sunset.

152 Discussion

153 As expected, the prevalence of stereotypic behaviour on the five Danish farms, as well as the
154 WelFur-Mink assessments of the European farms, was higher when observations were carried out
155 before feeding than before sunset, which supports our first hypothesis. However, the rank of the five
156 Danish farms was not the same for the observations before feeding and before sunset. Also, the
157 variation in the prevalence of stereotypic behaviour was larger before feeding compared to before
158 sunset, which was evident both in the results from the five Danish farms and the European farms.
159 This makes it problematic to derive a conversion factor, as the relationship between prevalence of
160 stereotypic behaviour before feeding and before sunset may not be the same on all farms. Thus, as
161 the prevalence of stereotypic behaviour before sunset alone cannot predict the prevalence before
162 feeding, we reject our second hypothesis.

164 When looking at Figure 1, one of the five Danish farms stood out. At the observation before sunset,
165 the clear roofing plates on this farm were covered by snow, and it was very dark inside the shed.
166 Hence, the mink’s perception of time of day may have affected the observations on this farm, as the
167 mink may have experienced the sunset as earlier than it actually was. The threshold between night
168 and day for mink is less than 1 lux (Travis et al. 1971, 1972), and in Danish latitude, a snow cover
may well accelerate the time of perceived sunset. This means that the mink’s activity level on this
farm may have been closer to the activity peak at dusk than on the other farms, thus increasing the
prevalence of stereotypic behaviour. This also highlights the difficulty of timing the assessments
before sunset, as the timing should be in regards to the mink’s perception of sunset. This means that
not only the time of sunset but also the light-reducing conditions have to be taken into account.

The large difference between the prevalence of stereotypic behaviour observed before feeding and
before sunset on the five Danish mink farms in this study was not evident in the results from the
European WelFur-Mink assessments or the Danish WelFur-Mink assessments alone. The possible
effect of individual assessors was not taken into account in the evaluation of the results from the
European farms. However, inter-observer agreement has been found to be higher for stereotypic
behaviour than for other animal-based measurements (Møller et al 2012), thus the effect of assessor
is expected to be low. Also, when comparing the results from the five Danish mink farms to the
WelFur-Mink assessments on the European farms, it is evident that the five farms included in this
study are not representative of either European farms or Danish farms in general, as the prevalence
of stereotypic behaviour on the five Danish farms was in the high end of all the assessed farms in
Europe (including the Danish farms).

On most of the European farms (including the Danish farms), the prevalence of stereotypic
behaviour observed before feeding was lower than what has been reported in Danish studies
(Hansen and Damgaard 2009; Malmkvist et al 2013). In these studies, the mink were fed energy-
restricted during winter, which means that the mink were fed a low energy diet in order to slim the
breeding animals to prepare them for mating. The degree of energy restriction has been linked to the
time spent performing stereotypic behaviour in this period (Damgaard et al 2004). This procedure is
commonly applied in Denmark, but how common and how restricted it is in Denmark and in
Europe is unknown. If most farms use a less energy-restricted feeding strategy, this would explain
the relatively lower prevalence of stereotypic behaviour observed before feeding on the majority of
the European farms (including the Danish farms). Dammaard et al (2004) found that feeding close to
ad libitum reduced the time spent performing stereotypic behaviour before feeding in the winter
period compared to a more energy-restricted feeding. This difference in stereotypic behaviour
between feeding strategies was not evident shortly after feeding. Unfortunately, the study did not
include any registrations of stereotypic behaviour before sunset. In WelFur-Mink, the prevalence of
stereotypic behaviour before feeding is the reference value. If the prevalence of stereotypic
behaviour based on observations before feeding is affected by the degree of energy restricted
feeding, while the prevalence of stereotypic behaviour from observations before sunset is not, a
conversion factor would depend on the feeding strategy, which would be difficult to apply in
practice.

Mink must be active in order to display stereotypic behaviour, but when calculating the prevalence
of stereotypic behaviour at farm level in WelFur-Mink, activity is not taken into account. As there
may be fewer active mink at the observations before sunset, this may reduce the number of mink
showing stereotypic behaviour. And as both active and inactive mink are included in the
denominator, this may reduce the prevalence of stereotypic mink before sunset. In the present study,
it was not possible to investigate this further, as the number of active mink was not registered.

In order to derive an exact conversion factor, the farms must be ranked the same way regarding the
prevalence of stereotypic behaviour independent of the timing of the observations, and the
relationship between the prevalence of stereotypic behaviour observed before feeding and before
sunset must be similar on all farms. Thus, based on the findings of this study it was not possible to
derive a conversion factor for observations of stereotypic behaviour before sunset in the winter
period in WelFur-Mink. We suggest that future studies should focus on how different factors (eg
feeding strategies and light conditions) affect the prevalence of stereotypic behaviour observed
before feeding and before sunset. Also, as activity and stereotypic behaviour are linked, the number
of active mink could be taken into account in order to develop a conversion factor.

Animal welfare implications
In this study, the prevalence of stereotypic behaviour from more than one third of all European
mink farms is presented. There was a large variation in the prevalence of stereotypic behaviour on
the European farms. This illustrates that it is possible to farm mink with a low prevalence of
stereotypic behaviour in the winter period.

Acknowledgements
We thank the farmers for participating in this study and Julie Jespersen for help with data
collection. AF Marshall was responsible for study design, data collection and data analysis. All
authors contributed to the manuscript. Fur Europe and Aarhus University financed this project.

References
AWIN 2015 AWIN welfare assessment protocol for horses. DOI: 10.13130/AWIN_horses_2015

Damgaard BM, Hansen SW, Borsting CF and Møller SH 2004 Effects of different feeding
strategies during the winter period on behaviour and performance in mink females (Mustela vison).
Results - Paper 3

Hansen SW and Damgaard BM 2009 Running in a running wheel substitutes for stereotypies in mink (Mustela vison) but does it improve their welfare? Applied Animal Behaviour Science. 118: 76-83


Travis HF, Davis HR, Sprague DC, Bleiler DF and Duby RT 1971 Development of practical procedures for altering the fur growth and reproductive cycle of mink using an artificial environment. Part II. The intensity of light required to give a 'Light Response' stimulation. Report no 1, Cornell University

Travis HF, Gardner WJ, Pilbeam TE and Cole RS 1972 The intensity of light required to give a 'Light Response' stimulation to growing kits. Report no 4, Cornell University


Consortium, Lelystad, Netherlands
Figure 1. The prevalence of stereotypic behaviour observed within 1.5 hours before feeding and 1.5 hours before sunset on five Danish mink farms. The same 150 mink on each farm were observed twice. Each symbol represents a farm.

Figure 2. Distribution of prevalence of stereotypic behaviour from WelFur-Mink assessments in the winter assessment period 2018 on 1166 European farms (a) of which 650 farms were Danish (b). The median, interquartile range, minimum and maximum of the prevalence of stereotypic behaviour for observations within 1.5 hours before feeding and 1.5 hours before sunset are shown.


5 Discussion

The overall aim of this PhD study was to address central challenges regarding the validity, reliability and feasibility of WelFur-Mink which became apparent during the development and tests of the protocol when preparing for large scale implementation. Three challenges were identified and addressed in three different parts of this PhD study. The first part concerns the challenge of feasible and representative sampling of mink for the on-farm assessment of welfare, the second part the possible change in the results of the welfare assessment with date of assessment during the on-farm assessment periods of the winter and growth periods and the third part the timing of the observation of the measurement ‘Stereotypic behaviour’ on the day of the on-farm assessment in the winter period. This chapter first provides an overview of the results of each of the three parts of the study including a discussion of the hypothesis. This is followed by a discussion of the possible welfare improvements resulting from the implementation of WelFur-Mink in Europe, how the aggregation system affects the interpretation of the results, the use of conversion factors in animal welfare assessment systems, the challenge of balancing validity, reliability and feasibility and the generalisation of the results to European conditions. Some methodological considerations are also included.

5.1 Discussion of hypotheses

5.1.1 Hypothesis 1 – Sampling method

This part focussed on the development and test of a new method for sampling cages with mink for the on-farm assessment of animal welfare in WelFur-Mink. The new sampling method is considered semi-random as the cage sections (clusters of six adjacent cages) in the sample are selected based on a systematic distribution between the sheds followed by a random selection within the sheds. The representativeness of samples selected with this new method was tested by simulated sampling on a model farm and evaluated in regard to the physical characteristics of the farm and the mink’s characteristics.

Based on practical tests, the first version of this new sampling method was found feasible to use in practice. Moreover, the risk of assessor bias was avoided as the assessor was not responsible for selecting the cage sections. The test of the revised version of this new sampling showed a high trueness, which means that there was no systematic skewness. However, the precision of the method was more variable due to variation within sheds. For
individual factors, the estimated probability of selecting a representative sample was quite high. However, the estimated probability of selecting a sample that was representative according to all factors was rather low. Thus, it cannot be expected that a sample selected with this method is representative according to all factors, but it will be representative according to some or most factors.

The hypothesis

_A feasible and unbiased method for taking a representative sample of mink for welfare assessments according to WelFur-Mink can be developed. We expect that samples selected with this method will reflect, but not necessarily be an exact representation of, the farm on which they are taken._

was therefore accepted.

The first version of the new sampling method has been used as a part of WelFur-Mink since the implementation on the European mink farms was initiated in January 2017. The assessors use an excel sheet developed for the purpose. The mobile application used to enter the results of the WelFur assessments is currently being updated, and the plan is to integrate the revised version of the new sampling method (described in Paper 1) in the mobile application so the assessors have all information and necessary tools gathered in one application. The discussion of the required representativeness is one example of how feasibility and the validity and reliability of a welfare assessment system must be balanced. This will be further elaborated in Section 5.2.4. In this case, it was not possible to use the original sampling method in practice, and a more feasible method had to be developed. The documented representativeness of samples selected with this revised version of the new method should therefore be used in the discussion of whether this new method should be used in WelFur-Mink as it is or if more efforts should be put into further developing this new method or yet another sampling method.

The new sampling method has proven to be relevant for sampling mink for other purposes than WelFur-Mink and for sampling in other species as well. The Danish authorities have considered using the new method for the sampling of mink for animal welfare inspections. The new sampling method has also been considered for sampling mink for evaluating compliance with the ‘Canadian Code of practice for the handling and care of farmed mink’. A similar method has also been developed and tested for sampling of foxes for WelFur-Fox assessments (Koistinen et al. 2017).
5.1.2 Hypothesis 2 - Changes in welfare with date of assessment

In this part, the possible change in the assessment of welfare with date of assessment during the on-farm assessment periods of the winter and growth periods was investigated. The focus was on the results of six welfare measurements which were expected to change with date of assessment either in the winter or in the growth period. These were ‘Body condition score’, ‘Fur chewing’ and ‘Stereotypic behaviour’ in the winter period and ‘Skin lesions or injuries to the body’, ‘Diarrhoea’ and ‘Temperament test’ in the growth period. Based on repeated welfare assessments on eight Danish farms, the changes in welfare with assessment date during each on-farm assessment period were investigated as the differences in results between sub-periods. The results were evaluated based on the prevalence of occurrences of too thin mink, fur chewing and stereotypic behaviour in the winter period and injuries, diarrhoea and exploratory mink in the growth period and the associated WelFur-Mink scores and the overall assessment at farm level.

The results showed a change in the results of some measurements. As expected, there was an increase in the odds of fur chewing with sub-period in the winter period. However, the odds of too thin mink in the winter period decreased (i.e. the change was in the opposite direction of what was expected), and there was no change in the odds of stereotypic behaviour. As discussed in Paper 2, the unexpected decrease in the odds of too thin mink may be related to the categorisation of body condition which differs between January and February. As expected, there was an increase in the odds of injuries, diarrhoea and exploratory mink with sub-period in the growth period. The effect of these changes on the WelFur-Mink scores on the higher levels were decreasing for every aggregation step. At the principle level, the changes were limited and only on one farm in one period, the overall assessment was somewhat affected by the changes. When including the actual values for all variable measurements, some of these changes were masked by variation in other measurements. This also showed a change with sub-period in the scores for the measurements ‘Cage enrichment’ in the winter assessment period and ‘Nest box material and bedding/nesting material’ in the growth assessment period. However, this had none or only a limited effect on the associated scores at criteria level and no effect at the scores at principle level.

The hypothesis

_Date of assessment in the winter and growth assessment periods affect the result of some measurements in WelFur-Mink. These changes do not necessarily affect the welfare_
scores or the overall categorisation.

was therefore accepted.

The changes with date of assessment during the on-farm assessment periods in the winter and growth periods shown in Paper 2 are limited compared to the changes during the reproduction and nursing period (Henriksen and Møller 2015). Moreover, some problems increase with assessment date while other decrease. However, date of assessment can affect the outcome of the assessment at farm level, if the principle scores are close to a threshold between two categories. The results imply that stratifying the visits between the on-farm assessment periods so that each farm is visited late in one period, in the middle in another period and early in another period, as suggested by Henriksen and Møller (2015), is not a relevant solution. It therefore seems that a conversion factor to counteract potential changes with assessment date may be the best way to level out the effect of date of assessment, taking the differences in the changes between seasons and measurements into account. The use of conversion factors will be further discussed in Section 5.2.3.

5.1.3 Hypothesis 3 – Stereotypic behaviour in the winter assessment period

This part focussed on the consequences of the timing within the day of the assessment for results of the assessment of stereotypic behaviour in on-farm assessment period of the winter period in WelFur-Mink. More specifically, the difference in the prevalence of stereotypic behaviour when assessed before feeding or before sunset was investigated. Moreover, the possibility of deriving a conversion factor to adjust for such difference was explored. Repeated assessment of stereotypic behaviour on five Danish mink farms as well as results from European WelFur-Mink assessments were the basis of this study. The detailed results are presented in Paper 3.

As expected, the prevalence of stereotypic behaviour was higher when assessed before feeding compared to before sunset on the five Danish farms. However, the relationship between the prevalence of stereotypic behaviour before feeding and before sunset varied between the five farms, which makes it challenging to derive a reliable conversion factor. The difference between the prevalence before feeding and before sunset on the five Danish farms was higher than on the European farms. Furthermore, as the prevalence of stereotypic behaviour on the five Danish farms was in the high end of all the assessed farms in Europe, they do not seem representative for mink farms in Europe.
The hypothesis

*The prevalence of stereotypic behaviour in the winter assessment period in WelFur-Mink depends on the time of observation, but it is possible to derive a conversion factor to adjust for the differences.*

was therefore only partly accepted as it does not seem possible to drive a reliable conversion factor based on present knowledge.

The results of Paper 3 show that the assessment of stereotypic behaviour in the winter period needs to be adjusted for the timing of the assessment in order to ensure the reliability of the assessment. The use of conversion factors to adjust for such differences will be further discussed in Section 5.2.3. However, based on the results of the present study, it does not seem possible to derive a reliable conversion factor, because the relationship between the prevalence of stereotypic behaviour observed before feeding and before sunset differs between farms. Moreover, it seems that the five Danish farms are not representative for European farms, which is further discussed in Section 5.2.5. This study also highlights the importance of gaining more knowledge about the variation in the prevalence of stereotypic behaviour during the day in the winter. This is further discussed in Section 5.2.6

### 5.2 General discussion

The three different parts of this PhD study represent three different challenges of using WelFur-Mink on a large scale. As described in Section 1.1, some of the challenges approached in this study were foreseen during the development phase (i.e. changes in welfare with date of assessment) while others were the results of the later tests in practice where WelFur-Mink was used on a large scale (i.e. sampling method and the timing of the observation of stereotypic behaviour). Implementation of a welfare assessment system such as WelFur-Mink as an integrated part of an animal welfare assurance scheme may lead to animal welfare improvements. For a large scale implementation to be successful, feasibility of the welfare assessment system is crucial. However, if compromises regarding the validity and reliability are made in order to maintain or improve the feasibility of the system, documenting areas where the validity or reliability of the system may be impaired is of uppermost importance, and if possible, solutions to maintain or restore the validity and reliability should be developed. Knowledge of how the validity and/or reliability is impaired will enable the development of such solutions. If there are systematic variations
in the welfare with known factors, conversion factors may be used to adjust for this. This study was mainly based on data collected in Denmark, but the aim was to address challenges of using WelFur-Mink all over Europe. It is therefore important to take the possible differences across Europe into account when evaluating validity, reliability and feasibility as well as when developing robust solutions for challenges demonstrated in such analysis.

5.2.1 Animal welfare improvements
Animal welfare assurance schemes where animal-based products are labelled or otherwise marketed based on the originating farms’ certified degree of animal welfare may lead to market-driven animal welfare improvements. Different welfare improvement tools may also be incorporated as a part of the scheme, which may lead to animal welfare improvements on the scheme member farms.

Animal welfare is an aspect of product quality that cannot be experienced directly by the consumer when purchasing a product. Instead, the consumer must rely on external information (Verbeke 2009), and a Eurobarometer survey from 2015 showed that the majority of the European citizens are concerned about animal welfare and would like more information about the conditions of farmed animals in their respective countries. In addition, more than half of the European citizens are willing to pay an additional price for products produced in an animal-welfare friendly housing system (European Commission 2016). Animal welfare labels or other marketing make it possible for consumers to identify products corresponding to specified animal welfare requirements. The idea is that, if consumers purchase products with improved animal welfare at a slightly higher cost, more animals will be produced under the improved conditions. When it comes to the labelling of food products, this has been found to help some consumers decide which product to choose (Verbeke 2009). Consumers of products originating from fur animals includes both the buyers at the auction houses, the fashion designers as well as the buyers of the final product. In the WelFur animal welfare assurance scheme, the included auction houses have committed only to sell pelts from farms categorised as ‘Acceptable current practice’ or higher. Thus, the buyers at the auction houses will not have an option of choosing between products with different welfare levels. Due to the comprehensive sorting of the pelts according to a range of quality traits, it was chosen not to differentiate between welfare categories as a quality trait. However, the consumers at the later stages may have the possibility of choosing between products made of pelts from welfare certified farms or not.
Many animal welfare assurance schemes also aim to improve farm animal welfare within their scheme members as an integrated part of the scheme. This can be achieved by incorporating a continuous improvement approach as described by Main et al. (2014). This may be done by providing the farmers with the results of the on-farm welfare assessment, for example by benchmarking them to other farms, and provide advice on how to improve the welfare based on the results (Main et al. 2014). The WelFur assurance scheme has incorporated such approach. The results of the welfare assessment are available for the farmers directly after each assessment; hence, the farmers can adjust their management according to the results. The assessed cages are physically marked and the results are stated per cage, so the farmer can see exactly what was registered. This gives full transparency and can help the farmers identify what to change in order to improve animal welfare. Advisors have also been educated by the WelFur scientist to be able to advice the farmers on what to change in order to improve animal welfare on their farms based on the result of the WelFur assessments as described for mink by Møller et al. (2017a). These advisors have been trained to identify the areas where changes will have the largest effect on the WelFur scores and, thereby, on the animal welfare. Henriksen et al. (2015) also suggested that stable schools where farmers work towards a common aim, exchange knowledge and give practical advice to each other can be used to integrate the results of a WelFur-Mink assessment into farm management. The benefit of stable schools is that the farmers learn from and give advice to each other, and everybody is in the same situation. However, as there is no ‘expert’ involved, the overview of what to change in order to gain the largest welfare improvements may be missing as this may require in depth knowledge of WelFur-Mink. However, when it has been identified what to improve, stable schools may be used to develop farm specific solutions or to adjust a known ‘standard’ solution to local conditions. The seasonality of mink production introduces a time lag where the necessary changes identified one year, often cannot be introduced until the following year (Møller et al. 2003). This provides ample time to find solutions, and the farmers included in the study by Henriksen et al. (2015) did not consider this a challenge, as they are used to thinking one year ahead.

Input from external persons or sources may help identify welfare problems or solutions as well as providing inspiration (Vaarst et al. 2011), and Henriksen et al. (2015) suggested that the WelFur-Mink assessment reports could be used as external input for stable school discussions. In my personal experience, having someone coming to your farm to assess the welfare may lead to distribution of knowledge about welfare. For example, when testing
WELFUR-MINK on one European farm, the owners of the farm were interested in what we looked at. We explained that we basically look at the same things as they do themselves every day, including the physical condition of the mink and the available resources, such as which enrichments are available for the mink. At first, the two owners did not understand what we meant by enrichments. They were not familiar with the concept and had not considered that the mink could benefit from having enrichments in the cage. Thus, by having someone coming to their farm to assess the welfare, attention was drawn to an area, where the welfare of the mink could easily be improved once the farmer became aware of it. Today, this will of course be highlighted by the following WELFUR-MINK assessment report. Thus, I believe that the large scale implementation of WELFUR-MINK will make some farmers consider why the included measurements are important for the welfare of the mink. I also believe that some farmers may take this potentially new knowledge into consideration and use it to improve the welfare of the mink on their farm.

5.2.2 Consequences of the aggregations
In WELFUR-MINK, each measurement taken in each assessment period is supposed to uncover some aspect of the welfare according to the welfare principles and criteria defined in Welfare Quality®. Thus, one measurement alone should not be able to explain the overall assessment of the welfare at farm level, but the result of each measurement should to some degree contribute to the overall assessment of welfare at farm level.

Not all changes at measurement level have an equal effect on the calculated measurement scores. For example, a change in the prevalence of mink with a given welfare problem may have different effects on the measurement score for this problem, depending on whether the change is in the 'good' (low prevalence of problems) or in the 'bad' (high prevalence of problems) area, due to the non-linear reasoning of the experts as described in Section 3.1.3. In other words, a change in the prevalence of a given problem may have a different effect on the welfare score for that measurement, depending on which area the change is in. Most often a change in the 'good' area will result in a larger change in the score than a change in the 'bad' area. Furthermore, the percentage rule assigned to the decision trees which are used to combine several sub-measurements as described in Section 3.1.3 may result in some changes at measurement level having no effect at all. If for example the measurement score is decided by the 10 percent in the worst situation, it does not matter if it is 10 of 100 percent of the mink that are in the worst situation – the farm will get the same score as 10 percent is considered the boarder in this example. This rule also implies
that if only 9.9 percent of the mink are in the worst situation, this will not directly affect the overall score. Thus, changes in welfare will only affect the score if the mink in the worst situation found on the farm is affected, and the number of mink that must have their welfare changed in order to change the score depends on the percentage rule assigned in the specific situation. Therefore, when evaluating the effect of changes in welfare by using data collected on farms (i.e. as in Paper 2), the results may be affected by the area the assessed welfare is in. This is also relevant for the generalisation of the results of this study to European conditions, which is further discussed in Section 5.2.5. In Paper 3, the measurement scores were not calculated. However, ‘Stereotypic behaviour’ in the winter period is one of the measurements where the experts followed a more linear reasoning; thus, a change in the prevalence of stereotypic behaviour will have a similar effect of the associated score independent of which area (i.e. ‘good’ or ‘bad’) the change are in (Møller et al. 2015).

The effect of an individual measurement score on the aggregated scores at the higher levels is reduced for every aggregation step. The size of the reduction depends on both the expert’s evaluation of the relative importance of each individual score, the score’s values in the aggregation and the number of scores that are aggregated. As described in Section 3.1.4, when aggregating a number of scores, the lowest score is the starting point, and this score can partly be compensated by the higher scores in the aggregation. Thus, the lowest scores, which are associated with the worst welfare, have a large impact on the aggregated score. The more scores that are included in an aggregation, the more the effect of individual scores is reduced. The potential reduction differs as there is a different number of measurements per criteria score, and a different number of criteria per principle score. For example, this gives a high weight to the measurement ‘Body condition score’ as this is the only measurement for the criteria ‘Absence of prolonged hunger’ which is one of only two criteria that make up the ‘Good feeding’ principle (Møller et al. 2015). The reduction was visible in the results of Paper 2 and in the study by Henriksen and Møller (2015) where the effect of changes in the prevalence of welfare problems decreased for every aggregation step. Similarly, Kirchner et al. (2014) found that the consistency with time of the Welfare Quality® assessments of 63 farms with beef bulls was higher in the aggregated scores compared to the results at measurement level, which means that the changes in welfare at measurement level with time were less visible at the aggregated levels. Compared to Welfare Quality®, the reduction of the effect of individual scores is higher in the WelFur system due to the additional aggregation across the three periods, as described in Section
3.1.1. A similar situation would arise if the Welfare Quality® assessments of sows and piglets, growing pigs and finishing pigs were to be aggregated into a welfare assessment of the full life cycle of pigs.

When interpreting results from studies where WelFur-Mink scores are calculated, it is therefore important to be aware, that the effect of changes in welfare depends on several factors in addition to the degree of the individual problem. The calculations and aggregations in WelFur-Mink are similar to the ones used in Welfare Quality® which have been criticised for giving too much weight to individual measurements. For example, de Vries et al. (2013) found that a limited number of measurements had a strong influence on the classification of dairy farms in a study including 196 farms. However, as described in the previous section, it is important to take into account that the results of individual measurements may have a limited effect on some areas and a large effect on others due to the aggregation process. In another study including Welfare Quality® assessment of 92 Dairy farms in England and Wales, Heath et al. (2014a) found that the criteria ‘Absence of prolonged thirst’ alone could predict the overall score. It was suggested that this was an unintended consequence of the aggregation system. However, in the study by Heath et al. (2014a), there was more variation between farms in the principle ‘Good feeding’ than in the three other principles. The score for the criteria ‘Absence of prolonged thirst’ were also highly variable between farms, ranging from 3 to 100 points. Thus, it may be reasonable that ‘Absence of prolonged thirst’, which is given a high weight in the aggregation with ‘Absence of prolonged hunger’ into the principle ‘Good feeding’, has a significant effect on the overall assessment, if the variation in the other principles is limited. However, the individual effect of the criteria ‘Absence if prolonged thirst’, which is one of only two criteria making up the principle of ‘Good Feeding’, may still be higher than other criteria where three or four criteria make up each principle. As emphasised in the previous section, it is important to take into account the variation in other scores as well. Therefore, it would be interesting to repeat the study by Heath et al. (2014a), including farms with more variation in the other principles or criteria, to see if the effect that they found of ‘Absence of prolonged thirst’ on the overall assessment will be reduced. The effect of individual measurements on the overall assessment in WelFur-Mink remains to be investigated but could be investigated using the results from all WelFur-Mink assessments in Europe. This would maximise the variation in the assessed welfare.
5.2.3 Conversion factors in animal welfare assessments

The use of conversion factors to account for systematic differences in results has been suggested as a practical solution to ensure the reliability of welfare assessments. For example, when using data collected at slaughter-houses to provide information about the animal welfare at the farm level, the possibility of deriving conversion factors to adjust for the effect of the individual slaughter house, which may be caused by fixed differences in recording practices, have been investigated (Nielsen et al. 2017). Another example is the evaluation of animal density in cattle herds, where conversion factors to take body weight and body surface into account have been suggested (Jóhannesson and Sørensen 2000). In the present study, conversion factors are suggested as a possible solution to adjust for systematic changes in the welfare assessment caused by differences in the timing of the assessment across days (Paper 2) or within the day of the assessment (Paper 3). Thus, conversion factors are suggested to be used in order to ensure that the required feasibility does not impair the reliability of the assessment, that is in regards to the length of the on-farm assessment period (Paper 2) and the possibility of observing stereotypic behaviour either before feeding or before sunset (Paper 3). A reliable conversion factor when adjusting for changes in welfare across or within days should ensure that that each farm gets the same or a similar assessment independent of the timing of the assessment. When evaluating a group of farms, the farms should be ranked similarly independent of the timing of the assessment. This requires that the relationship between the prevalence of the problems observed at the different times is similar on all farms. If this is not the case, using conversion factors to adjust for changes in welfare across or within days may lead to an over- or underestimation of the welfare on individual farms. This challenge was evident in Paper 3 where the rank of the farms differed between the timing of the observations.

Conversion factors can be applied at different levels in the assessment system. An argument for applying conversion factors at measurement level is the transparency of the assessment and the value of the assessment as a management tool to improve animal welfare. If a farmer is given a report with a low prevalence of a problem, there are no reason to improve the welfare. However, if the prevalence of the problem is related to the timing of the assessment, and the farmer is made aware that the prevalence is expected to be higher at another time, the farmer may take actions to improve the welfare. As shown in Paper 2, date of assessment does not have the same effect on all measurements as some welfare problems increased with date of assessment while other decreased. If only looking at the WelFur scores at the higher levels, the adjustment may be less necessary as the effect
of differences in individual measurement results is reduced due to the aggregation process described in Section 5.2.2. However, conversion factors applied at measurement level may be considered a way to bring the result of each measurement closer to the true value to make sure that the aggregated result is as correct as possible. However, if the focus is only on the reliability of the overall assessment, only changes in the principle scores are relevant to take into account. In this study, the results of Paper 2 show that the changes with date of assessment at principle level are limited during the winter and growth periods. Thus, in order to ensure the reliability of the overall assessment, it does not seem necessary to derive a conversion factor to adjust for date of assessment in these on-farm assessment periods. However, the effect of date of assessment during the reproduction and nursing period was larger than in the winter and growth periods, and a conversion factor may still be needed in this period (Henriksen and Møller 2015).

To ensure the value of WelFur as a management tool, I suggest to aim at deriving conversion factors to adjust for changes in welfare across or within days at measurement level. The study by Henriksen and Møller (2015) was based on repeated assessments on four Danish farms, Paper 2 on repeated assessments on eight Danish farms and Paper 3 on repeated observations of stereotypic behaviour on five Danish farms. However, as indicated in Paper 3, the results from the Danish farms may not be representative of all European farms. Thus, the possible differences in animal welfare within Europe should be taken into account to ensure that the conversion factors can be applied all across Europe. Another possibility is to take the change in welfare with date of assessment or within the day of the assessment into account in the experts evaluation of the welfare, that is to let the experts’ interpret the welfare by taking date of assessment or the timing within the day of the assessment into account. For example, the experts may interpret a given prevalence of stereotypic behaviour differently depending on the timing within the on-farm assessment period or the day of the assessment. For the experts to be able to do such an interpretation, information about the changes in the welfare with date of assessment or within the day of the assessment should still be available. This may allow for a less linear interpretation of the welfare. For example, a farm with a relatively high prevalence of a welfare problem in a part of the period where this problem is normally not present may be judged significantly harder than what would have been the case when assessing the welfare based on an adjusted value later in the period. Thus, the experts may reason that when a welfare problem is present at a time where it is normally not present the welfare on the farm is truly compromised.
5.2.4 Balancing validity, reliability and feasibility

The necessary feasibility when using an animal welfare assessment system on a large scale may challenge the validity and reliability of the system. In WelFur, this balancing act has been present at all levels, that is from the selection of which measurements to include in the assessment protocol to the adjustment of the protocols and development of practical guidelines for the assessment. It can be argued that if the validity and reliability are truly compromised, the value of the assessment is missing. On the other hand, if the assessment system is not sufficiently feasible, it will not be used or implemented on a large scale. The challenge is therefore to find solutions that ensure the feasibility without compromising the validity and reliability (too much).

One example of this balancing act is the discussion of when a sample is considered representative. As described in Section 2.2.7, if the sample that the welfare assessment is based on is not representative for the farm, the assessment may not be valid. However, as discussed in Paper 1, it is most often not possible to get an exact representation of a farm in WelFur-Mink due to the selection of the cages in sections (clusters of six adjacent cages). If considering a deviation of ± one cage section representative, the chance of selecting a sample that was representative according to all factors was rather low while the chance of selecting a sample that was representative according to some or many factors was rather high when testing the method on the complex test farm. Based on the discussion in Paper 1, I suggest that this method balances present requirements and possibilities for feasible and representative sampling in WelFur-Mink. However, the consequences of the reduced representativeness for the assessment of the welfare may still need to be investigated. This will be further discussed in Section 5.2.6. The probability of selecting a representative sample may be improved by increasing the sample size or selecting the cages in smaller clusters. Both suggestions would, however, strongly affect the time needed for the assessment. However, as the time needed for the on-farm assessment on many farms in practice is shorter than originally anticipated, this could be an option within the given framework of the present system. The main benefit of the new sampling method presented in Paper 1 was the feasibility of the method. Also the reliability was increased compared to the original sampling method due to the independency of assessor. Other sampling methods balance validity, reliability and feasibility differently. For example, the original sampling method was, in theory, representative, but there was a risk of assessor bias, and the feasibility was low on some farm.
'Stereotypic behaviour' in the winter period is an example of a measurement where the feasibility of the assessment in some cases was compromised when used in practice. The practical solution was to observe stereotypic behaviour before sunset if observations before feeding was not possible. However, as discussed in Paper 3, this will require that the result of the registrations is adjusted for the timing of the assessment, and based on the present knowledge, this does not seem possible. The question is now whether the feasibility is compromised too much if only allowing observations to be carried out before feeding, or if the reliability is compromised too much if continuing to allow for observations before sunset if it is not possible before feeding without assigning a conversion factor. In a worst-case scenario, it can be argued that the measurement ‘Stereotypic behaviour’ in the winter period should be excluded from the assessment protocol. However, if the observation of stereotypic behaviour is restricted to take place only before feeding, one possible solution could be to require that on farms that feed early in the morning, feeding must be postponed until the observation of stereotypic behaviour is completed. As previously mentioned, postponing feeding up to 1.5 hours has been found not to affect the frequency of stereotypic behaviour (Møller et al. 2014). I therefore suggest that the most feasible and reliable solution is to restrict observations of stereotypic behaviour to be performed only before feeding or during postponed feeding until more knowledge about the variation within the day in the prevalence of stereotypic behaviour in the winter period is available. However, if many farmers are not willing to postpone feeding due to welfare concerns or changes in routines, this must be included as a requirement for being a part of the WelFur-Mink assurance scheme. Whether measurements that may be affected by date of assessment (Paper 2) can be included in the assessment system if it is not possible to derive a reliable conversion factor, is a similar discussion. It could also be possible to consider such reliability challenges in the aggregations, which means that instead of excluding measurements, they may be given less weight in the aggregation into the overall assessment of welfare at farm level. This will give the farmers the information about the welfare in terms of the raw prevalence of problems, but they will have to be interpreted in regard to the reduced reliability of the measurement.

5.2.5 Generalisation of results
WelFur-Mink is developed with the aim of being able to assess the welfare of farmed mink in Europe. Mink are housed under similar conditions across Europe, as described in Section 2.1.1. However, the farms may differ significantly in size, which makes it difficult to select a ‘standard’ farm. When testing the new sampling method as described in Section
3.2.1 and Paper 1, it was decided to use a model farm that was considered a ‘worst case scenario’. As described in Section 3.2.1, this farm was larger than the average Danish and European mink farm, but this was intentional as there are examples of even larger farm across Europe. Thus, the results is not an example of how the method is expected to perform on most mink farms in Europe, but an example of how good or bad it performs in a worst-case scenario. Thus, on most European farms, the new sampling method is expected to perform better than on the worst-case scenario farm.

Despite the similar housing conditions, there may still be differences in the welfare challenges across Europe which may be related to for example differences in thermal conditions and available resources. For example, the cold weather during winter in Finland and the warm weather during summer in Greece lead to different challenges regarding the thermal comfort of the mink. Another example is related to the available bedding material, where straw is commonly used in Denmark whereas wood shavings are more common in Finland. In addition, the farming traditions and the available knowledge may differ between countries. For example, in some countries, there are no available advisory services, while in other countries it is mandatory to have a veterinarian visiting the farm several times a year. Thus, the welfare challenges and prevalence of welfare problems may differ between countries and regions. This means that the results obtained from the Danish farms in this study cannot necessarily be generalised to European conditions. As mentioned in Section 5.2.3, conversion factors aiming to adjust for changes in welfare across or within days need to take the possible variation at European level into account to ensure that the conversion factors can be applied all across Europe. As discussed in Section 5.2.2, the same change in for example the prevalence of animals with a given welfare problem may have different effects on the overall assessment, depending on whether there are many or few animals with the problem on the farm. Therefore, if the prevalence of welfare problems on Danish mink farms differs from the rest of Europe, this may imply that results on aggregated scores cannot be generalised to European conditions either. One of the stated aims of WelFur is to document the welfare of fur animals across Europe. Thus, the results from the welfare assessments conducted in Europe may be used to investigate this in terms of the individual measurement results as well as at the aggregated levels, which will be further discussed in Section 5.2.6.

The prevalence of stereotypic behaviour on five Danish farms included in Paper 3 was in the high end of all assessed farms in Denmark and the rest of Europe. Only brown first-
Discussion

Year female mink were included on the five Danish farms. This limits the generalisation to WelFur-Mink in practice as a regular sample farm will include both males, first-year and older females and different colour types if present on the farm. There are examples that older females spend more time performing stereotypic behaviour than first-year females (Malmkvist and Schou 2017). This implies that the prevalence of stereotypic behaviour on the five farms may be higher in a regular WelFur-Mink assessment.

5.2.6 Methodological considerations

The performance of the new sampling method presented in Section 3.2.1 and Paper 1 was tested in regard to the characteristics of the mink and their housing environment on the farm. This approach was chosen as the aim was to replace the original sampling method where each sample was stratified according to a range of factors which are known or believed to be related to the welfare of the mink. These factors are often confounded with shed or area on the farm; thus, the distribution of the sampled mink in regard to these factors is highly important. This approach may, however, still be considered a limitation since the overall aim is to end up with a welfare assessment that reflects the welfare of all the mink on the farm and not only selected characteristics. As described in the previous section, the test of the new sampling method was based on simulated sampling on only one model farm in one assessment period, which may also be considered a limitation of the generalisation of the results to other farms and periods. It could be possible to take this into account by simulating different farms with different prevalence of welfare problems in different assessment periods. The layout of existing farms could be taken as the starting point. The variation in welfare measured by resource-based measurements related to less variable housing conditions such as cage size or nest box insulation capacity should follow the actual distribution on the farms, as this is most often confounded with shed or areas of the farm. The prevalence of different welfare problems measured by direct animal-based measurements and some of the more variable resource-based measurements could be assigned and distributed randomly, aiming to reflect the variation in welfare as indicated by the European WelFur-Mink assessments. The performance of the method could be evaluated by calculating the agreement between the welfare of all the mink on the farm and the welfare of the mink in a number of simulated samples. This would give the opportunity to evaluate the performance of the method on farms of different sizes, in different assessment periods and with different prevalence of welfare problems. However, possible associations between welfare problems, such as specific housing conditions are not
considered by this approach unless such associations are taken into account in the distribution of welfare problems on the farm.

As described in the previous section, Paper 2 is based on a limited subset of Danish farms; thus, there is a risk that the variation found in this study is limited compared to the variation across Europe or even across Denmark. This risk was highlighted by the results of Paper 3, where the five Danish farms only represented a limited range of the prevalence of stereotypic behaviour in Denmark and in Europe. However, the results of Paper 2 may still be used to indicate which measurements can be expected to change with date of assessment during the on-farm assessment periods of the winter and growth periods. It may be possible to use the results from the WelFur-Mink assessments in Europe to estimate the possible variation and changes in welfare with date of assessment in each on-farm assessment period. This will not provide information about the changes within individual farms, but may provide a more general picture of the changes in welfare with date of assessment in Europe. Results from the WelFur-Mink assessments in Europe may also be used to evaluate the changes in the overall assessment of welfare, taking the changes in welfare with date of assessment during all three on-farm assessment periods into account. Differences between countries or regions could also be highlighted this way and indicate if there is a need for more in-depth investigations in different parts of Europe. Such investigations could include selected farms in Europe focusing on the measurement changes indicated in Paper 2 and the study by Henriksen and Møller (2015). The risk of misclassification of farms may also be included. As in Paper 3, the results can be compared to results from WelFur-Mink assessments in Europe to see how well the results correspond to a larger subset of mink farm in Europe.

The results of Paper 3 clearly stated the limitations of the study in regard to the generalisation of the results from the five Danish farms, as the five Danish farms only represented a limited range of the prevalence of stereotypic behaviour registered during WelFur-Mink assessments on farms in Denmark and in the rest of Europe. The results of Paper 3 therefore highlight the importance of gaining more knowledge about the variation in the prevalence of stereotypic behaviour during the day in the winter period, and how this is related to for example differences in feeding strategy and activity level. As described in Paper 3, occurrences of stereotypic behaviour in the winter period are often related to hunger but may also have other underlying reasons. It is therefore also relevant to investigate if the occurrences of stereotypic behaviour observed before feeding and before

156
Discussion

sunset are caused by the same underlying frustrations. In other words, it could be that the
stereotypic behaviour, occurring before feeding or before sunset, in fact has different
reasons and represents different welfare problems.
6 Conclusions

This PhD study addressed some of the challenges regarding the validity, reliability and feasibility of WelFur-Mink which became apparent during the development and tests of the protocol.

The new sampling method, that was developed as a part of this PhD study was found to be sufficiently feasible when used on a large scale. Moreover, assessor bias was avoided as the selection of cages is independent of the individual assessor. Testing showed that the revised version of the new method has no systematic skewness. It also showed that samples selected with this new method cannot be expected to be representative for the farm according to all factors related to the mink’s characteristics and housing environment. However, the samples can be expected to be representative for the farm according to some or most factors. Thus, it is suggested that this revised version of the new sampling method fulfil the requirements for feasible and representative sampling in WelFur-Mink.

The investigation of the possible changes in the WelFur-Mink assessment with date of assessment showed a change in the results of some of the measurements during the on-farm assessment periods of the winter and growth periods. Some changes were positive in regard to the assessed animal welfare while others were negative, and in some cases, the overall assessment of welfare at farm level may be affected. Conversion factors are suggested to be derived to be used to adjust for these changes with date of assessment at measurement level, thereby aiming to ensure the most correct assessment of all measurements independent of date of assessment and to ensure the value of the results as a management tool.

The results showed a large variation in the prevalence of stereotypic behaviour during winter on mink farms in Europe. The results also showed that the prevalence of stereotypic behaviour was higher when observations were carried out before feeding than before sunset. However, it may be challenging to derive a reliable conversion factor to adjust for this difference as the relationship between the prevalence of stereotypic behaviour observed before feeding and before sunset may vary between farms. Until more knowledge about the variation within the day in the prevalence of stereotypic behaviour in the winter assessment period is available, it is suggested that the most feasible and reliable solution
may be to restrict observations of stereotypic behaviour in WelFur-Mink to be carried out before feeding only or when feeding is postponed up to 1.5 hours.

As WelFur-Mink at present is in the process of being implemented as a part of an animal welfare assurance scheme for mink pelts produced in Europe, it is paramount that the system is highly feasible, but it is equally important that the validity and reliability are preserved. This PhD study is an example of how such a large scale implementation may require that the validity, reliability and feasibility of the animal welfare assessment system are re-evaluated and balanced.
In this PhD study, three challenges regarding the use of WelFur-Mink on a large scale in Europe were addressed. The results of this study may be used to decide on the direction of the future development of WelFur-Mink, for example if the new sampling method should be further developed, if conversion factors should be developed in order to adjust for the effect of date of assessment on some measurements during some on-farm assessment periods or if observations of stereotypic behaviour in the winter assessment period should be restricted to be performed only before feeding until we know more about the relationship between feeding strategy and the prevalence of stereotypic behaviour. The WelFur protocols were developed to be used under European conditions. However, it is possible to use the WelFur-Mink protocol for assessing the welfare of mink in non-European countries as well, especially where mink are housed in conditions similar to what is found in Europe. If the housing system differs significantly, some adjustments of the protocol may be needed, as there may be situations that cannot be described with the existing categories. However, most animal-based measurements are somewhat independent of the assessment system, as they measure the mink’s response to the production system. The results of this PhD study may also be relevant for official control of compliance with legislation on mink farms. For example, depending on what is included in the control, the date of the control visit may affect the result of the control in the winter and growth seasons as illustrated in Paper 2. The same has been pointed out for the reproduction and nursing season by Henriksen (2015). It is likely that challenges similar to the ones addressed in this study arise when using welfare assessments on a large scale in other farm animal production systems. The approach used to address the challenges in this study may be used as an inspiration for welfare assessments in other farm animal production systems.

Alongside this study, other challenges regarding the use of the WelFur-Mink protocol in practice have been addressed. For example, the test of WelFur-Mink in Europe raised the question of the possible impact of some cage designs on the result of the temperament test where the mink’s reaction to a wooden tongue spatula inserted into the cage is recorded. For example, a shelf in the front of the cage where the mink could lay and touch the spatula during the test was suspected to affect the mink’s reaction. This was therefore studied by Henriksen et al. (2017) who found that a shelf in the front of the cage increased the prevalence of exploratory mink. The guideline of how to perform the test was therefore
adjusted, and it is now emphasised that the spatula must be inserted into the cage in such a way that the mink cannot lay on a shelf and touch the spatula at the same time. The quality of the data for the measurement ‘Mortality’ has also recently been addressed. For this measurement, the farmers have to deliver mortality records by the end of each assessment period. Thus, this measurement cannot be taken as a part of the on-farm assessment, and the quality of the recordings made by the farmer may vary. Therefore, it has recently been decided to include an assessment of the quality of the mortality recordings as a part of the on-farm assessment. This means that the farmer will have to explain and show to the assessor how mortality is registered on the farm. If the assessor finds that the mortality recordings are not trustworthy, this will be taken into account in the assessment of this measurement by reducing the WelFur-Mink score. A similar assessment of the quality of the information may also be relevant for other measurements where the farmers provides the information such as ‘Frequency and duration of handling and transportation’ and ‘Age and procedures at weaning’.

Other challenges of using WelFur-Mink in practice still need to be solved. In this study, the feasibility of observing stereotypic behaviour in the winter assessment period was addressed, and the consequences of allowing observation of stereotypic behaviour to be performed either before feeding or before sunset were investigated. Similar challenges exist in the on-farm assessment periods of the reproduction and nursing period and the growth period. However, in these periods, feeding is expected to have less effect on the activity of the mink as they are most often fed *ad libitum* or close to *ad libitum*. Furthermore, the prevalence of stereotypic behaviour in these periods is significantly lower than in the winter period. However, the prevalence of stereotypic behaviour may still vary over the course of the day; thus, the timing of the observation of stereotypic behaviour may affect the result of the welfare assessment, and must thus be taken into account. The interpretation of thermal comfort in different climatic zones is also a challenge that must be addressed. Most of the available knowledge regarding the thermal comfort of mink comes from studies on farms in the northern part of Europe. However, as mink are also kept in warmer regions in the south of Europe, more knowledge about the mink’s thermal comfort in these conditions is needed in order to ensure a correct interpretation of the mink’s welfare, for example in regard to the assessment of the measurements ‘Nest box material and bedding/nesting material’ and ‘Protection from exceptional weather conditions’. As discussed in Paper 2, the categorisation of mink body condition in the winter assessment period (where body condition score 1 and 2 in January and 1 in
February are associated with prolonged hunger) may be the reason why the odds of too thin mink were found to decrease during this period. It should therefore be investigated if a more gradual categorisation could be developed, which may reduce the change in the odds during this period.

Continuous updating of the assessment protocols taking the most recent knowledge into account is an integrated part of the system. However, at some point it is expected that the increasing knowledge will affect the interpretation of mink welfare which may make is necessary to significantly change the structure or content of the assessment protocol. In this case, a new expert evaluation of the welfare will be needed. If the assessment protocols are to undergo a thorough updating process, it may also be relevant to address the structure of the entire system with the aim of improving the overall feasibility of the assessment protocols. For example, the AWIN welfare assessment system has adopted a two-step approach where an initial screening including selected key measurements, determines if a more detailed assessment is needed (AWIN 2015a, b, c, d, e). The results of the European WelFur-Mink assessments may be used to identify such key measurements and the consequences of such approach. It may also be relevant to address the structure of the aggregation system in order to enhance the transparency of the somewhat complex aggregation system. It may for example be relevant to investigate some of the consequences of the present aggregation system, such as the reduction of the effect of individual measurements and the size of the possible compensation of a low welfare in one area with higher welfare in another area.

In the WelFur aggregations, all measurement scores are aggregated across the three assessment periods before they are aggregated into the 12 criteria scores and further into the four principle scores. This means that all animal groups on the farm over a production year (i.e. adults, juveniles and kits) are included in the assessment of each measurement at farm level. The difference in mink welfare between assessment periods was illustrated in a study by Henriksen et al. (2016), and they concluded that results from more than one assessment period are needed to classify the welfare on mink farms over a production year. A more transparent solution could be to remove the aggregation across the three assessment periods and calculate the criteria scores, principle scores as well as the overall assessment of welfare independently for each assessment period. By using this approach, it will be easier to identify which of the assessment periods are the most challenging. Moreover, as one aggregation step is removed, the possible compensation of low welfare in
one area with higher welfare in another area is also reduced. An overall assessment of the welfare across the three assessment periods could then be assigned based on expert evaluations. Assessing the welfare in each assessment period separately resamples how the welfare is assessed in for example the Welfare Quality® assessment of pig and cattle, where the welfare in each production phase is assessed individually, as described in Section 2.2.3.

Education of WelFur-Mink assessors is carried out by the scientists involved in the WelFur project. To be allowed to carry out the assessments, all assessors must follow a one-week course, which includes practical as well as theoretical training, and pass an exam based on pictures and videos. The assessors must also participate in a calibration day before the beginning of each assessment period (Møller et al. 2017b). However, the inter- and intra-observer reliability of assessors educated based on this programme still remains to be tested. The assessors are carrying out farm assessment for approximately two months after the period specific calibration day; thus, the assessment results may drift over this period. As a consequence, the reliability of the assessment may be challenged. The private company undertaking the WelFur assessments has an internal quality control system, but even so, it would be interesting to investigate inter- and intra-observer reliability of assessors to highlight areas where the training programme could be adjusted in order to ensure the reliability of the entire system, as described in Section 2.2.7.
8 Additional publications from the PhD study


Marsboll, A.F., Henriksen, B.I.F. & Møller, S.H. 2016. It is possible to take a representative sample of animals based on the number of cages with mink in each shed. (English: It is possible to take a representative sample of animals based on the number of cages with mink in each shed.) Annual Report 2015, Kopenhagen Research, pp. 15-23.


Møller, S.H., Marsbøll, A.F. & Henriksen, B.I.F. 2017. Svar på ofte stillede spørgsmål til WelFur-Mink: Hvorfør tæller halve haler, der er helet op, med som ”sår og skader”? (English: Answers to frequently asked questions to WelFur-Mink: Why do half tails that have healed count as a ‘Skin lesion or injury’?) Dansk Pelsdyravl, vol. 80, no. 7, p. 25.


Marsbøll, A.F., Henriksen, B.I.F. & Møller, S.H. 2017. Svar på ofte stillede spørgsmål til WelFur-Mink: Er det ikke uretfærdigt, at forekomsten af stereotyp adfærd på nogle gårde bliver registreret før fodring, hvor dyrene jo er mest aktive, mens det jo på andre gårde bliver først registreret inden solnedgang? (English: Answers to frequently asked questions to WelFur-Mink: Isn’t it unfair, that the occurrence of stereotype behaviour on some farms is registered before feeding where the animals are most active, while on other farms, it is registered before sunset?) Dansk Pelsdyravl, vol. 80, no. 8, p. 29.


## 9 Activities resulting in ECTS course credits

<table>
<thead>
<tr>
<th>Title</th>
<th>Institution</th>
<th>Year of completion</th>
<th>Workload (ECTS)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PhD course: Animal welfare and risk assessment</td>
<td>Aarhus University</td>
<td>2015</td>
<td>1</td>
</tr>
<tr>
<td>PhD course: Basic statistical analysis in life and environmental sciences</td>
<td>Aarhus University</td>
<td>2015</td>
<td>4</td>
</tr>
<tr>
<td>PhD course: Animal pain</td>
<td>Aarhus University</td>
<td>2016</td>
<td>5</td>
</tr>
<tr>
<td>PhD course: Mixed models</td>
<td>Aarhus University</td>
<td>2016</td>
<td>3</td>
</tr>
<tr>
<td>PhD course: Animal experimentation</td>
<td>Aarhus University</td>
<td>2016</td>
<td>2.5</td>
</tr>
<tr>
<td>PhD course: Animal welfare – from fork to farm</td>
<td>NOVA university network</td>
<td>2016</td>
<td>3</td>
</tr>
<tr>
<td>PhD course: Measuring and assessing animal welfare at herd level</td>
<td>Aarhus University</td>
<td>2017</td>
<td>5</td>
</tr>
<tr>
<td>Other activities: Visit at Fur Europe in Brussels</td>
<td>Fur Europe</td>
<td>2017</td>
<td>1.5</td>
</tr>
<tr>
<td>PhD course: Scientific writing and communication</td>
<td>Aarhus University</td>
<td>2018</td>
<td>4</td>
</tr>
</tbody>
</table>
10 Other dissemination activities

Teaching bachelor and master courses at Aarhus University

- Assisting teacher in ‘Velfærdsvurdering’ (English: Welfare assessments) at the BSc course ‘Husdyrproduktion’ (English: Animal production) in 2015.
- Teaching ‘Introduktion til mink production’ (English: Introduction to mink production) at the BSc course ‘Husdyrproduktion’ (English: Animal production) in 2015 and 2016.
- Teaching ‘Introduction to mink production’ at the MSc course ‘Introduction to animal production’ in 2015 and 2016.
- Teaching ‘Introduction to mink production’ at the MSc course ‘Animal production practices’ in 2017.

Educating WelFur-Mink assessors

- WelFur assessor training course for mink, AU-Foulum, December 2017.

Supervision

- AP graduate student in Agro Business and Landscape Management, Business Academy Aarhus, 15 ECTS project on stereotypic behaviour in mink.
- BSc student in Animal Science, Copenhagen University, 15 ECTS bachelor project on WelFur.

Presentations for visitors at AU-Foulum

- Assisting at the UNF Biotech camp visit to the stables.
- Introducing high school students to mink production and research on mink at AU-Foulum.
- Presenting the Agrobiology education at Aarhus University for high school students
- Introducing students from Wageningen University to mink production and research on mink at AU-Foulum.
- Introducing participants at the ‘51st Congress of the International Society for Applied Ethology’ to the welfare assessment system WelFur.
Conferences and seminars

- Oral presentation (‘Kan minks velfærd i vinter-og vækstperioden vurderes uafhængigt af observationsdatoen?’ (English: Can mink welfare in the winter and growth assessment periods be assessed independently of date of assessment)) at ’Mink-Temadag 2015’ at AU-Foulum, Denmark.

- Oral presentation (‘Do the WelFur-assessment of mink in the growth and the winter period change with the date of assessment?’) at ‘NJF Seminar 485 - Autumn Meeting in Fur Animal Research 2015’, Turku, Finland.

- Oral presentation (‘Det er muligt at udtage en repræsentativ stikprøve af dyr ud fra antallet af bure med mink i hver hal’ (English: It is possible to take a representative sample of animals based on the number of cages with mink in each shed)) at ‘CPH Mink Seminar 2016’, Copenhagen, Denmark.

- Oral presentation (‘It is possible to take a representative sample of animals based on the number of cages in use in each mink shed’) at the ‘XIth International Scientific Congress in Fur Animal Production’, Helsinki, Finland.

- Oral presentation (‘Hvordan kan vurderingen af stereotyp adfærd i WelFur-Mink standardiseres i praksis?’ (English: How can the assessment of stereotypic behaviour in WelFur-Mink be standardised in practice?)) at ‘CPH Mink Seminar 2017’, Copenhagen, Denmark.

- Oral presentation (‘Can faecal cortisol metabolites be used as an Iceberg Indicator in the on-farm welfare assessment system WelFur-Mink?’) at the ‘51st Congress of the International Society for Applied Ethology’, Aarhus, Denmark.

- Poster presentation (‘Date of assessment affects the WelFur-assessment of mink in the winter- and growth period?’) at the ‘7th International Conference on the Assessment of Animal Welfare at Farm and Group Level’, Wageningen, Netherlands.

- Participating at the ‘7th International Conference on the Assessment of Animal Welfare at Farm and Group Level’ innovation day with a stand about WelFur-Mink, Wageningen, Netherlands.

- Oral presentation (‘Hvordan kan vurderingen af stereotyp adfærd i vinterperioden standardiseres i praksis i WelFur-Mink?’ (English: How can the assessment of stereotypic behaviour in the winter period be standardised in practice in WelFur-Mink?)) at ‘Mink-Temadag 2017’ at AU-Foulum, Denmark.
11 References

AWIN 2015a. AWIN welfare assessment protocol for donkeys.
DOI: 10.13130/AWIN_DONKEYS_2015.

AWIN 2015b. AWIN welfare assessment protocol for goats.
DOI: 10.13130/AWIN_GOATS_2015.

AWIN 2015c. AWIN welfare assessment protocol for horses.
DOI: 10.13130/AWIN_HORSES_2015.

AWIN 2015d. AWIN welfare assessment protocol for sheep.
DOI: 10.13130/AWIN_SHEEP_2015.

AWIN 2015e. AWIN welfare assessment protocol for turkeys.
DOI: 10.13130/AWIN_TURKEYS_2015.


References


COOP n.d. Dyrevelfærd så det kan mærkes (English: Animal welfare so it can be labelled) [Online]. Available at: http://dyrevelfaerd.coop.dk/dyrevelfaerdshjertet/ [Accessed 22 August 2018].

Council of Europe 1999. Recomendation concerning fur animals. Available at: https://www.coe.int/t/e/legal_affairs/legal_co-operation/biological_safety_and_use_of_animals/farming/Rec%20fur%20animals%20E%201999.asp.

Dalmau, A. 2017. Development of a certification schema on animal welfare based on
References


Danish Veterinary and Food Administration 2017. New animal-welfare label will win the hearts of Danes [Online]. Available at: https://www.foedevarestyrelsen.dk/english/Animal/AnimalWelfare/Pages/New_animal_welfare_label_will_win_the_hearts_of_Danes.aspx [Accessed 22 August 2018].


Fur Europe 2017. 4,000 European fur farms to be WelFur certified by 2020 [Online].


Forkman, B. & Keeling, L. 2009b. Assessment of animal welfare measures for layers and broilers Welfare Quality Reports No. 9, Cardiff university, Cardiff, UK.


reducing the number of parameters within the Welfare Quality® assessment protocol 
Heath, C.A.E., Lin, Y., Mullan, S., Browne, W.J. & Main, D.C.J. 2014b. Implementing 
Welfare Quality® in UK assurance schemes: evaluating the challenges. Animal Welfare, 
23, 95-107.
Henriksen, B.I.F. 2015. Assessment of animal welfare in mink farms - based on the 
Henriksen, B.I.F., Anneberg, I., Sørensen, J.T. & Møller, S.H. 2015. Farmers' perception of 
stable schools as a tool to improve management for the benefit of mink welfare. 
Livestock Science, 181, 7-16.
Henriksen, B.I.F., Malmkvist, J. & Møller, S.H. 2017. The effect of different stimuli on the 
'stick test' should be taken into account when measuring the mink's temperament. In: 
Jensen, M.B., Herskin, M.S. & Malmkvist, J. (Eds.), Proceedings of the 51st Congress of 
the International Society for Applied Ethology, Aarhus, Danmark, Wageningen 
Henriksen, B.I.F. & Møller, S.H. 2015. The reliability of welfare assessment according to 
the WelFur-protocol in the nursing period of mink (Neovison vison) is challenged by 
Henriksen, B.I.F., Sørensen, J.T. & Møller, S.H. 2016. Classification of animal welfare on 
mink farms differs between three annual production periods. Animal Welfare, 25, 265- 
273.
Text, CABI, pp. 255-271.
N. (Eds.), Mink... biology, health and disease, University of Guelph, Ontario, Canada 
af velfærd (English: How good is the overall welfare? - about aggregation of welfare). 
In: Sørensen, J.T., Houe, H., Rousing, T. & Sandøe, P. (Eds.), Vurdering af dyrevelfærd 
i en husdyrbesætning, Center for bioetik og risikovurdering, Frederiksberg, pp. 41-49. 
Jóhannesson, T. & Sørensen, J.T. 2000. Evaluation of welfare indicators for the social
References

herd level: many goals, many methods. Acta Agriculturae Scandinavica, Section A -
Animal Science, 51, 26-33.
Keeling, L., Evans, A., Forkman, B. & Kjaernes, U. 2013. Welfare Quality® principles and
criteria. In: Blokhuis, H., Miele, M., Veissier, I. & Jones, B. (Eds.), Improving farm
animal welfare: Science and society working together: the Welfare Quality approach,
Wageningen Academic Publishers, Wageningen, pp. 91-114.
Kirchner, M.K., Schulze Westerath, H., Knierim, U., Tessitore, E., Cozzi, G. & Winckler, C.
2014. On-farm animal welfare assessment in beef bulls: consistency over time of single
measures and aggregated Welfare Quality® scores. Animal, 8, 461-469.
and feasibility issues and future perspectives with special regard to the Welfare
Welfur fox good health score in the current semi-random sampling method. In: De
Jong, I. & Koene, P. (Eds.), Proceedings of the 7th International Conference on the
Assessment of Animal Welfare at Farm and Group Level Wageningen, Netherlands,
Wageningen Academic Publishers, p. 182.
Korhonen, H., Hansen, S.W., Malmkvist, J. & Houbak, B. 2000. Effect of capture,
immobilization and handling on rectal temperatures of confident and fearful male
beautiful fur animals - genetics of colours, fur, defects and diseases. NJF Working
Group for Fur Animals.
Lundmark, F., Berg, C. & Rocklinsberg, H. 2018. Private animal welfare standards-
optimaries and risks. Animals 8, 1-16.
Lundmark, F., Berg, C., Schmid, O., Behdadi, D. & Röcklinsberg, H. 2014. Intentions and
values in animal welfare legislation and standards. Journal of Agricultural and
Environmental Ethics, 27, 991-1017.


RSPCA n.d.-b. RSPCA welfare standards - Is there an RSPCA Assured label on the


reliability of animal-based measures on growing pig farms. Livestock Science, 151, 35-45.


Scandinavica, Section A - Animal Science, 51, 73-77.
   In: Newsletter 6 – April 2018, Welfare Quality Network.
Welfur-Mink is an on-farm animal welfare assessment protocol for farmed mink which is used to certify mink farms in Europe. The aim of this PhD study was to address central challenges regarding the validity, reliability and feasibility of the Welfur-Mink assessment protocol which became apparent when the protocol was developed and tested. These challenges were in regards to the sampling of mink for the on-farm assessment and the robustness of the protocol to date of assessment as well as the timing within the day of the assessment. The findings of this study contribute to ensure a feasible and correct assessment of welfare on mink farms in Europe.