The development of a quantitative and qualitative protocol for assessing the welfare of laboratory mice

Ivone Campos-Luna BSc, MSc

Published work submitting to Newcastle University for the degree of Doctor of Philosophy in the School of Natural and Environmental Sciences

December 2019

Table of contents

List of figures

List of tables

Acknowledgements

I would like to thank specially to my supervisors, Dr. Mathew Leach and Dr. Amy Miller for their support, teaching and patience even outside of this work. They have gone beyond their duties to make this work possible. I would like to thank all of the staff from the Comparative Biology Centre at Newcastle University for their invaluable help during my PhD. Especially to the vet team, Dr. Rocio Fernandez Palacios-O'Connor and Dr. Kathy Murphy for their patience and knowledge. I would also like to acknowledge Andrew Beard for his help and support in this process. I would like to thank to Professor Francoise Wemelsfelder for her valuable help in developing chapters 3 and 5 of this thesis. An important acknowledge all of the animals that one way or another contribute to this work, they have taught me invaluable lessons in these four years. Finally, to my sponsor, Colciencias, for giving the funds for starting this work.

Los más importantes agradecimientos son para mi familia, mi madre y mi padre, que desde diferentes sitios me han acompañado y apoyado desde que inicie este proceso ya casi 10 años atrás. Gracias, porque sin su amor, apoyo y sin sus enseñanzas esto no habría sido posible. A mis hermanas y mis sobrinos por estar siempre ahí, desde la distancia y por ser parte de mi vida siempre a cada paso que doy. A mi abuela y mi tía Romelia, por su apoyo espiritual incondicional por animarme a seguir persiguiendo mis sueños. A mis amigas Diana Morales, Laura Sariego, Laura Pinzon y Catalina Medrano por estar para mí en las buenas y en las malas por ser las mejores amigas del mundo y levantarme cuando no podía más y celebrar juntas cuando estaba en la cima.

Abstract

Mice are the most commonly used species for biomedical research. Maintaining high levels of welfare for these animals is critical for the individual animal and for ensuring the scientific data collected from them is of high quality. An efficient protocol for assessing laboratory mouse welfare should include physiological, psychological and environmental indicators, and should be able to recognise both welfare strengths and shortcomings. The protocols that are currently used often lack indicators which assess the animal's psychological state, which is necessary for a holistic assessment and critical for animal welfare. This study aimed to develop a holistic protocol for the assessment of laboratory mouse welfare when the assessment is made in an audit or every day. The development of the protocol involved three stages: [1] validity, practicability and reliability assessment of existing indicators, through a Delphi consultation technique, [2] the validation of a novel indicator (Qualitative Behavioural Assessment –QBA) for the assessment of emotional expressions, [3] the reliability and practicability assessment of the developed laboratory mouse welfare protocol. The Delphi consultation identified hunched posture, coat condition, the exhibition of normal and abnormal behaviour and usage of nesting material as the most valid, reliable and practical indicators for welfare assessment in a laboratory environment. The most common handling method for laboratory mice is tail handling, which produces anxiety thus new techniques have been developed for handling laboratory mice such as tube handling. QBA showed a high intra and inter-observer agreement in the assessment of emotional states for laboratory mice that had undergone handling by two different techniques. This study demonstrated that mice handled by their tail were described as "anxious/fearful" whereas mice handled by a tunnel were described as "confident/ playful." The final protocols for an audit and everyday welfare assessment of laboratory mice demonstrated to be practical and reliable between assessors with experience in laboratory mice. These protocols can be used in laboratory facilities for assessing laboratory mice welfare, incorporating psychological state using QBA. The 3-phase method used to construct the final protocol showed to be a comprehensive process which explored well-know and novel indicators and comprises all aspects of mouse laboratory welfare including QBA as a novel method of assessing psychological state. The process of developing a protocol for assessing laboratory mouse welfare is critical and considering experts, previous and novel indicators can improve the assessment quality by including all the components of the definition of welfare.

Chapter 1. Welfare in laboratory animals: definition and assessment

1.1. Definition of animal welfare

The longstanding relationship between humans and animals has influenced the definition of animal welfare. In 1964, Ruth Harrison published “Animal Machines”, which questioned industrial farming and its effects on animal welfare by illustrating how extreme changes in a controlled environment could have significant effects on welfare. This was the first milestone publication regarding the definition of animal welfare, which led to the 1965 Brambell report. This report looked into the welfare of animals kept under intensive livestock husbandry systems and was an essential step forward in terms of animal welfare, as for the first time there was a recognition of animals as sentient beings, *“animals show unmistakable signs of suffering from pain, exhaustion, fright, frustration and can experience emotions such as rage, fear, apprehension, frustration and pleasure*” (Brambell, 1965). This report also led to the publication of the five freedoms (Table 1.1) by the Farm Animal Welfare Council which helped further define animal welfare, asserting that equal consideration should be given to physical and mental wellbeing (Farm Animal Welfare Council, 1992).

FREEDOM	DESCRIPTION
Freedom from hunger and thirst	By giving access to water and a diet to maintain health and vigour.
Freedom from discomfort	By providing an appropriate environment.
Freedom from pain, injury and disease	By preventing or rapid diagnosis and treatment.
Freedom to express normal behaviour	By providing sufficient space, proper facilities and the appropriate company of the animal’s kind.
Freedom from fear and distress	By ensuring conditions and treatment which prevent mental suffering.

Table 1.1. General description of the Five freedoms based on the Farm Animal Welfare Council (1992).

The following years marked great advances for animal welfare science. The first scientific definition of animal welfare was documented in 1986. Donald Broom defined animal welfare as ‘The state of an individual in regards its attempts to cope with the environment’ (Broom, 1986). In this definition, ‘cope’ refers to physiological and mental stability based on the animals biological functioning and is an objective approach of animal welfare (Broom, 1988; Broom, 1991a). Although Broom’s definition of animal’s welfare included physiological and mental stability with regard to environmental changes, the means of assessment of the physiological and mental stability of the animals were not defined, as they were not deemed to be relevant during that time.

A few years later, another scientific definition was offered by Duncan (1991). Contrary to the definition provided by Broom, this definition involved animals mental stability related with feelings as a unique component of animal welfare: “Our thesis is that animal welfare is dependent solely on the mental, psychological, and cognitive needs of the animals concerned” (Duncan and Petherick, 1991). This definition was criticised by Broom who stated that feelings were denoting only one part of the behavioural repertoire that animals used to cope with the environment without taking into account the physical status (Broom, 2011). It can be argued that Duncan’s definition complements Broom’s definition of welfare, as it provides more information about the importance of the animal’s mental and cognitive states. However, it cannot be considered as a stand-alone definition, as welfare comprises of both the animal's physiologic function and the impact of the environment as proposed by Broom.

The definition that was widely accepted was the one that includes both mental and physical well-being given by Broom. However, at the time, limited research was carried out studying animal emotions and psychological states (Dawkins, 1983; Dawkins, 1985; Dawkins, 1988). Preference tests and operant conditioning tests were used to assess motivational priorities in animals showing that behavioural deprivation can cause suffering (Treit *et al.*, 1980; Dawkins, 1988; Kirkden and Pajor, 2006). Other indicators such as fear as a negative feeling (Scott *et al.*, 2009) and animal management procedures (e.g. tail docking, cross-fostering) were assessed, demonstrating they can negatively impact animal welfare (Algers, 2012). As a result of the research on animal’s mental states, another definition of welfare was proposed in 1997 which stated that “the welfare of animals consists in the satisfaction of preferences over states of affairs, a more preferred environment results in a higher level of welfare” (Jensen and Sandøe, 1997). Jensen and Sandøe (1997), definition was based on the presence of a pleasant mental state and the absence of unpleasant negative states. This definition along with the development of new technology used for the assessment of welfare (e.g. MRI, infrared cameras) leads to a rapid expansion of research into animals emotions (conscious experience, where there is a mental activity with degrees of pleasure or displeasure). This research, however, is initially focused on negative mental states such as pain (Weary *et al.*,

2006) and anxiety (Ohl *et al.*, 2008). Research focusing on animal cognition (capacity of process and acquiring knowledge and understanding through experience, though and senses), communication and feelings (Douglas-Hamilton *et al.*, 2006; Pepperberg, 2006; Mendl *et al.*, 2009), cognitive bias (pattern of judgment of a situation affected by previous experiences or emotional states) (Mendl *et al.*, 2009) and preference and motivation tests (Kirkden and Pajor, 2006) were used for the assessment of animals emotions. At the end of the 2000s, a new approach started to emerge in animal welfare science, the study of positive emotions, pleasure and happiness as welfare indicator in animals (Boissy *et al.*, 2007a; Balcombe, 2009).

To date, Broom's definition of animal welfare is the most accurate, as it integrates the environment, the biological functioning and the mental well-being of the animal (Figure 1.1) (Fraser and Broom, 1997; Fraser, 1999; Duncan, 2005a). This definition states that animal welfare is the animal's ability to preserve their biological functions (physiological parameters), their psychological estate (mental well-being) and the ability to maintain their evolutionary needs taking into account the environment where the animal is being kept. (Dawkins, 1990; Broom, 1996; Fraser and Broom, 1997; Dawkins, 1998; Hubrecht, 2014).

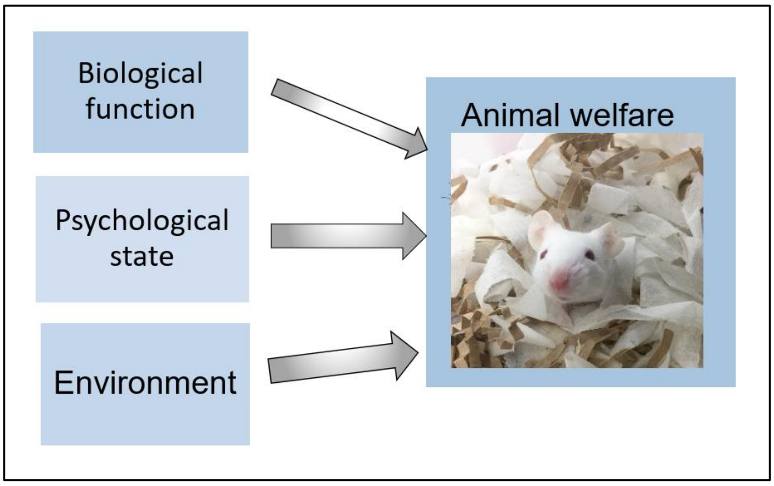


Figure 1.1. Holistic animal welfare definition that encompasses three main components

Understanding the psychological component of animal welfare is considered the most challenging component to assess, as it involves studying emotional states in animals. Animal feelings have been studied in the past (Dawkins, 1990; Mendl and Paul, 2004; Panksepp, 2004; Paul *et al.*, 2005; Broom, 2010) including the assessment of positive feelings (Boissy *et al.*, 2007b; Mendl *et al.*, 2009). Recently, new techniques including studies of the role of the observer (observer bias) in animal welfare assessment have been developed. Techniques such as Qualitative

Behavioural Assessment (QBA) are intended to incorporate the observer bias into the assessment protocol and use them as indicators to assess animal welfare (Wemelsfelder, 2007; Wemelsfelder *et al.*, 2012). QBA is used to assess psychological states of animals by measuring the behavioural expressions which represent the underlying emotions of animals (Wemelsfelder, 1997). QBA assesses how an animal is carrying out different activities involving behaviour, interaction with conspecifics and with the environment (Wemelsfelder, 2007).

1.2. The role of ethics and legislation laboratory mice welfare

The definition of animal welfare is intimately related to the human-animal interaction and the different roles that animals have had through human history. Furthermore, ethics underpins the concept of welfare as it acts as a starting point for how we define welfare. Scientific evidence about the impacts of our actions on animal welfare, and how it should be assessed and measured, is crucial to allow informed decisions regarding how animals should be treated. Animal Welfare as a scientific discipline that was created because of ethical concerns and the necessity to know more about which parameters should be assessed (Fraser *et al.*, 1997). Consequently, it is essential to clarify what ethics means and the role that both ethics and law play in human-animal interactions.

There are three main approaches to ethics concerning animal welfare: contractarianism, animal rights and utilitarianism (Regan, 1987; Morton *et al.*, 1989; Cohen, 1997). According to contractarianism, animals have no moral status. It is a system based on contracts where each individual enters into one with another. The limitation of this approach is related to the animal's lack of communication channels with humans - they cannot be part of a contract because they lack intellectual skills and linguistics (Cohen, 1997), therefore by this definition, they would have no moral status. However, strictly speaking, this theory can also be equally applied to humans who are not able to communicate such as very young children and people with communication disabilities. In relation to animal welfare, this approach can be used indirectly if humans feel that the animals have value (i.e. we care for them) and therefore include them in a contract. Additionally, contractarianism stated that any form of animal use is ethically acceptable as bring benefits to humans. This ethical view is highly anthropocentric since any measures taken to protect animals will depend on and secondary to human concerns (Palmer, 2010).

The animal rights approach states that animals have fundamental rights because they experience life and have consciousness (self-awareness and awareness of external objects or others) (Regan, 1987). Based on this theory, animals have moral equivalence with humans because they can have feelings, memory and have intentionality (Hubrecht and Kirkwood, 2010). Supporters of this theory such as Tom Regan believed that animals have inherent rights and animal experimentation is not acceptable because of the moral dilemmas that it raises (i.e. when experiments involve pain and discomfort to animals) (Regan, 1987; Mukerjee, 1997). However, the moral dilemma that involves animal experimentation has been contrasted with the ecological balance between species in relation to competition or predation which is incompatible with the possession of intrinsic rights (Matfield, 1991).

The last approach that is used to make decisions about animal welfare is Utilitarianism. Certainly, this is the approach that applies directly to laboratory animals. According to this approach, morality in animals has one single rule: always act to maximise the well-being of the individuals (Morton *et al.*, 1989). This approach states that actions can be considered right if they raise more benefits than cause harms (Morton *et al.*, 1989) and states that a greater benefit of an action outweighs a greater amount of suffering or harm (Singer, 2011). The main contribution of utilitarianism is related to the assessment of the outcome of the actions (e.g experiments) rather than the acts that are made to obtain a final consequence. In contrast with the animal rights approach, where animals possess intrinsic value for their moral status as a living creature, the utilitarian approach takes into account both humans and animals, but not equally (Regan, 1987; Francione and Garner, 2010). This approach is considered intermediate between animal rights and contractarianism because it does not disregard the belief that animals have rights but does result in some categorisation (Regan, 1987). One of the fundamental premises of the utilitarian approach is harm-benefit analysis regarding the animal-human relationship. It is important to note that most of the current guidelines and legislation about animals, in particular, animals used in research are based on this ethical approach. Legislation such as the EU directive (Directive for the Protection of Vertebrates Animal Used for Experimental and other Scientific Purposes 2010/63/EU) and the Animal (Scientific Procedures) Act 1986 are based on this ethical principle. The primary objective of the legislation is to protect animals and guide animal research by providing a code of practice for the housing and care of laboratory animals and developing a series of standard and minimum requirements

(Croney and Millman, 2007). Similarly, several laws have been created internationally to protect animals used in research and to protect researchers who are conducting experiments by giving them a framework for studying animals and justify their work to others (e.g. the general public, ethical committees). In addition to laws which protect animals, Three main ethical principles were developed over 50 years ago to provide a legislative framework (Hau and Schapiro, 2002). These three principals were produced to underpin 'Humane' animal experimentation and are Reduction, Refinement and Replacement (Russell *et al.*, 1959). Replacement refers to the substitution of a conscious living vertebrate for a non-sentient material (e.g. microorganism, tissues, and higher plants) when it is possible. When it is not possible to replace the animals used in experiments, then, reducing the number of animals should be considered (i.e. power calculation, adapting the experimental design to reducing variability). Refinement states that if reduction and replacement are not possible, the refining of the experimental studies to reduce to an absolute minimum the distress caused is paramount (Russell *et al.*, 1959). One of the first relevant changes in legislation protecting animals gave them the status of "sentient beings" (an individual who perceived and respond to sensation such as pain) through the Treaty of Amsterdam in 1997 (European Union, 1997). However, the first piece of legislation in relation to the protection of animals in research was development in 1986 with the introduction of Animal Scientific Procedure Act.

However, this utilitarian approach has a disadvantage as the weighting of harms and benefits is not usually straightforward as they have different currencies that are difficult to equate. As a general rule, harms are quantifiable in the animals used, but benefits are quantifiable in both, humans and animals which in some cases can be unknown. For example, for the quantification of harms, score systems to assess pain in laboratory animals are used. However any positive experience of the animals are often unquantifiable because they are yet to be determined (Vieira de Castro and Olsson, 2015).

In relation to laboratory animal welfare, the utilitarian along with the contractarianism approaches are usually used. This mixed approach consists of the implementation of absolute limits (contractarianism) and some weighting of the harm-benefit assessment (utilitarianism) (Perry, 2007). The Principles of Humane Experimental Techniques was a landmark publication regarding laboratory animal welfare (Russell *et al.*, 1959). This publication used this mixed approach as a starting point about ethics and experimentation in laboratory animals. The advantage of this approach in laboratory animals is that the contractarianism theory defines the end of the experiment taking into account endpoints or final limits whereas the Utilitarian theory uses the harm-benefit analysis. This utilitarian approach supports the use of animals in research only if the benefits of the research outweigh the harm to the animals (Singer, 1995; Francione and Garner, 2010; Vieira de Castro and Olsson, 2015). A harm-benefit analysis is carried out prior to experimental research to assess if it is ethically acceptable (Francione and Garner, 2010; Vieira de Castro and Olsson, 2015). In laboratory animals, it is essential to analyse the potential benefit of the scientific research taking into account the potential benefits/outcome of the model to be studied (e.g. animal, mathematical, cellular) and its validity and then to weight them against the potential harms done to the animals.

Refinement is particularly relevant to this thesis as it is directly related to welfare and is the focus of this research. Refinement in the scope of the 3Rs refers to the inclusion of methods that minimise pain, suffering, distress or lasting harm that may be experienced for the animals used in research (Russell *et al.*, 1959). Practically, this principle can be applied in laboratory animals, for example, when appropriate analgesia and anaesthesia are provided during scientific procedures (Tannenbaum and Bennett, 2015) to reduce the severity of that procedure. Another example provided by the Association for Assessment and Accreditation of Laboratory Animal Care (AAALAC) involves the modification of husbandry or experimental procedures that enhance animal welfare (e.g. pain assessment score sheets, post-surgical care) (AAALAC, 2011). However, one of the main requirements for refinement in laboratory animal research is the recognition of animal welfare status and the identification of factors that influence welfare, including legislation, ethics and context in which the animal is being used.

1.3. Factors that influence the definition of laboratory animal welfare

As in other species, laboratory animal welfare is closely influenced by legislation, ethics and the context in which the animal is used, i.e. farm, laboratory, zoo companion or wild animal. Ethical principles in laboratory animals are the driving force behind the creation of legislation. In addition to laws and ethics, appropriate measurement tools to assess welfare are crucial to allow informed decisions on how such animals should be treated when used in research (Fraser *et al.*, 1997).

In the UK, the legislation relevant to laboratory animals comprises the EU Directive for the Protection of Vertebrate Animal Used for Experimental and other Scientific Purposes 2010/63/EU, 2013 and the Animals (Scientific Procedures) Act 1986 (ASPA), which provide rigorous protection for laboratory animals (Wells, 2011). In other parts of the world, countries such as the United States of America, Korea, Japan, China and Mexico similar legislation have been created. Controversially, the protection provided to laboratory animals by the Animal Welfare Act (1966) in the United States which has not been updated, excludes laboratory rodents and birds (among other species) from protection. However, this gap in protection is to some degree bridged by Federal laws and guidance such as Public Health Service and National Institutes of Health which offer protection to all species used for human health research (Mukerjee, 1997). Despite the differences in legislation between countries, the laws protecting laboratory animals act as guidelines defining the direction and the minimum standards that we need to adhere to and offer some protection for laboratory animal welfare.

Ethics in laboratory animals are applied when there is a conflict between the value of the research to humans against the harm to the animals involved, which is common in science as laboratory animals are used for the benefit of humans. In this case, the potential pain, distress and suffering caused to each animal must be considered. However, even if the potential pain and distress of each animal are considered, if the benefits of the community outweigh the suffering of the individual, it is still considered to be acceptable (Donnelley, 1989). Although the general ethical guidance accepts the shared goals over individual goals, scientists need to be able to design protocols, taking into account the individual animals used, the level of animal suffering and scientific value of the study (Donnelley, 1989).

In order to maintain a high level of welfare in laboratory animals, we must be able to assess it. The assessment of welfare in a laboratory context requires to be measured by the use of validated indicators. These indicators are closely related to people's experiences and expertise. Therefore, the assessment method chosen needs to be reliable when used by different people with different backgrounds in the same context. Also, special attention should be given to ensure that all components of welfare are assessed and not only one animal welfare component (e.g. veterinarians assessing welfare including only physiological measurement and underestimated environmental factors).

The context and the assessor's background also have an impact as these two factors can affect the indicators chosen for the assessment of welfare. For example, scientists tend to select indicators which represent biological concepts (e.g. stress, immune competence, fitness) such as cortisol levels and disease incidence (Stafleu *et al.*, 1992) whereas the general public adopts a more interdisciplinary definition involving not only the biological function but health, natural history and subjective experiences (Hagelin *et al.*, 2003). Furthermore, a survey carried out in 2014 for Ipsos MORI about the attitudes towards animal research showed that animal welfare is an essential concern for the government, researchers and the general public. This ethical concern is based upon people's beliefs that animals consciously experience suffering comparable to that of humans (Dawkins, 1998) and ensures that laboratory animal welfare remains a high priority (Dawkins, 1998). However, opinions are not necessarily straightforward, for example, a survey carried out in Denmark (2012) show divided results. In this study, three different opinions were obtained: the approvers (30%), the disapprovers (20%) and the reserved (50%) (Lund *et al.*, 2014). The reserved group decided not to approve or reject animal experimentation until they saw each particular case and weight both the human and animal harm-benefit to assess each specific case (Lund *et al.*, 2014). This attitude towards animal research evidences the importance of the public opinion about the welfare of these animals, deciding on animal research based on the harm-benefit analysis.

It is clear that these definitions are intimately related to an individual's background and interaction with animals, along with how ethics and society have shaped their experiences (Broom, 2011). Furthermore, the context of laboratory animals is highly relevant when it is compared with the same animal, living in the outside world, as it seems to define how these animals are treated. As Moore (2001) wrote "*when mice are part of the everyday life in people houses the use of a painful and chemical substance to exterminate them (anticoagulants) seems to be justifiable because they are considered a plague. However, when they are part of medical research that it is made for human interests in laboratory facilities, we give them rights".* As Moore states, in the laboratory context, the mouse might play a role in improving human welfare and therefore the regulation and ethics around this context need to be very high compared animals used in other contexts.

To conclude, the definition of welfare for laboratory animals should not be different from any other animal. The basics that underline that definition are the same (biological functioning, physiological well-being and adequate environment).

However, additional inputs related to legislation, ethics and experimental conditions need to be taken into account. After the definition of laboratory mouse welfare is provided, the means of assessment of this welfare needs to be defined, taking into consideration additional factors that are determinant for the assessment including environmental conditions and the specific experiments that are performed on these animals.

1.4. Why laboratory mouse welfare is important

Mice are the most common species used in scientific research. An estimated 4.6 million mice are used annually in regulated research globally (based on an estimate from Taylor *et al.* (2008). Of these, the UK used 3.08 million of mice (Home Office, 2013), Europe used ~7 million mice (European Commission, 2013) (European Commission 2013) and USA used ~7 million mice (based on a conservative estimate from (Animal and Plant Health Inspection Service, 2015)) annually. As a consequence of the number of mice used, their welfare should be a priority. Over 3 million scientific procedures involving animals were carried out in the UK during 2017, 89% of which involved mice (Home Office, 2018). A scientific procedure is defined as an animal experiment that could potentially cause pain, distress, suffering or lasting harm to protected animals (Hollands, 1986a) . This high proportion of mice used in research is likely to be due to the relatively simple housing conditions needed (Lane-Petter, 1966), the ability to study almost every aspect of mammalian physiology, the isolation of single genes in mice, and their genetic similarity to humans. These features make them suitable for modelling a range of human diseases across a range of disciplines (Matfield, 1991; Asif T. Chinwalla et al. , 2002; Vieira de Castro and Olsson, 2015).

1.5. Factors that influence the welfare assessment of laboratory mice

The two main factors that influence the assessment of welfare in laboratory mice are the type of study that it is enrolled in (i.e. the experimental procedures that are carried out) and the environment (housing and husbandry) in which it lives.

In a laboratory animal context, mice may undergo experimental procedures where pain, distress or discomfort (i.e. their welfare is comprised) can be a very significant component of the studies. Therefore, the assessment of their welfare should be carried out taking into account harm-benefit analysis on a case by case basis, if the decision of treatment for alleviating the welfare problem is raised (Griffin *et al.*, 2014). These animals required special attention in terms of welfare, not only because of the number of animals used but also because of the strict regulation that protects them, the moral duty of such research and the high importance that they have for human health.

The refinement of the studies, to reduce pain and discomfort (i.e. improve welfare) to the absolute minimum is paramount for laboratory mice (Russell *et al.*, 1959). In order to alleviate pain or distress, we must first be able to recognise it effectively. The effective recognition of pain in laboratory animals is based on different indicators such as changes in behaviour, activity levels and knowledge of the physiological pathways of pain (Russell *et al.*, 1959). The provision of analgesia is close relate to the ability to recognising pain and it should be based on individual assessment considering the specific procedures that have been carried out (Flecknell, 2002). Historically, animals have not received analgesia following surgical procedures, indicating a possible lack of appropriate assessment of pain affecting the welfare of the animal (Richardson and Flecknell, 2005). Richardson and Flecknell (2005), carried out a systematic review of published, peer-reviewed, scientific papers where animals had undergone a surgical procedure. Provision of analgesia was recorded during two periods (1990-1992 and 2000-2002). Only 3% of the papers published from 1990 to 1992 report some usage of analgesia. This figure increased to 20% for the papers published from 2000 to 2002. For papers in which no analgesic use was reported, authors were contacted via email to inquire about the reason for the lack of usage/reporting of analgesic treatments, i.e. was this due to lack of reporting or were no analgesics given. The majority of people (71%) did not report any analgesia regimen because they did not use any. The main reasons for the lack of analgesia were related with the absence of signs of pain observed by the researchers (35%) and that the analgesics were considered unnecessary (35%) (Richardson and Flecknell, 2005). This study highlighted the lack of appropriate use of analgesia in laboratory mice, explained by the inability of researchers to recognise pain in these animals.

1.5.1. The effect of environment on laboratory mouse welfare

The environment in which laboratory mice are housed also has a considerable impact on their welfare. Husbandry and housing are considered to be a very important part of laboratory mouse welfare as they together constitute their main experiences, as mice like all laboratory animals spend most of their life simply being housed (Fox *et al.*, 2006). In laboratory facilities husbandry refers to all standard 'care' procedures that are carried out including handling, feeding, watering and cleaning cages. These standard procedures can be stressful (e.g. handling, cage cleaning), thus affecting animal welfare, particularly when they are performed by inexperienced staff (Meaney *et al.*, 1996; Balcombe *et al.*, 2004) and has for example been shown to cause a reduction in immune status (Sharp *et al.*, 2003). Environmental factors are often regulated by legislation, for example in the UK by the Animals (Scientific Procedures) Act and its code of practice for the housing and care of rodents (Home Office, 2014a). These documents provide guidelines for the adequate care of these animals taking into account different factors such as the behaviour of the animals in alignment with the 3Rs.

Environmental factors include a wide range of conditions and procedures such as; handling as part of routine husbandry, cleaning cages, housing practices, cage complexity, and environmental conditions in the cages, rooms and facility.

Handling is likely to be the most common procedure experienced by laboratory mice, as it an integral part of all most all the other procedures carried out on them. However, there are no specific rules or instructions related to handling mice in laboratory facilities; new research has suggested that the standard method of handling mice by the tail is stressful, increasing anxiety levels (Hurst and West, 2010) and affecting the results of behavioural experiments (Clarkson *et al.*, 2018). Traditionally, animals have been handled by tail as it considered to be a practical and rapid technique, but the animal welfare implications have only recently been assessed. New research suggests that the handling methods in laboratory mice need to refine as mice that are handled by a tube or cupped showed low anxiety levels comparing with tail handling (Hurst and West, 2010).

Another very common husbandry procedure that affects laboratory mouse welfare is the cleaning of cages, which is necessary for all confined animals to remove soiled bedding, preventing behavioural and health problems (Balcombe, 2006). The frequency of cleaning cages can have an impact on welfare affecting the animal's communication and social behaviour (Hurst and West, 2010). The information about the adequate frequency is contradictory as decreasing the frequency of the cage changes (e.g. every two weeks) can have benefits as a reduction of the incidence of handling and environmental disturbance (Baumans, 2005a). However, it can also be a problem, affecting physical welfare by increasing ammonia levels in the cage (Reeb-Whitaker *et al.*, 2001).

Another environmental factor that potentially affects the welfare of mice is the home cage. The Home Office Code of Practice for the Housing and Care of Animals for Scientific Purposes give recommendations for minimum cage size, cage distribution and complexity, bedding material, nesting material and stocking density. Small cages for housing laboratory mice can negatively affect their welfare, reducing normal behaviour and resulting in altered brain functions (Harvey and Chevins, 1985; Braastad, 1998; Meek *et al.*, 2001). A study published by Würbel (2001) contributed to the understanding of the importance of housing conditions especially in behavioural neuroscience, demonstrating that inadequate housing in early life interferes with brain development, behavioural response rules and disruption of environmental adaptations. As housing conditions for mice are not enforced by law, but they are part of recommendations in the code of practice, inadequate housing may occur, leading to impacts on behaviour and welfare (Würbel, 2001). These housing conditions are closely related with cage density (the number of animals per unit space) affecting mouse behaviour when they are housed in inadequate social groups (e.g. males from different litters) (Lawlor, 1983; AAALAC, 2011). Groups should be stable and harmonious, with barriers and hiding places provided to minimise aggression (Van de Weerd and Baumans, 1995; Van Loo *et al.*, 2002). As it has been shown that placing incompatible individuals (e.g. aggressive) or altering the social group causing confrontations which result in fights and bite wounds which can compromise animal welfare (Blom *et al.*, 1992; Champy *et al.*, 2004; Deacon, 2006).

Furthermore, this social alteration can be reflected in the health of the animal, causing stress and a reduction in body weight (Haemisch and Gärtner, 1994). Single housing can also lead to the manifestation of abnormal behaviours because of the alteration of social distribution. The term 'stress isolation' (or isolation syndrome) is described as a group of abnormal behaviours that are present when social conditions are not fulfilled. These behaviours include aggression, stereotypies, nervousness, and handling difficulties (Nevison *et al.*, 1999; Würbel, 2001). Single-housed mice show behavioural effects such as anxiety-like behaviours (Berry *et al.*, 2012) and stereotypes [(define as a behaviour pattern which is irregular, repetitive and has no apparent function: (Mason, 1991)], low body weight, low body fat and reduction in bone mineral density which is detrimental for the health of the animal (Nagy *et al.*, 2002). In addition, long periods of social isolation (3 weeks) can produce a reduction in the performance of behavioural tests (active and exploration tests), increase anxiety-like behaviour and reduce performance in the conditioning test (Berry *et al.*, 2012). These results confirm that the significant disadvantages in terms of behavioural studies, normal behaviour and physical health for laboratory mice when they are singly housed (Voikar *et al.*, 2005).

The content of the cage or cage complexity (e.g. presence of nest box, nesting material) can also impact on welfare. A more complex environment can improve welfare and lead to increased neuronal activity (Turner *et al.*, 2002), decreased disease progression (Hockly *et al.*, 2002) and improve positive behaviours such as exploration (Leach *et al.*, 2000) and play (Marashi *et al.*, 2003). Preference studies have demonstrated that mice select to spend time in more complex environments that include nesting material, rather than more barren environments, even if they must work to get access to those environments (Blom *et al.*, 1992; Van de Weerd *et al.*, 1998a; Fraser and Nicol, 2011). These studies show that an adequate complex cage is essential for the welfare of the laboratory mice.

The presence of bedding material (substrate for the base of the cage used for covering the acrylic floor) where animals can perform natural behaviours such as exploring and foraging is required in the home cage for maintaining animal's health and welfare (Balcombe, 2006). Mice prefer bedding material that they can manipulate which consists of large fibrous particles, rather than small particles that can also cause respiratory problems (Blom *et al.*, 1996a). Nest building material should also be provided in the home cage (Balcombe, 2006; AAALAC, 2011). Nesting material has changed from an accessory used in the cages to enrich the environment and forms a mandatory part of the housing for laboratory mice (AAALAC, 2011; Home Office, 2014a). Nesting material is important as it is considered a critical part of the environment providing the resources for modifying their environment (Baumans, 2005a) and helping to the animals to cope with the cold temperatures in the laboratory facilities. It appears that the temperature we routinely keep mice (i.e. 20-24^{o}C) is below that of their optimal preferred temperature of 30^{o}C (Gaskill *et al.*, 2009). A comparison of temperatures inside and outside the nest have demonstrated the benefits of the provision of nesting material allowed to control temperature within the cage specifically within the nest (26^{o}C and 29^{o}C) (Gaskill *et al.*, 2009; Gaskill *et* *al.*, 2013a) and also can improved breeding (Van de Weerd *et al.*, 1997b; Van de Weerd *et al.*, 1998a).

One of the most common methods for improving laboratory mouse welfare is environmental enrichment. This refers to any changes in the housing system that increases the number and diversity of positive natural behaviours (Olsson and Dahlborn, 2002). The benefits of environmental enrichment for animal welfare includes decreasing of abnormal behaviours (Van Praag *et al.*, 2000; Baumans, 2005a) and improving animal's ability to cope with living in a confined environment (Broom, 1993; Leach *et al.*, 2000). The code of practice for housing and care of animals for scientific proposes states that the environment should have opportunities for performing physical exercise, foraging, and cognitive activities (activities that engage animals cognition) appropriate for each species (Baumans, 2005b). This enrichment needs to be valuable for the animal, cost-effective to produce, easy to clean and maintain, should not be a risk to the animal's health, and should not affect animal husbandry, procedures and experimental protocols (Scharmann, 1991). Examples for mice include cardboard tubes (Heizmann *et al.*, 1998), nest boxes (Van de Weerd *et al.*, 1998b), paper towels as nesting material (Sherwin, 1997), and wood chips (Blom *et al.*, 1996b). The provision of additional structures within the cage for allowing animals to perform natural behaviours (different materials for chewing and gnawing) and to modify their environment (bedding, nesting material, tubes, boxes) can be considered for improving housing conditions (Baumans, 2005a). Sensory enrichment including visual, tactile, auditory, taste and olfactory stimuli are also possible enrichment tools. Additional sources of food (i.e. seeds scattered into the bedding) can give the animal the opportunity to forage which is an essential component of mice evolutionary needs (Baumans, 2005a). It is important to underline that environmental enrichment is very important for laboratory mouse welfare, however, a single approach might not suitable as different strains of mice have different needs (Baumans, 2005a). Further research is needed concerning suitable enrichment for specific strains in laboratory mice.

Along with housing and husbandry, the environmental conditions (ventilation, humidity, lighting, and noise levels) in which mice are kept also affects their welfare (Home Office, 2014a). Mice are especially sensitive to changes in temperature and humidity due to their small body mass (AAALAC, 2011). Mouse home rooms should maintain a temperature between 20ºC and 24ºC with humidity between 45 and 65%

(Home Office, 2014a). If the environment does not provide natural light, it should be controlled and consistent to represent, as far as possible, natural conditions (recommended 12:12 hour light/dark cycle) (Home Office, 2014a). Alteration in the light/dark cycle can affect the animal's welfare by affecting animal's circadian rhythm impacting on behaviour (Valentinuzzi *et al.*, 1997; Van der Meer *et al.*, 2004). Light intensity, either above or below normal levels, can additionally affect their health, including physiological changes (e.g. retinal damage) especially in albino mice (Wolfensohn and Lloyd, 2008). The adequate level recommended by the Home Office in the UK is between 350 to 400 lux on the bench (i.e. Room level) (Home Office, 2014a). At cage level, the recommended intensity of light should not exceed 60 lux especially for albino animals (Wolfensohn and Lloyd, 2008).

Other environmental conditions such as air quality and air changes, especially in Individually Ventilated Cages (IVCs) should be considered. The main propose of the ventilation in animal rooms or cages is to provide good quality air and to minimise the levels of gases, odours and infectious agents. When the quality of the air is compromised, it can have an impact on health, increasing the risk of respiratory diseases, so it is advisable to have 10 to 20 air changes per hour in a fully stocked room (Reeb-Whitaker *et al.*, 2001). If the ventilation in the cage is too high, this can also result in problems, as the odour cues that they used for communication can be affected, and the temperature can decrease considerably within the cage (David *et al.*, 2013). Humidity is another environmental factor that affects animals welfare by affecting their behaviour (Nelson *et al.*, 1990) and health by increasing the proliferation bacteria and the risk to respiratory infections (Reeb *et al.*, 1998; Chesler *et al.*, 2002). A mouse holding room should be kept between 40% and 60% of humidity (Home Office, 2014a).

1.5.2. The effect of experimental procedures on laboratory mouse welfare

Experimental procedures in laboratory mice include common practices carried out on the animals related to the experiments such as handling, dosing, sampling, anaesthesia, surgery, and euthanasia (Kaliste, 2004).

Handling, dosing and sampling are the most routinely used experimental procedures that laboratory mice undergo (Balcombe *et al.*, 2004). Handling and restraint of mice for simple procedures such as dosing or sampling can cause stress and anxiety if it is not done competently (Balcombe *et al.*, 2004). These activities can also cause physiological changes including increased heart rate and blood pressure (Sharp *et al.*, 2003). Dosing is also a common practice in laboratory mice and so the selection of the appropriate techniques (i.e. routine and volume of administration) and dosing site as well as the appropriate equipment (e.g. adequate needle and syringe for the species) can help to reduce this welfare effects in the animals (Hem *et al.*, 1998). Staff training is important for reducing the possible welfare implication related to stress, anxiety and health problems (Kaliste, 2004). Sampling, such as blood collection, can also have welfare and physiological implications, such as increasing glucocorticoids levels (Tuli *et al.*, 1995) and stress linked with handling, restraint and pain (Balcombe *et al.*, 2004).

Some experimental procedures require the use of anaesthesia and/or surgery. The term anaesthesia means the 'loss of sensation' which can involve the loss of consciousness (general anaesthesia) or the loss of sensation to a specific part of the body (local anaesthesia) (Stokes *et al.*, 2009). The administration of anaesthetic agents requires special training, including the adequate selection of the anaesthetic regimen, doses and monitoring of the anaesthesia and recovery. Adequate anaesthetic management for the experimental procedures improves animal welfare and the validity of the results collected (Arras *et al.*, 2001) by reducing secondary effects, such as gastrointestinal problems (Richardson and Flecknell, 2005). In order to carry out surgical procedures, specific training is required, which encompasses not only surgical skills and the intraoperative care but detailed knowledge of pre-operative and post-operative care for the animals. This includes elements such as the provision of control warmed temperature, analgesia, adequate or special diet if it is needed and close monitoring of the animals.

Euthanasia in laboratory animals is considered an important procedure and has the potential to have a considerable impact on welfare if carried out in an inhumane manner as almost all laboratory animals will be euthanised. As a general rule, an adequate euthanasia procedure must be reliable and irreversible, induce loss of consciousness with no, or temporary signs, of pain, distress and anxiety (AAALAC, 2011).

1.6. The assessment of welfare in laboratory mice

Welfare assessment in laboratory mice has been carried out using different methods, indicators and techniques. A limited number of studies have focused on the selection of the most appropriated indicators for measuring laboratory mouse welfare (Mertens and Rülicke, 1999; van der Meer *et al.*, 2001b; Jegstrup *et al.*, 2003; Baumans, 2004; Leach *et al.*, 2006; Spangenberg and Keeling, 2015). Some of these have included the development of welfare protocols with indicators obtained from consultation with experts (Leach *et al.*, 2008a). For example, this study involved the assessment of laboratory mouse welfare using 119 measures, divided into two categories: animal-based and resource-based. Although the results of this study showed that the protocol developed assessed the welfare of the mice accurately using these two types of indicators, it also demonstrated that this was time-consuming to do, and it needs further research about what sample size would be most appropriate (Leach and Main, 2008). As a consequence, this protocol has not seen wide uptake by the laboratory animal industry. In contrast, the study of Spangenberg and Keeling (2015) used only one type of indicator for assessing laboratory mouse welfare. Their protocol was solely based only on the assessment of the animal-based indicators with no resource-based indicators used. This protocol was based on the Welfare Quality program, developed for farm animals and the study aimed to develop a protocol that can be used in the everyday welfare assessment through observation of the animals in the home cage and with no equipment or additional resources needed (Spangenberg and Keeling, 2015). The results showed some difficulties in identifying animal indicators that reflect thermal comfort and positive emotional states in laboratory mice. More recently, specific protocols have been developed for assessing welfare changes as a consequence of specific experimental procedures, particularly in relation to pain, suffering and distress. For example, the assessment of pain using grimace scales (Leach *et al.*, 2012), and telemetry for heart rate in post-laparotomy pain assessment (Arras *et al.*, 2007). In addition, special protocols have been developed in relation to the different strains of laboratory mice (Hawkins, 2002) and more specifically for genetically modified mice (Wells *et al.*, 2006b).

It is essential to know which aspects of laboratory mouse welfare need to be assessed and how they can be measured either at the groups or individual level. One of the first challenges that we face when assessing laboratory mouse welfare is related to the number of animals, as in a laboratory facility we are dealing with a large number of animals. From a practical perspective, the assessment of every individual seems to be unrealistic as it is not possible to score every single animal in the facility on a daily basis. However, it is important to underline that the animals that are being used in an experimental procedure should be assessed at individual level at least daily for the duration of the experimental study as they may require additional interventions due to the increased likelihood of their welfare being compromised. The use of individual indicators that assess laboratory mouse physical welfare and psychological state, known as animal-based indicators (Leach *et al.*, 2008a), is often time-consuming and are not practical for the assessment of big populations. From a practical perspective, a system of assessing the welfare of laboratory mice at group level, by selecting individuals as a sample of the population, or assessing factors that affect all animals such the environment within the laboratory seems more appropriate. This often includes assessment of environmental measures, which are referred to as resource-based indicators (Hurst and West, 2010) and together with the animal-based indicators they form the basis of an effective laboratory mouse welfare assessment system.

Taking into account the definition of laboratory mouse welfare proposed at the beginning of this chapter, welfare in laboratory mice has three main components or principles: biological functioning, psychological states and environment (which represents the ability to maintain evolutionary needs). The assessment of welfare of laboratory mice should be based on the definition, and all three components should be assessed accordingly, as these indicators are interrelated with one another, and the lack of one can have negative impacts on other. For example, a lack of environmental enrichment can negatively affect both psychological and physical welfare (Baumans, 2005a).

1.6.1. The assessment of the environment (Resource-based) as a laboratory mouse welfare indicator

The main objective of managing environmental factors which impact laboratory animals is to prevent biological problems and to do so, key indicators can be assessed (Appleby and Hughes, 1997). Even if these indicators are the main focus of many legislation programs and they are considered to be necessary for animal's welfare, they alone are not an adequate animal welfare predictor (e.g. environmental enrichment as tool for improving welfare can vary between strains) (Sørensen and Fraser, 2010). It has been said that these indicators do not have acceptable validity and reliability (Sørensen and Fraser, 2010). They are considered as risk factors and possible causes that affect welfare. Therefore, they are considered as part of an interaction that along with animal-based indications outlines animal's welfare

(Dawkins *et al.*, 2004). The process to choose which environmental conditions are assessed depends on the basic environmental principles such as housing, husbandry and environment.

The most suitable method for assessing these environmental indicators is the comparison of the current practices in the laboratories facilities against the recommended standards in the codes of practice (AAALAC, 2011; Hawkins *et al.*, 2011; Guillen, 2012; Home Office, 2014a). However, most of these recommendations are based on experience and general knowledge, and only a few of them based on scientific evidence. The fact that some of these recommendations have been historically set based on experience and not based on the actual preferences can be difficult as they may not necessarily reflect the animal's actual needs. However, more recent research is being carried out relating to different resource-based factors that affect animal welfare such as handling methods (Gouveia and Hurst, 2017), bedding and nesting material preferences (Heizmann *et al.*, 1998) and temperature (Gaskill *et al.*, 2009). Such specific research about environmental preferences in laboratory mice for nesting material and appropriate cage size is likely to improve welfare in most strains of mice (Olsson and Dahlborn, 2002).

1.6.2. The assessment of biological health (Animal-based) as a laboratory mouse welfare indicator

The indicators include in the biological assessment of laboratory mouse welfare are based on the expected normal physiological measurements for a given species, sex, and age of the animals and they are compared against the values obtained within the laboratory facility. The factors that affect laboratory mouse welfare including environment and experimental procedures can influence the biological health of the animals. These indicators function as an output as they represent the animal's response to the environment and therefore by assessing biological health of the animals we can assess welfare (Spangenberg and Keeling, 2015) (Leach *et al.*, 2008b). The biological health of the animals includes physical health and compromises environmental and experimental challenges. These changes are monitoring through veterinary care and observations by technical staff who care for the animals on a daily basis.

A comprehensive health monitoring program which includes veterinary care, the development and management of protocols during experiments (including pain management and euthanasia) is relevant for maintaining higher welfare (Home Office, 2014a). The laboratory care program should ensure that animals enrolled in studies are healthy as a requisite for good welfare and therefore good models for animal-based science. As a general rule, daily animal care observations are performed by trained laboratory personnel for recognising signs of disease, injury, distress and behavioural changes. Provision of appropriate treatment, when required, is also essential (AAALAC, 2011). One important example is pain and distress which may occur as a direct result of the research being carried out, e.g. following a surgical procedure or as a consequence of injury, i.e. fighting. It is widely accepted that animals experience pain (Fraser *et al.*, 1997; Gillingham *et al.*, 2001; Sherwin, 2001; Flecknell, 2002; Flecknell and Roughan, 2004; Guillen, 2012) and as it is considered a stressor, it is imperative from an ethical, professional and legal perspective that adequate pain management is provided (Home Office, 2014a). Pain recognition and alleviation using anaesthetics and analgesics should be relevant in all experimental studies and protocols to identify and alleviate pain and distress should be created in the laboratory and reviewed routinely (Bateson, 1991; Flecknell and Roughan, 2004).

Indicators such as the presence of wounds (Spangenberg and Keeling, 2015), body temperature (Vogel *et al.*, 2016), and Body Condition Score (Ullman-Culleré and Foltz, 1999) can be assessed. The assessment of these indicators is usually done at individual level, especially for animals that are included in experimental studies. However, when the assessment is done in the entire population as part of the daily checks of the animals, biological health indicators can be used at group level by observing the cage as a single unit. For example, cage odour and excessive urine and faeces in the cage are biological health indicators which assess animal’s health using the cage as a single unit for the assessment of the group of animals within the cage. Cage odour can be used to assesses the levels of ammonia in urine, as an indirect and basic measure of health problems such as kidney diseases (Ishii *et al.*, 1998) while excessive urine and faeces in the cage assess animal biological health as an indicator of digestive or urinary system disease (Spangenberg and Keeling, 2015).

1.6.3. The assessment of psychological state (animal-based) as a laboratory mouse welfare indicator

Psychological state is a component of welfare which uses animal-based indicators for the assessment. These indicators assess the psychological effect that the environment and the experimental procedures have on the animal. Psychological state assessment involves determining the animal's ability to maintain their mental well-being and is usually measured through the animal's behaviour (Nevison *et al.*, 1999).

Despite laboratory animals being bred in controlled environments for many generations, animals still show behaviours that are present in their wild ancestors (e.g. nest building) (Nevison *et al.*, 1999; Gaskill *et al.*, 2013b). Internal mechanisms that govern those behaviours are the same for wild and laboratory animals (Dawkins, 1990), so laboratory animals are still being motivated to exhibit many of the behaviour patterns seen in their wild counterparts, even though they are no longer required for survival under captive conditions (Dawkins, 1990; Würbel *et al.*, 1996). Therefore, if such behaviours are not exhibited by animals, their welfare will be diminished, manifesting in behavioural disorders, stress and other pathological conditions (Dawkins, 1990). Behaviours such as exploration, foraging, nest building, grooming, and digging are part of an animal's normal behavioural repertoire (Dawkins, 1988; Baumans, 2004; Baumans, 2005a). Therefore, it is necessary to provide an environment which allows them to perform such behaviours to maintain a reasonable level of welfare (Baumans, 2005a).

The basic assessment of an animal psychological state is made through behavioural observations. Normal behaviour can be divided into three categories: behaviour that animals want to do, behaviour that they want to do but only when external situations require it, and behaviour they do not want to do (Dawkins, 2003). An animal's behaviour is a manifestation of how well it has been adapted to a specific situation or if they have failed to adapt by showing stress signals (Dawkins, 2003). However, if it is questionable whether behavioural indicators are sufficient to be able to recognise these signals or if additional measures are needed. Prior to using behavioural indicators to assess welfare, it is necessary to know the normal behaviour of the mice, which is achieved by developing an ethogram (normal behaviour catalogue) (Banks, 1982). The comparisons between behaviour perform in the natural living, and the captive environments are usually made when assessing welfare as behaviour can be observed and quantified and then compared with the ethogram, to determine if there are any significant changes (Potter and Broom, 1987). For example, mice usually built a nest in the wild to provide shelter from predators, retreat from aversive environmental conditions and to protect the young (Latham and Mason, 2004). Therefore, nest building behaviour is associated with survival in wild mice and laboratory mice it remains a highly motivated behaviour even if they have been removed from the wild for many generations. Consequently, the absence of a good quality nest, for example, could be an indicator of reduced welfare even in the laboratory environment (Gaskill *et al.*, 2013b). An alternative approach to assessing nest quality was proposed by Rock (2014). The time to integrate nest test (TINT) involves the assessment of time that mice spend building the nest following placement of new nesting material into the cage. It is considered successful if the mice integrate the material into the nest within 10 minutes. Mice with undergoing surgical procedures exhibited negative results in these tests showing a delay in nest building of two days in TINT compared with same mice before surgery (Rock *et al.*, 2014).

The assessment of animal welfare using behavioural indicators needs to take into account not only the type of behaviour that is performed but the function of the behaviour within the environment and the internal motivation of the behaviour (Appleby and Hughes, 1997). Behavioural indicators are divided into categories: normal behaviour suppression, behaviours that indicated negative emotional states (pain and distress), abnormal behaviours, and preference/aversions (Appleby and Hughes, 1997).

The first category is normal behaviour suppression occurs in preparation for flight, defence or hiding in response to a stressful situation (Broom, 1988). The aim of these changes is the replacement of the normal behaviour for behaviour more appropriate but if the aversive situation continues, or if the situation is intense this suppression of behaviour might continue even if the disturbance has ended (Appleby, 2011). Besides the basic normal behaviours that animals need to perform to ensure their survival (e.g. feeding, drinking, breeding) there are other behaviours that are not necessary to survive but are considered to be innate and animals “need” to perform them to guarantee their welfare such as nest building in laboratory mice (Gaskill *et al.*, 2013b).

The assessment of pain and distress can be measured through behaviour as well as physical indicators. These behaviours involved the ones that animals do not normally perform and include fleeing, hiding, immobility and distress signals (Fraser, 1989). Fleeing, hiding and immobility are present when animals are avoiding an object in a stressful situation. Welfare is thought to be poor if avoidance behaviour is strong when the situation or the object in question is present (Broom, 2007). In the case of distress signals can be observed in animals in behavioural patterns such as vocalizations (Portfors, 2007), food and water consumption, locomotion activity (Flecknell and Liles, 1991; Liles and Flecknell, 1992; Flecknell and Roughan, 2004) and more recently facial grimace in laboratory mice (Miller *et al.*, 2016). In addition, specific 'pain' behaviours are usually assessed taking into account the location of the pain (e.g. belly pressing in abdominal surgery) (Miller *et al.*, 2016).

The third category of behavioural indicators are the abnormal behaviours related to the inability of an individual to cope with environmental changes, abnormal behaviour such as stereotypes (Broom, 2007). Stereotypies are considered the main abnormal behaviour present in laboratory animals including mice (Mason *et al.*, 2007). The most common examples of stereotypes that have been observed in mice are bar gnawing and bar jumping (Nevison *et al.*, 1999). It is important to underline that it is not the behaviour itself that is considered abnormal, but the frequency and the context where the behaviour is performed (Mason, 1991). These abnormal behaviours can often involve individuals who perform activities that can lead to self-injuries or that produce injury in other conspecifics (Broom, 1988). However, these abnormal behaviours are not easy to interpret like in the case of stereotypes which are easier to identify when they lead to injuries (e.g. barbering in mice) but when there is no evidence of injuries it makes the interpretation about their welfare consequence more difficult. Also, it has been suggested that some of the repetitive behaviours are not necessarily bad for the animal as they could be a coping mechanism to reduce distress and anxiety (Mason, 1991). In laboratory mice, abnormal behaviour such as stereotypes, sudden fear or aggression, vocalisations, a decrease of grooming (leading to chromadacryorrhea-red secretions around the eye and nose) are considered as indicators of poor welfare (Baumans, 2005b). Factors that affect laboratory animal welfare and contribute to the development of abnormal behaviours are related with the environment such as restrictive housing conditions and which prevents animals from performing natural behaviours and from controlling their physical environment, which has been demonstrated to lead to a stress reaction

(Wiepkema and Koolhaas, 1993). This stress reaction is manifested by the appearance of abnormal behaviour patterns such as stereotypes and the suppression of normal behaviours (e.g. decreased general activity level). In the case of gnawing, for example, it is considered normal behaviour (used for food consumption), however, in a suboptimal environment, it may be exhibited both out of context (e.g. towards metal bars) but also at an abnormally high frequency (Leach *et al.*, 2000). It is possible that this behaviour could be an indicator of animals trying to escape from suboptimal environments (Würbel *et al.*, 1996).

The final category is represented by preference/aversion test and new indicators for assessing laboratory animal's psychological states. These indicators include those that have been used in humans which include qualitative judgment in animal behaviour (Stevenson-Hinde *et al.*, 1980). Tests of cognitive bias were developed in laboratory rodents, in the form of a novel cognitive task designed to measure biases in judgment when exposed to ambiguous cues (i.e. if the animal is optimistic or pessimistic; (Boissy *et al.*, 2007a; Bateson and Nettle, 2015) and preference tests which provided to the animals difference options between different environmental conditions (e.g. temperature, bedding materials) from which they are free to choose (Gaskill *et al.*, 2012). As suggested by Dawkins (2003), one of the issues in animal welfare research is to find indicators that can directly measure animal's welfare scientifically and objectively. Dawkins discusses that animals behaviour is the key element required to answer such question about welfare (Dawkins, 2003). This affirmation was supported by Patterson- Kane et al., (2000), who also agreed that behaviour really becomes a welfare tool by answering want animals want. They add that is not only important what they want but how much they want it. Preference testing can be used for assessing what laboratory animals prefer. For example, laboratory rats choose to be with conspecifics over isolation and will work to achieve this (e.g. press a lever may time) if needed (Patterson- Kane et al., 2000). Another example of preference testing in laboratory mice is related to environmental temperature (Gaskill *et al.*, 2009). In this study, mice were given free access to the different home cages of varying temperatures (20ºC, 25ºC and 30ºC), and the occurrence of maintenance, inactive and active behaviours where measured. The study found that mice preferred to spend time in the cages with high temperatures, especially female mice (Gaskill *et al.*, 2009). Aversion tests are also used for the assessment of welfare, particularly when we are trying to understand more about what housing and environmental conditions can compromise welfare. For example, the CO_2 as a method for euthanasia in laboratory rodents has been shown to be aversive based on aversion studies (Leach *et al.*, 2004).

1.7. The aggregation of indicators into a laboratory mouse welfare protocol

Protocols for welfare assessment of laboratory mice are routinely used to identify when animals have their welfare compromised (Hawkins *et al.*, 2011). These protocols should involve the aggregation of the three main aspects of animal welfare assessment (biological functioning, psychological states and environment). This allows interventions to be made such as revising animal husbandry, enrichment and humane endpoint decisions (Hawkins *et al.*, 2011). The protocol used needs to be valid (it measures what is intended to measure, in this case, laboratory mouse welfare), reliable (produces the same results when it is performed by different people at different times) and practical (not time-consuming and limited resources needed). Also, special protocols will need to be developed to include indicators tailored for a specific study when it impacts on a specific aspect of animal welfare (e.g. pain assessment in surgery studies (Wells *et al.*, 2006a)). An optimal welfare assessment protocol for mice will need to be dynamic and include indicators that account for the type of animals that are being assessed (e.g. stock or experimental), the size of the facility and the specific characteristics of the research.

As a standard, animal welfare assessment protocols should be designed taking into account animals individuality (e.g. species, sex, age, strain) (Hawkins *et al.*, 2011). They also should provide consistent information which is easy to interpret and analyse (Leach *et al.*, 2008a). Preferably, a protocol should include a combination of indicators from each domain of welfare (i.e. biological functioning, psychological states and environment) to enable a more comprehensive assessment of welfare. For example, for abdominal surgery in mice, behavioural indicators that assess pain such as hind leg behaviour, abdominal pressing or hunched posture should be included as a post-operative welfare assessment (Wright-Williams *et al.*, 2007). After selecting the most appropriate indicators, the optimal means of carrying out the assessment should be determined (i.e. observation of the animals/facility, questionnaire given to staff), which depends on the nature of the indicator (i.e. humidity is usually assessed by direct observation of humidity readings within the laboratory facilities) (Leach *et al.*, 2008a).

Traditionally, developments in the effective assessment of welfare have focused on farm animals, where the protocols used today for the assessment of welfare in different species are based on the five freedoms (Botreau *et al.*, 2007b; Blokhuis *et al.*, 2010), for example the Welfare Quality in poultry (Welfare Quality®, 2009b), cows (Welfare Quality®, 2009c) and pigs (Blokhuis *et al.*, 2010). The use of the five freedoms is considered effective in farm animals as it has been successful in identifying welfare problems and made suggestions (based on the results) for potential refinements to promote welfare, e.g. reducing lameness and injuries due to environmental problems on the farms) (Whay *et al.*, 2003). However, there are still concerns about the weighting of each welfare aspect in such protocols in terms of the overall assessment of welfare and the differentiation of acceptable or unacceptable welfare (Veissier *et al.*, 2011). In laboratory animals, there is not a standard protocol model for welfare assessment. However, the guidelines created by Morton and Griffiths (1985) have been used as a baseline for the development of protocols for these animals. It has been argued that welfare assessment in laboratory mice should include both resource-based and animal-based indicators which are used to cover the three main aspects of animal welfare (physical health, psychological state and environment) (Leach *et al.*, 2008a). However, the limited number of studies on this topic to date, many have included only animal-based indicators under the premise that those are sensitive enough for detecting any welfare problems, even when the environment negatively impacts welfare (Spangenberg and Keeling, 2015).

Moreover, animal-based indicators are key for the success of animal welfare protocols as they assess the animal's ability to cope with environmental changes and represent the animal's reaction to the aspects of the environment we can measure (i.e. resource-based measures). Spangenberg and Keeling (2015) suggest that there is no need for monitoring the environment on a regular basis as there is a specific code of practice and legislation dictating the 'environmental' aspects and the impact of these can be assessed using animal-based indicators. However, even if the vast majority of indicators are animal-based, the environment and other resources that we provide to the animals (e.g. husbandry) also need to be assessed to make sure the minimum standards that animal need related with the environment are always met.

A complete animal welfare assessment protocol which encompasses the three main components of welfare by including valid, reliable and practical indicators will allow us to provide a comprehensive assessment, recognising areas where welfare can be improved as well as good practices.

The welfare of laboratory mice has been shaped alongside the evolution of the definition of animal welfare and the legislation that governs people's actions in relation to laboratory animals. There is now a significant understanding of the need and importance of laboratory mouse welfare as illustrated by the development of codes of practice, which include indicators and protocols for protecting animal welfare. More research is needed into the validity, practicability and reliability of different indicators as most of the normal values for the indicators used to date have been established by experience, but they have not been validated in experimental studies. Additional research about indicators that assess the psychological states of mice as one of the foundations of animal welfare assessment needs to be developed. Novel indicators for the assessment of animals psychological states has been developed recently especially in farm animals. Qualitative Behavioural Assessment (QBA) has been used successfully for the assessment of positive emotions in farm animals and it has been included in the welfare quality program (Wemelsfelder, 1997; Council, 2009a; Welfare Quality®, 2009c). This indicator uses the animal's subjective experience along with behaviour and integrates them in a dynamic and communicative whole agent used for expressing the quality of their own experience (Wemelsfelder, 1997).

1.8. Objectives

The main objectives of this thesis are to:

1. Develop a protocol that includes quantitative and qualitative mouse welfare indicators through scientific inquiry and expert opinion that will allow those who care for (i.e. technicians & veterinarians), regulate (i.e. HO inspectors) and use (i.e. scientists) animals in research to assess the welfare of laboratory mice practically and reliable.
2. Investigate whether a novel welfare assessment technique; Qualitative Behavioural Assessment, may offer an effective and practical means of assessing laboratory mouse welfare states.
3. Apply the developed protocol to assess welfare in a laboratory animal facility to test its reliability and practicability.

Chapter 2. Validation of mouse welfare indicators: a Delphi consultation survey

2.1. Introduction

Mice are the most commonly used species of laboratory animal worldwide (Ormandy *et al.*, 2009). Due to the large number of mice used in research (over 1 million in the UK during 2017) (Home Office, 2018), the refinement of their welfare is critical. However, this refinement is dependent on our ability to efficiently assess their welfare, as without this we cannot identify instances when a refinement is needed or if any refinement applied has been effective. Consequently, there has been increasing interest in developing new methods to effectively assess the welfare of mice at both individual and group level within animal facilities (Branchi *et al.*, 2001; Proctor and Carder, 2014).

The welfare of laboratory mice is routinely assessed using a combination of animal-based (e.g. physiological, psychological changes) or resource-based (e.g. environmental conditions, staff training) indicators, along with relevant procedure linked indicators(e.g. pain management). Resource-based assessment is carried out using indicators that given information about the environment that we provide to the animals (Dawkins, 1990; Broom, 1996; Fraser and Broom, 1997; Hubrecht, 2014). These indicators include environmental indicators relating to animal housing (e.g. temperature, humidity) as well as everyday husbandry activities (e.g. handling, cleaning cages) (Baumans, 2005b). Animal-based assessment involves the measurement of animal's behaviour and physiological reactions as they reflect how animals cope with the environmental changes, preserving their biological and psychological functions. The aggregation of all the aspects of laboratory mouse welfare (i.e. based on overarching definition in Chapter 1: biological function, psychological state and environment) into a welfare protocol, is paramount to provide an overall assessment. There have been a number of studies gathering information from experts about appropriate indicators and methods for assessing animal welfare. However, the majority focus on species other than mice, including cows (Geist, 2010), horses (Collins *et al.*, 2009), pigs (Bracke, 2006), laying hens (Whaytt *et al.*,

2003) and their main focus was developing policies and recommendations for welfare (Bennett *et al.*, 2004; Rikkonen, 2008; More *et al.*, 2010). For example, the Delphi consultation for obtaining information about the development of policies for agriculture in Ireland (More *et al.*, 2010) or the expert's opinion study about environmental enrichment in pigs (Bracke, 2006).

The Delphi consultation is a widely used survey technique that seeks information from experts about a specific topic (Adler and Ziglio, 1996; Collins *et al.*, 2009; Keeney *et al.*, 2010). The questions are presented anonymously, through a series of rounds with the aim of achieving consensus within the group (Linstone and Turoff, 2002). This methodology has been used in a diverse range of animal science fields including the assessment of the impact of DEFRA policy on welfare (Bennett *et al.*, 2004); the implication of animals diseases on productivity (Der Fels-Klerx *et al.*, 2002); and for the selection of a subset of species to have their habitat protected (Hess and King, 2002). This technique has also shown to be an effective method of gaining information about welfare assessment in farm animals (Whaytt *et al.*, 2003; Bracke, 2006; Collins *et al.*, 2009). The Delphi methodology is used as a preliminary source for assessing 'face' validity, which is defined as the subjective opinion of experts about the extent to which the measure is meaningful in terms of providing information on the animal's welfare (Nevo, 1985; Sireci, 1998; Blokhuis, 2013). This face validity is based on the assumption of "safety in numbers" where a group of people are less probable to come to a wrong conclusion than an individual (Hasson *et al.*, 2000).

This study aimed to determine, through a modified Delphi consultation, which indicators of mouse welfare are considered valid, reliable and practical to use in laboratory facilities as part of either an audit of welfare by an independent assessor or the daily assessment of welfare by facility staff. This study used the Delphi consultation technique as a tool to identify potential measures for assessing mouse welfare to be used in practice in later studies. In this consultation, a level of 70% agreement for global consensus (i.e. across all of the indicators) and over 60% agreement for individual indicator consensus was required. There are no specific guidelines for defining a consensus in Delphi studies, as it is argued that it is dependent on the nature of the research (e.g. medical decision, development of new policies) being carried out (Keeney *et al.*, 2006). Studies using this technique in nursing and animal welfare contexts have used a level of 70% consensus as a standard (McKenna, 1994; McKenna *et al.*, 2001; McKenna and Hasson, 2002; Leach *et al.*, 2008a). Many studies interestingly do not provide any information about the level of consensus needed (Hess and King, 2002; Bennett *et al.*, 2004; Collins *et al.*, 2009) or if they are required to have 100% of agreement (Der Fels-Klerx *et al.*, 2002), Since there are no guidelines to set a consensus level in Delphi studies a level of 70% was used as this is aligned with other peer-reviewed, animal welfare research (More *et al.*, 2010; Wentholt *et al.*, 2012).

2.2. Overview of Delphi methodology

A Delphi consultation process was conducted to determine, through expert opinion, the most valid, reliable and practical indicators to be used in two specific scenarios for assessing laboratory mouse welfare; [1] an audit welfare assessment of a laboratory facility and [2] an everyday assessment of laboratory mice welfare. This study was comprised of four distinct sequential phases; [1] a scoping meeting, [2] a pilot survey, [3] Round one of Delphi consultation, and [4] Round two of Delphi consultation (Figure 2.1).

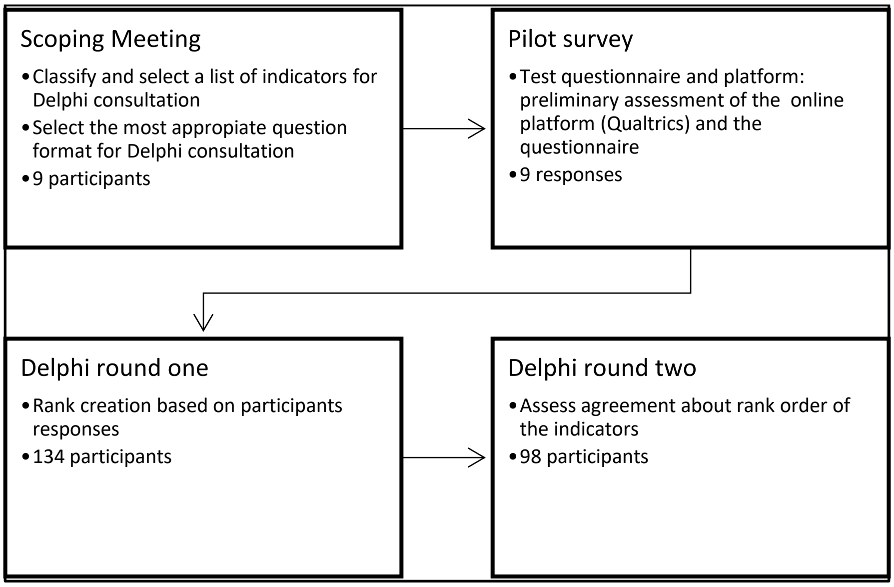


Figure 2.1. Delphi consultation process used in this study to obtain information from experts about laboratory animal welfare indicators.

2.3. Scoping meeting and pilot survey

2.3.1. Methodology

Participants for this scoping meeting were recruited via email. Information about the purpose of the meeting was sent to Newcastle University animal laboratory facilities to invite people who were working with laboratory animals in different roles (e.g.

technicians, researches, managers, facility directors, PhD students)to participated in the meeting. a total of 9 people agreed to participate in the meeting. The scoping meeting was divided into two sessions. In session one, participants were asked to generate as many indicators of mouse welfare as possible, considering their validity, reliability and practicability for the assessment of mouse welfare in the two specific contexts (Table 2.1). In session two, the groups were asked to rank the quality of a list of potential questions (Appendix A) that could be used in the first round of Delphi consultation. The aim of the scoping meeting was also to ensure that the wording in the questions was clear to guarantee the outcome the consultation will meet the aims. The pilot survey aimed to provide feedback on the type of questions, the length and the information contained within the questionnaire and was used to refine the questions and survey format for the first round of the Delphi consultation (Appendix B).

Theoretical scenario	Description
Audit welfare assessment	You are about to assess the welfare of conventional ‘stock' laboratory mice over audit (4 hours) in a laboratory animal facility that houses approximately 4,000 mice. The mice are currently being housed and are undergoing standard routine daily checks, but not being used in any procedure. They are housed in 7 animal rooms, each holding 4 racks, with each rack holding 50 cages and each cage housing between 1-4 mice (depending on sex, strain and age). The laboratory has a medium biological security barrier (the use of masks, overshoes and uniform are required to enter the animal rooms). You will have portable and calibrated monitoring equipment available, for example, thermometer, humidity gauge, light intensity meter etc.
Everyday welfare assessment	You are about to assess the welfare of conventional ‘stock' laboratory mice as part of your daily checks of the animals under your care. The mice are currently being housed but are not being used in any procedures. The laboratory has a medium biological security barrier (the use of mask, overshoes and uniform are required to enter animal rooms). You will have portable and calibrated monitoring equipment available, for example, thermometer, humidity gauge, light intensity meter etc.

Table 2.1. Theoretical scenarios used in the Delphi consultation

This pilot was launched using the online system Qualtrics platform (http://www.qualtrics.com) and was live for 2 weeks (December 2015) with 9 participants completing it. Participants were asked to complete the survey and assess the type of questions, their clarity and to indicate the amount of time needed to complete the questionnaire.

2.3.2. Scoping meeting results

The scoping meeting session one included 9 experts in laboratory animal science including a laboratory animal veterinarian, 2 laboratory animal technicians, 4 researchers that work regularly with laboratory animals, a licencing and regulatory administrator, and an animal unit manager. The main objective of this session was to generate a list of mouse welfare indicators that participants considered valid in a specific context (Table 2.1). The information about the indicators list provided by the 9 experts along with the justification of the selection and the means of measure the indicators is provided in Table 2.2.This list was considered for the final list of indicators used for the Delphi consultation.

Criterion	Justification	Measurement
Alertness	Health/physical indicator	Checklists, scales.
Response towards stimuli	Health/physical indicator	Behavioural assessment for each mice
Response towards people	Health/physical indicator	Behavioural assessment for each mice
Repetitive movements	Psychological indicator	Ethogram, observations
Bodyweight	Physical indicator	Scales
Social behaviour	Psychological indicator	Observation
House conditions (single, group)	No information provided	Observation
Coat condition	Health/physical indicator	Scales
Eye condition	Health/physical indicator	Scales, checklist
Cheek bulge	Health/physical indicator	Scales, checklist
Audible vocalisations	Psychological indicator	Automatic recordings
Balance test	Health/physical indicator	No information provided
Whiskers position	Health/physical indicator	Scales
Posture	Health/physical indicator	Scales
Ears position	Health/physical indicator	Scales
Body temperature	Physical and emotional indicators	Scan
Health indicators (secretions, injuries)	No information provided	Observation of Sentinel individuals
Mortality rate	Produce patterns over the time	Records
Environmental conditions	Effect on welfare directly	Observation
Bedding condition	Effect on welfare directly	Observation
Cage size	Effect on welfare directly	Observation
Enrichment	Effect on welfare directly	Observation
Food and water consumption	No information provided	Automatic measurement
Hormones test (hair, urine)	Short and long-term welfare indicators	No information provided

Table 2.2. List of mouse welfare indicators compiled by the participants in session one of the scoping meeting.

Participants were also asked about where gaps in knowledge and limitations for the effective assessment of laboratory mouse welfare. Table 2.3 provides the participant's option about the limitation for assessing laboratory mouse welfare. The limitations are mostly the validity and practicability of the indicators that are being used nowadays.

LIMITATION TO EFFECTIVE ASSESSMENT OF MOUSE WELFARE	
Time	The lack of time for making a complete welfare assessment
Expertise	The level of expertise of the person carries out the checks The lack of possibilities of train and educate the technicians Gaps in the expertise in the person doing the assessment Tools needed to measure sounds and ultrasonic range
Knowledge	The lack of scientific knowledge of some indicators. Scales and checklist to assess animal welfare: including their reliability and validity to assess animal welfare Communication between new research results and technicians: dissemination of the information
Research	Detailed information in terms of behaviour of individual strains Observation of behaviour in animals: repeatability, standardisation (e.g. how long do you need to watch for?) A sample size of animals to assess large animal welfare facilities: how many sentinels do you need? Which number is representative across the rack? How to pick up the number? Lack of research about sound pollution: how to measure it and their real effects in mice Lack of research about olfactory senses (e.g. effects of ammonium in mice). Communication by essences More research needed about social environment in mice Vocalisations: meaning in mice The importance of the experience of people doing the assessment Cleaning procedures in cages: how much do we know about their effects related to scents and communication Transgenic animals: difference in the strains/ages and, detail picture of the individual screen (phenotype) Air quality: what is appropriate, acceptable, differences in strains

Table 2.3. List of limitations to effectively assess welfare given by the participants of the scoping meeting

The objective of session 2 was to determine the most effective type of questions to ask in Delphi consultation. Participants analysed 6 pre-written questions for their potential effectiveness to gain further information on the topic (Table 2.4). Participants were asked to produce a rank order for the questions in terms of effectiveness, with a justification of the order chosen. Question 6 was deemed to be the most effective by the participants, and so was taken forward to be used as the main question in the first round of Delphi consultation with some modifications based on the participant’s comments.

Question	Participants comments	Final rank
Question 1. "For each specific measure of welfare listed below, please indicate (yes, no or maybe) whether you consider it to be appropriate (reliable, practical and valid) for assessing laboratory mouse welfare. Please provide a brief justification for your answer". The question has 59 animal welfare indicators divided into animal-based and resource-based in a table with three columns (indicator, classification, justification)".	Good question, but the justification needs to be changed to 'why' because justification needs an argument and why is a reason. People may put vague justifications otherwise. The justification each indicator can take too long so people maybe choose just a few indicators to avoid those long answers. Would everyone understand the terms (reliable, valid and practical)? Would everybody know the measurement terms? Takes too long to answer The question requires a lot of knowledge and a lot of thinking	3
Question 2. "Please list 20 measures that you consider to be appropriate for assessing laboratory mouse welfare and briefly describe why you think that they are important".	People might be able to write down not very good welfare indicators to complete the number as 20 might see a long number It seems to be more appropriate for an exam paper Inexperienced people may struggle because of the high number of criteria needed Would be worth it to add categories to the questions such as an animal-based indicator or environmental-based to give some guidance It doesn't provide any information about practicality and reliability of the measure Very similar to question two and people can provide very general (e.g. health) or very specific (e.g. Piloerection) measures.	5
Question 3. "In your opinion, which are the most appropriate measures that should be used to assess mouse welfare within a laboratory animal facility?"	Difficult to answer if the person does not have enough experience List the measures and then rate them based on practicality and reliability This was considered a good question	6
Question 4. "From the list of measures below please select 10 that you think is the most appropriate for effectively assessing the welfare of mice in laboratory animal facility". The question has 59 indicators and a small chart to add the 10 indicators selected.	Add bite wounds The 10 measures in the table can be organised into categories (e.g., animal-based, behaviours) to make it clearer. Good prompt. Many people could provide useful data. Change some terminology to make them more acceptable (based on knowledge level and language). It doesn't have any information about practicability and feasibility of the measure. The order or the list might influence choice, so randomise list or organised with heading (e.g. animal, environmental).	2

Question 5. "Please list 10 animal-based and 10 environment-based measurements of welfare that you would use to assess the mouse welfare in laboratory animal facilities. Please justify why you considered these to be appropriate".	Ranking the measure in terms of importance within the categories to see how important people think they actually are. Limited by knowledge-base from the participants. Ten measurements from each category might be difficult for junior staff. The number should say up to ten animal welfare measurements because if people don't know they can just mention general stuff. No sense about which is more important (environmental vs animal). People may write add extras that they are not considered important just to make up numbers. Change effectivity to reliability to make the question more concise and clearer	4
Question 6. "From the list below please indicate if you consider each of the measures are appropriate (yes or no) and how practical and effective you would consider them to be for assessing mouse welfare in laboratory animal facilities. To assess their practicality and effectivity please use a scale describe below: "(59 indicators organised in animal-based and resource-based; practicability and effectivity assess using a chart with a scale: 1 not practical/effective, 2 fairy practical/effective, 3 neither practical/effective or unpractical/ineffective, 4 slightly practical/effective, 5 very practical/effective).	It can be more complete if have information about the top 5 from all the list of measurements. The introduction to the question is very heavy and it better to have it in a separate section to make it easy to read. It can be useful to add information form the question 4 (e.g. select the 10 most important) then add information from question 1 (e.g. reliability and practicability scores) Definition of the terms would be useful, i.e. for an online survey using a clickable definition. Add to this question an open-end question for additional comments about the measurements.	1

Table 2.4. Rank order was given by the participants of the scoping meeting of the 6 potential questions to be used in the first stage of Delphi consultation.

Based on the scoping meeting, the indicators that participants considered the most appropriate for assessing laboratory mouse welfare were included in first round of the Delphi as well as some additional indicators following an extensive literature survey (Table 2.5).

INDICATOR	**SOURCE**
Body condition score	(Paster *et al.*, 2009)
Coat condition	(Lloyd and Wolfensohn, 1999)
Dehydration	(Lloyd and Wolfensohn, 1999)
Skin colour	(Irwin, 1968)
Respiratory rate	(Rogers *et al.*, 1999)
Body temperature	(Lloyd and Wolfensohn, 1999)
Weight change	(Leach *et al.*, 2006)
Gait	(Irwin, 1968)
Tail position	(Leach *et al.*, 2006)
Audible vocalisations	Scoping meeting
Usage of cage space	(Leach *et al.*, 2006)
Usage of nesting material	(Gaskill *et al.*, 2013b)
Pups outside the nest	(van der Meer *et al.*, 2001b)
Hunched position	(Lloyd and Wolfensohn, 1999)
Alertness	Scoping meeting
Bite wounds/marks	Scoping meeting
Wounds (excluding bite wounds)	Scoping meeting
Excessive urine & faeces in the cage	Scoping meeting
Ocular/nasal discharge	Scoping meeting
Swollen abdomen	(Leach *et al.*, 2006)
Facial expressions of pain	(Leach *et al.*, 2006)
Barbering (hair removal)	(Leach *et al.*, 2006)
Bloodstains in cage	(Leach *et al.*, 2006)
Food consumption	(Lloyd and Wolfensohn, 1999)
Water consumption	(Lloyd and Wolfensohn, 1999)
Exhibition of normal behaviour	Scoping meeting
Exhibition of abnormal behaviour	Scoping meeting
Approaching hand in cage	Scoping meeting
Ease of handling	(Leach *et al.*, 2006)
Urination/defecation during handling	(Irwin, 1968)
Rearing	(Leach *et al.*, 2006)
Cage odour	(Hubrecht *et al.*, 1993)
Positive interactions with conspecifics	Scoping meeting
Negative interactions with conspecifics	Scoping meeting
Faecal cortisol levels	Scoping meeting
Housing density	Scoping meeting
Cage dimensions	Scoping meeting
Complexity of the cage	(Leach *et al.*, 2006)
Floor-type	(Leach *et al.*, 2006)
Substrate type	Scoping meeting
Temperature	(Lloyd and Wolfensohn, 1999)

Ventilation	Scoping meeting
Humidity	(Leach *et al.*, 2006)
Intensity of lighting	(Leach *et al.*, 2006)
Light source	Scoping meeting
Ultrasonic noise levels	(Leach *et al.*, 2006)
Nesting material	Scoping meeting
Food type	Scoping meeting
Frequency of feeding	Scoping meeting
Type of watering system	(Leach *et al.*, 2006)
Type of cage cleaning procedure	(Leach *et al.*, 2006)
Frequency of cleaning cages	(Leach *et al.*, 2006)
Frequency of physical examination	Scoping meeting
Empathetic attitude of staff towards animals	(Leach *et al.*, 2006)
Staff training	(Leach *et al.*, 2006)
Frequency of veterinary procedures	Scoping meeting
Infanticide Rate	Scoping meeting
Mortality Rate	Scoping meeting
Diseases Rate	Scoping meeting

Table 2.5. List of laboratory mice welfare indicators used in Round one of Delphi consultation and their source.

2.3.3. Pilot survey results

Nine participants completed the pilot study. Participants were asked to complete the survey and assess the type of questions, their clarity and to indicate the amount of time needed to complete the questionnaire. This survey provided feedback on the type of questions, the length and the information contained within the questionnaire, and was used to refine the questions and survey format for the first round of the Delphi consultation (Table 2.6).

Question number	Comments
1	• Recommend to have the headings come down the page with you / repeat every 10 questions. • There are parts of the environmental factor list where 'reliability' seems to be meaningless (What do you mean by how reliable detecting a type of cage cleaning procedure is?) • When you miss out a question, point to where exactly that is, not the whole section. This is important in the big box-ticking sections. • I found the second section hard to complete, and apply the concepts of validation, practicality and reliability • Measures such as temperature, RH, ventilation etc. may have welfare implications, but it is the "cage" level that needs monitoring, and this information is not available from current systems.
2	• It would have been useful to have these three terms (validation, reliability and practicability) redefined on this section page.
3	• 400 mice seem like a lot. I may even have gone lower, depending on what the literature suggested the figures should be. • My decision to assess Body condition and Dehydration would depend on observations of coat condition, behaviour etc.
4	• The validity of the indicators has not changed, but a positive or negative could be valid (but not the converse) if that makes sense.

Table 2.6. Summary of the comments obtained from the pilot survey

2.4. First-round questionnaire

2.4.1. Methods

Participants were recruited using a diverse set of methods, including personal contacts, professional organisations relating to laboratory animal welfare, veterinarians working in laboratory animal facilities, literature searches of academics and researchers who have published on mouse welfare in the last 15 years. A total of 206 people agreed to take part in this first round. The questionnaire was then sent via an individual link to each participant using Qualtrics platform (http://www.qualtrics.com). Consenting participants were informed about the aim, methods and duration of the study. Data collected was only used for this specific research project. Ethical approval was granted from Newcastle University (Project ID 449).

The questionnaire was separated into 3 sections. In the first two sections, the participants were given two different theoretical scenarios (Table 2.1) as a guide to complete the questions (Table 2.7) relating to validity, practicability and reliability of potential welfare indicators (Table 2.5).

Theoretical Scenario	Question	Description	Additional information
Section 1: Audit welfare assessment	Question One	Selection of mouse welfare indicators (animal-based and resource-based)	Assessment of the validity, practicability and reliability of each indicator.
	Question Two	Selection of the ten most important welfare indicators	Justification of the selection in the context of validity, practicability and reliability
	Question Three	Stating the optimal means of measurement for each indicator	Selection of whether measurements will be made by observation, questionnaire or both and the proportion of animals/facilities needed for the measurement to be appropriate
Section 2: Everyday welfare assessment	Question Four	Selection of ten most important indicators used in an every-day welfare assessment	Assessment of each indicator in the context of validity, practicability and reliability.
Section 3:	Demographic information	Information related to the participants	

Table 2.7. Summary of the questionnaire used in the Delphi consultation Round one.

Validity was defined as '*an indicator that provides useful information about the animal's welfare*'. Practicability was defined as '*an indicator that can be measured in a reasonable amount of time, incurring a reasonable cost and is feasible within the constraints of a laboratory animal facility*'. Reliability was defined as '*an indicator that produces consistent information when used by different people assessing the same animal and the same person assessing the same animal in the same state on more than one occasion*'. Section three contained demographic information. This questionnaire was 'live' for two weeks (February 2016).

2.4.2. Data analysis

Descriptive statistics were used to analyse the data collected from participants on the validity, practicability and reliability data for the indicators. Different studies have used frequencies, mean, median for analysing data and provide final results (Schmidt, 1997; Hasson *et al.*, 2000; Whaytt *et al.*, 2003; Keeney *et al.*, 2006; Hsu and Sandford, 2007; Rikkonen, 2008; Frewer *et al.*, 2011).This method of analysis was in line with other similar published using the Delphi consultation process e.g. (Bracke, 2006; Leach and Main, 2008; Collins *et al.*, 2009; Geist, 2010; More *et al.*, 2010).

2.4.3. First-round questionnaire results

Demographics

The response rate for round one was 63% with 135 participants of 214.

Section one: audit welfare assessment

One hundred and thirty-five participants assessed the validity, practicability and reliability of 59 indicators in the first section of the round one questionnaire. Indicators that were selected by over 90% of the participant were considered to have high levels of validity, reliability and practicability. The information was divided into animal-based (Figure 2.2) and resource-based indicators (Figure 2.3) with 57 out of 59 indicators considered to be valid by over 50% of the participants. Animal-based indicators considered valid by a high number of participants included; *body condition score* (97.8%), *coat condition* (99.3%), *hunched position* (99.3%), *and exhibition of abnormal behaviour* (96.9%). Resource-based indicators considered valid by a high number of participants included; *staff training* (98.5%), and *mortality rate* (97%).

In terms of the practicability of the indicators; *temperature, nesting material* and *coat condition* were considered as the most practical by participants (>96%). In terms of indicator reliability; *hunched position, visible wounds (excluding bite wounds)* and *bite wounds/marks* were selected as the most reliable (>93%).

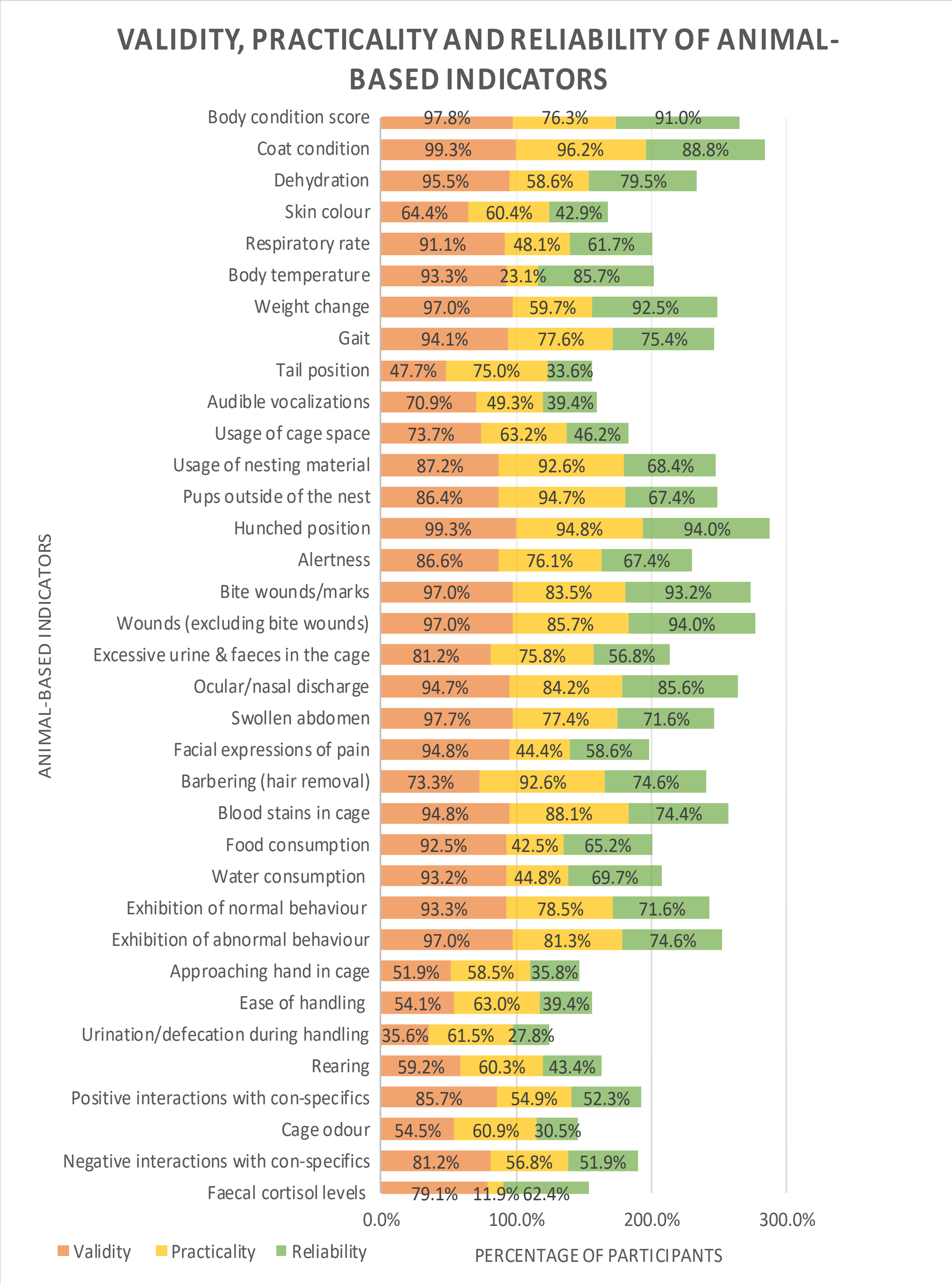


Figure 2.2. Percentage of participants that selected each animal-based indicator as valid (Orange), practical (yellow) and reliable (green).

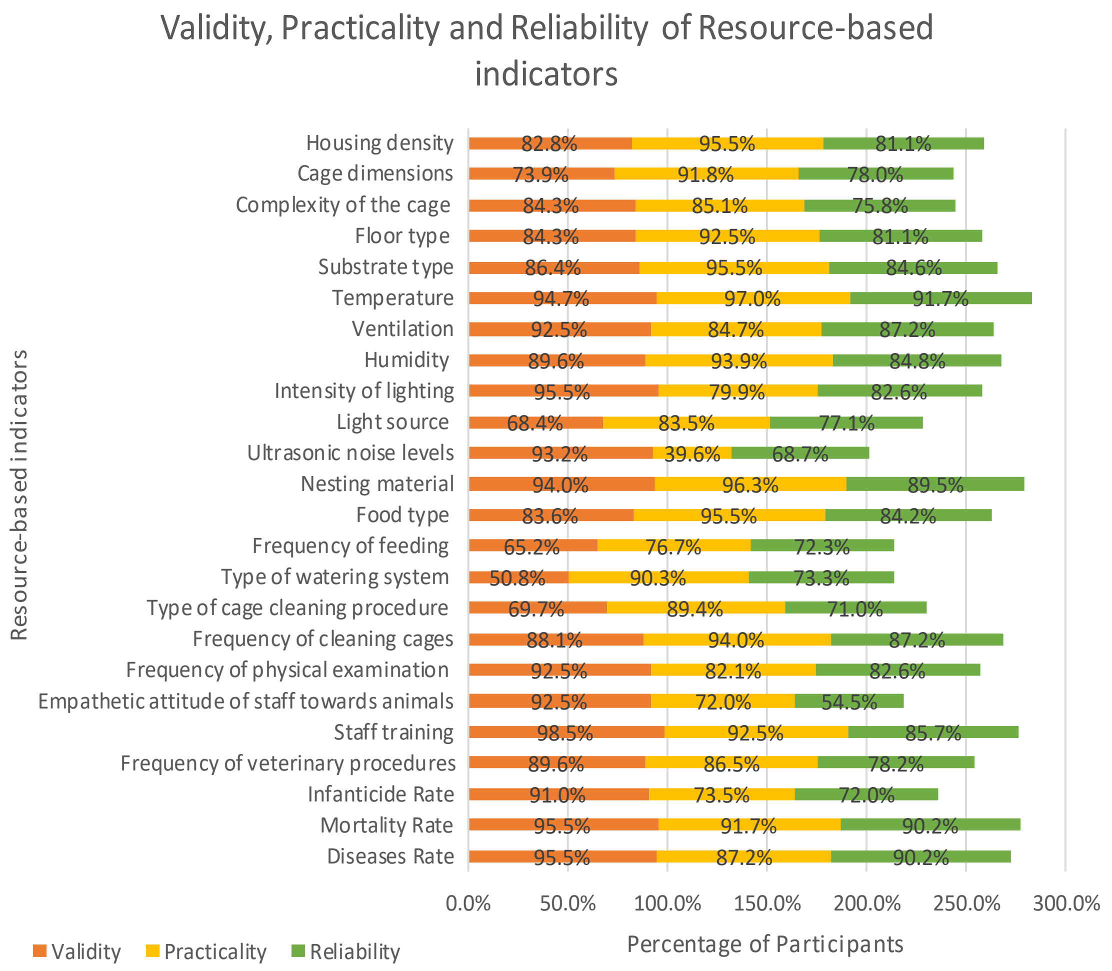


Figure 2.3. Percentage of participants that selected each resource-based indicator as valid (Orange), practical (Yellow) and reliable (Green).

The top ten mouse welfare indicators

The participants were asked to select their top 10 indicators for assessing mouse welfare in an audit welfare assessment. A wide range of indicators was selected by the participants as shown in Figure 2.4. *Hunched position* was the most selected indicator, with 66% of participants including it in their top 10 list.

Type of watering system, light source, and *urination/defecation during handling* were selected by only one participant, and the *frequency of feeding* and *rearing* were not selected by any of the participants. Therefore they were removed from the list going forward into round two of the consultation process. Two additional indicators were added to the list; *light cycle* (8%) and *noise levels* (7%) based on their frequency in the written comments provided by the participants in round one.

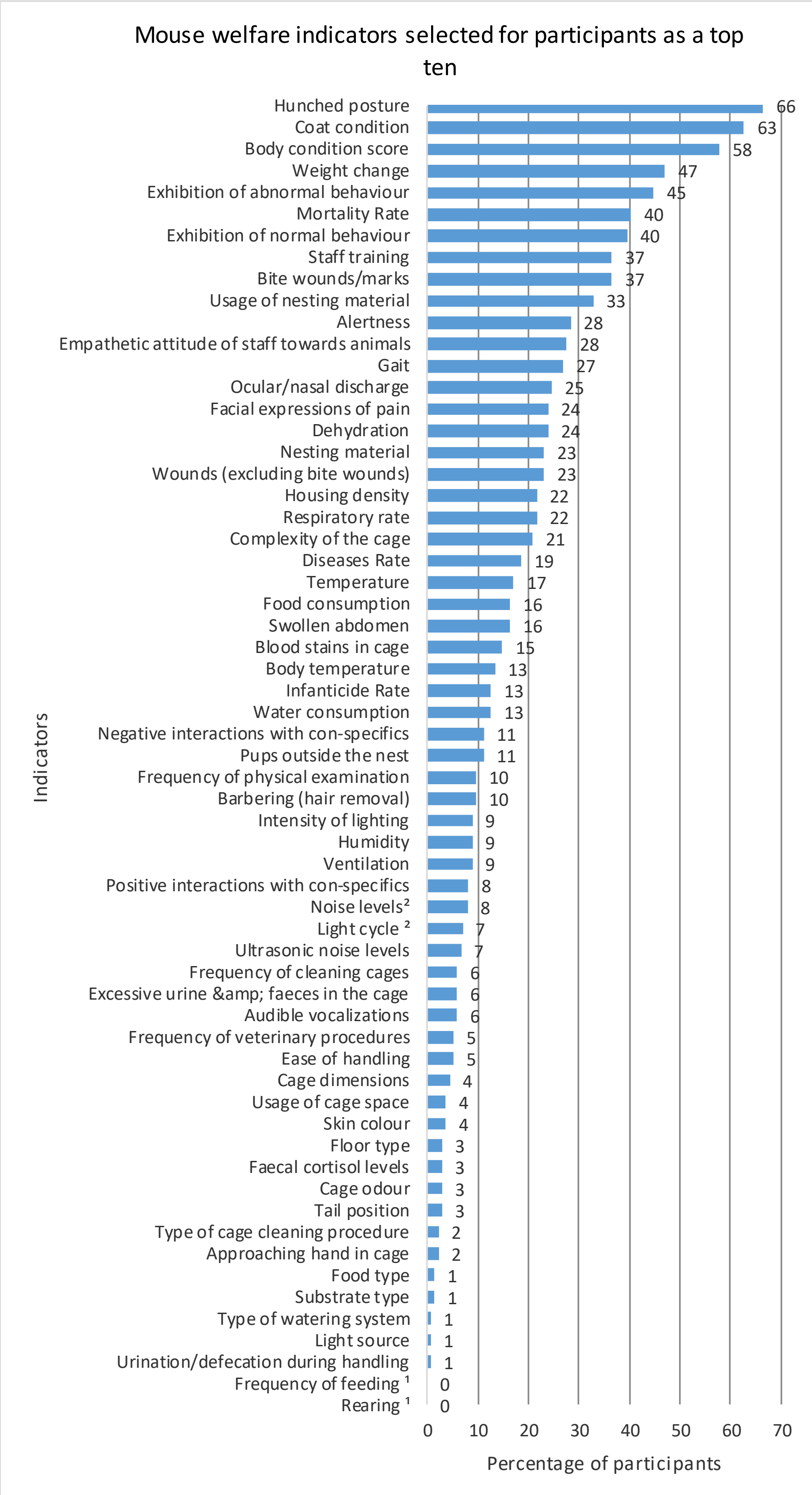


Figure 2.4. The percentage of participants who selected each welfare indicator to be included in their top 10 list, in round 1 of the consultation.

Method of measurement of each indicator

Participants were asked about the means for assessing each indicator either by observation of the animal or animal room, a questionnaire to the facility staff or both (Figure 2.5). Observation of the entire mouse population or facility was the most frequently chosen assessment method (for 28 indicators). The highest percentage of agreement among the participants was seen for *swollen abdomen* (97.9%), *alertness* (97.9%), *coat condition* (96.9%) being assessed by observation; *house density* (96.7%) being assessed by observation and questionnaire (100%) and *light source* being assessed by questionnaire (100%).

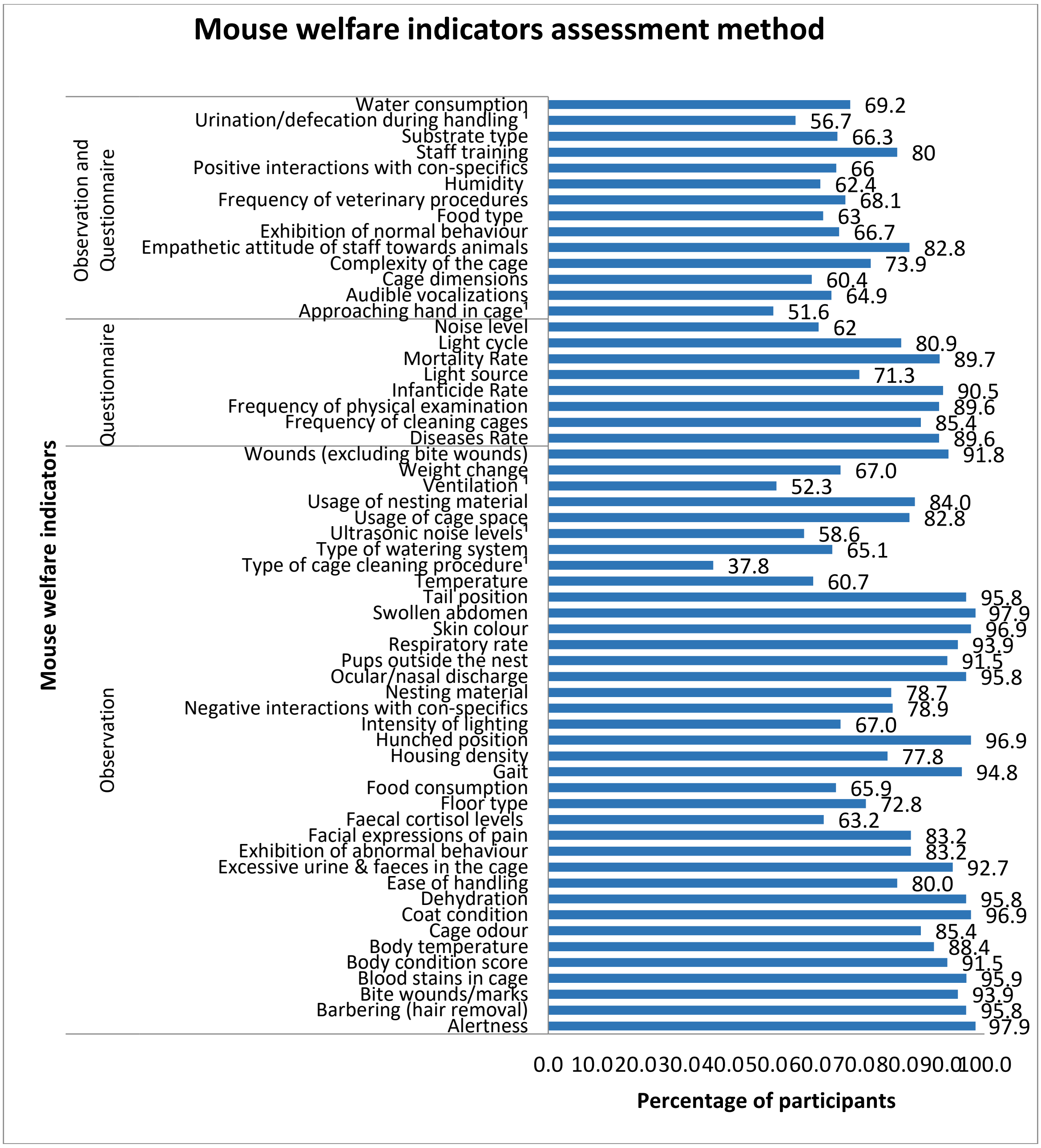


Figure 2.5. Percentage of participant's agreement for the selection of the most appropriate method of assessment.

Section two: Every-day welfare assessment

From the mouse welfare indicators that participants chose as valid in question one, they were asked to select the ten most important indicators for assessing mouse welfare in an everyday assessment (Figure 2.6). The indicators selected most frequently (by >70% of participants) were *coat condition, hunched position,* and *bite wound/marks*. In contrast, *urination/defecation during handling, faecal cortisol levels* and *type of cleaning procedure* were selected by only 3% of the participants.

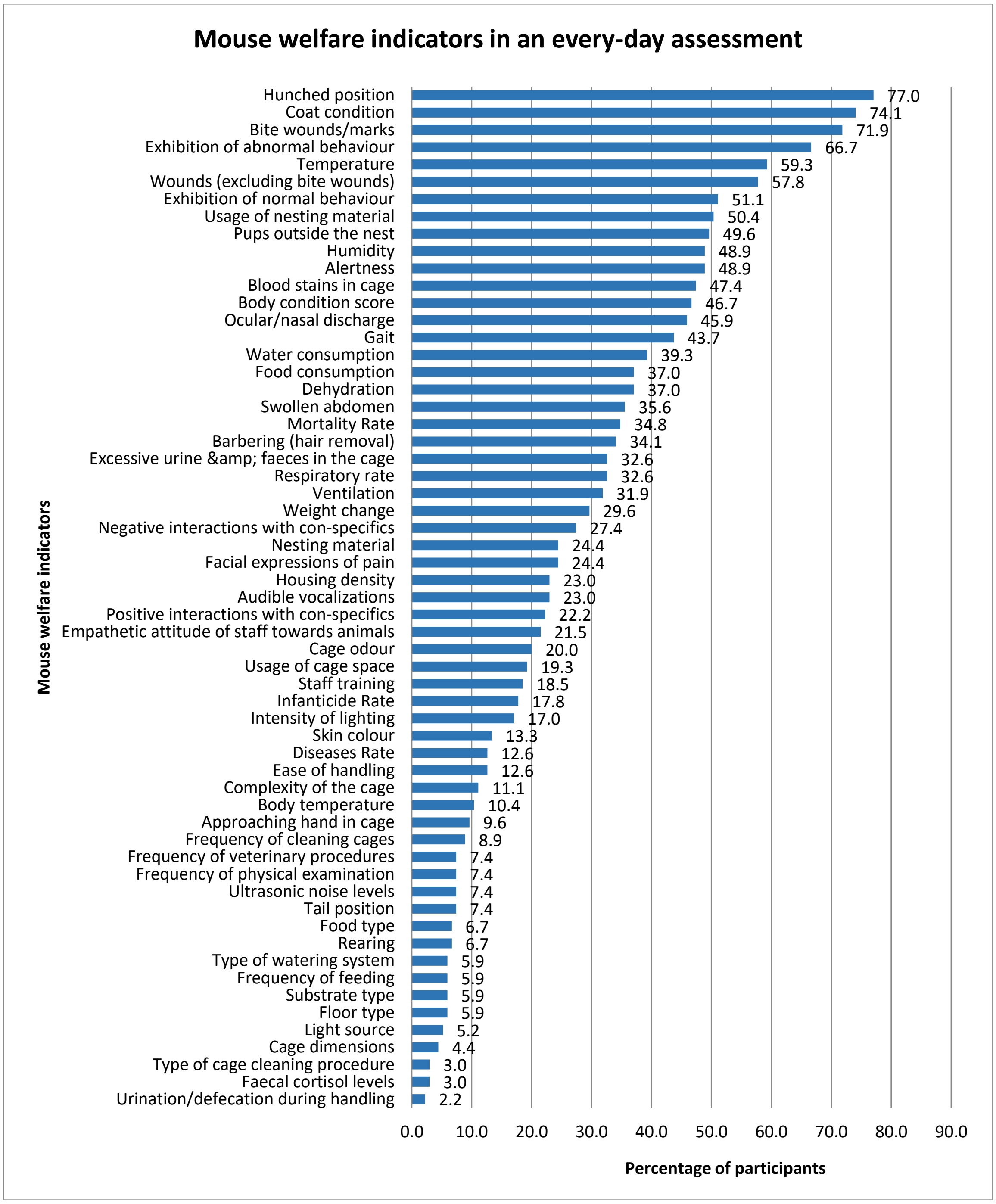


Figure 2.6. Indicators selected as the top ten most important for an every-day welfare assessment in the first round of Delphi consultation.

2.5. Second round questionnaire

2.5.1. Method

Participant's responses in the round one were used to generate the second-round questionnaire. The indicators that were selected as 'valid' or 'very valid' from the list of 59 indicators for both audit and everyday welfare assessment were then organised into a rank according to how frequently they were selected by participants from 1 to 59. The second-round questionnaire was sent out to the participants who completed the round one questionnaire using the Qualtrics platform with a personal link via email. This questionnaire was again 'live' for two weeks (March 2016). The second-round questionnaire was separated into two sections (Table 2.8) in which the participants were given the same two theoretical scenarios from round one (Table 2.1). In question one, participants were presented with the indicators for the audit welfare scenario in the rank order obtained from round one. The participants were asked to state if they agreed with this rank order. If they disagreed, they were then asked to reorder the indicators into the rank position they considered more appropriate and state the reason for the change (i.e. based on validity, practicability and/or reliability). The responses of the top ten indicators for audit scenario selected in the round one questionnaire were presented in rank order, and participants were asked if they were agreed with this final top ten list. If they disagreed, they were then asked to reorder the indicators into the rank position they considered more appropriate and state the reason for the change (i.e. based on validity, practicability and/or reliability). The above process was repeated for scenario 2 (everyday welfare assessment).

Theoretical Scenario	Question	Description	Additional information
Section 1: Audit welfare assessment	Question One	Indicators rank order taking into account participants responses from round one.	Modifying the rank order if is necessary for the context of validity, practicability and reliability of the indicator
	Question Two	Means and the percentage of animals/facilities to assess each indicator.	Indicators can be assessed through observation, questionnaire or both. Percentage of animals/facilities from 10% to 100%
Section 2: Everyday welfare assessment	Question Three	Indicators for everyday assessment in rank order taking into account participants responses from round one.	Modifying the rank order if is necessary for the context of validity, practicability and reliability

Table 2.8. Summary of the questionnaire used in the Delphi consultation Round two.

2.5.2. Data analysis

Descriptive statistics were used to analyse the data collected from participants on the validity, practicability and reliability data for the indicators in the same way as in round 1.

As the Delphi consultation process relies on a group of experts reaching consensus, the validity and rank order of the indicators given were compared between participant’s job role and years of experience working with laboratory animals using a Chi-square test. These two factors (job role and years of experience) were chosen from the demographic information of the participants because they were directly representative of the level of experience of working with laboratory mice. Other demographic data was also collected (e.g. age, gender, and academic level) which was used for reviewing the demographic information about the participants, but these data were not used to analyse the Delphi consultation results as they were not determinant for established the participant's experts level.

2.5.3. Second round questionnaire results

Demographics

The round two was sent to the 135 participants and the response rate was 73% with 98 participants completing this final round. Each participant had an individual code and their individual responses were followed through round one and round two for the data analysis.

Of the 98 participants who completed both rounds, 30% were veterinarians, 20% researchers, 19% laboratory facility managers, 11% technicians, 10% Name Animal Care Welfare Officers, and 8% were Animal Welfare researchers. The majority of the respondents were working in the United Kingdom (41%), followed by USA (13%), Australia (12%), Canada (10%) and Switzerland (8%).

Participant expertise was accessed considering the years of experience working with laboratory animals, qualifications and job position. Most of the participants have a PhD (35%), other qualifications were related to animal welfare (e.g. IAT, Diploma in animal science) (20%) and Master in Science degree (15%). This qualification did have an impact on the years of experience working with laboratory animals (x^2= 32; P> 0.008), being PhD and professional qualifications likely to have over 10 years of experience. Current job position of the participants also affects in the years of working with laboratory animals (x^2= 37.7; P> 0.009) with veterinarians and manager/directors likely to have over 10 years of experience. In total, from the 98 participants, more than 50% had over 10 years of experience with laboratory mice (58.2%), 17.3% with experience between 6 to 10 years and 19.4% with experience between 1 to 5 years. Only 2% of the people who participated in the consultation had less than one year of experience with laboratory mice. Additionally, the years of experience working with laboratory animals was not influenced by sex (72% female and 26% male) (x^2= 3.26; P> 0.51). Participants were widely distributed into several different job categories related to laboratory animals, with 32.7% being veterinary surgeons, 23.5% being researchers working with laboratory animals (of which 4.1% are professors), 19.4% being animal technicians (of which 7.1% are animal welfare officers) and 13.3% being laboratory facility managers.

The validity of some indicators chosen was associated to some extent with the participant's job position. Indicators such as *skin colour* (r^2= 28.3; p=0.019), *weight change* (r^2= 25.6; p=0.042), *usage of nesting material* (r^2= 25; p=0.05), *hunch position* (r^2= 21; p=0.02), *alertness* (r^2= 29; p=0.01), *wounds (excluding bite wounds* (r^2= 27; p=0.003), and *facial expressions of pain* (r^2= 28; p=0.02) were selected more by veterinarians. The participant's experience of working with laboratory mice was associated with the selection of some of the final top 10 indicators. There is a positive association between the selection of *body condition score* as top ten indicators and the length of time that participants have been working with laboratory mice (x^2=14.4; p=0.02). This indicator was chosen more by people with over 6 years of experience (48%). Similarly, *hunched posture* (x^2=12.2; p=0.01) and *mortality rate* (x^2=10.4; p=0.03) was also positively associated with the length of time that participants have worked with laboratory mice. These indicators were chosen more frequently (50%) for participants with over 6 years of experience.

Audit welfare assessment

The overall consensus for the rank order of the 59 indicators used in an audit welfare assessment was 77.2% (Figure 2.7). Based on these results, a consensus among the participants was reached in the second round as participants agreed with the rank order of the 59 indicators for the audit welfare assessment. The indicators with the highest level of consensus for the rank position were *hunched posture,* and *coat condition* ranked first and second with over 90% of agreement between participants.

The indicators with the lowest consensus were *staff training, alertness, empathetic attitude of staff towards animals*, and *facial expressions of pain* with 62% agreement.

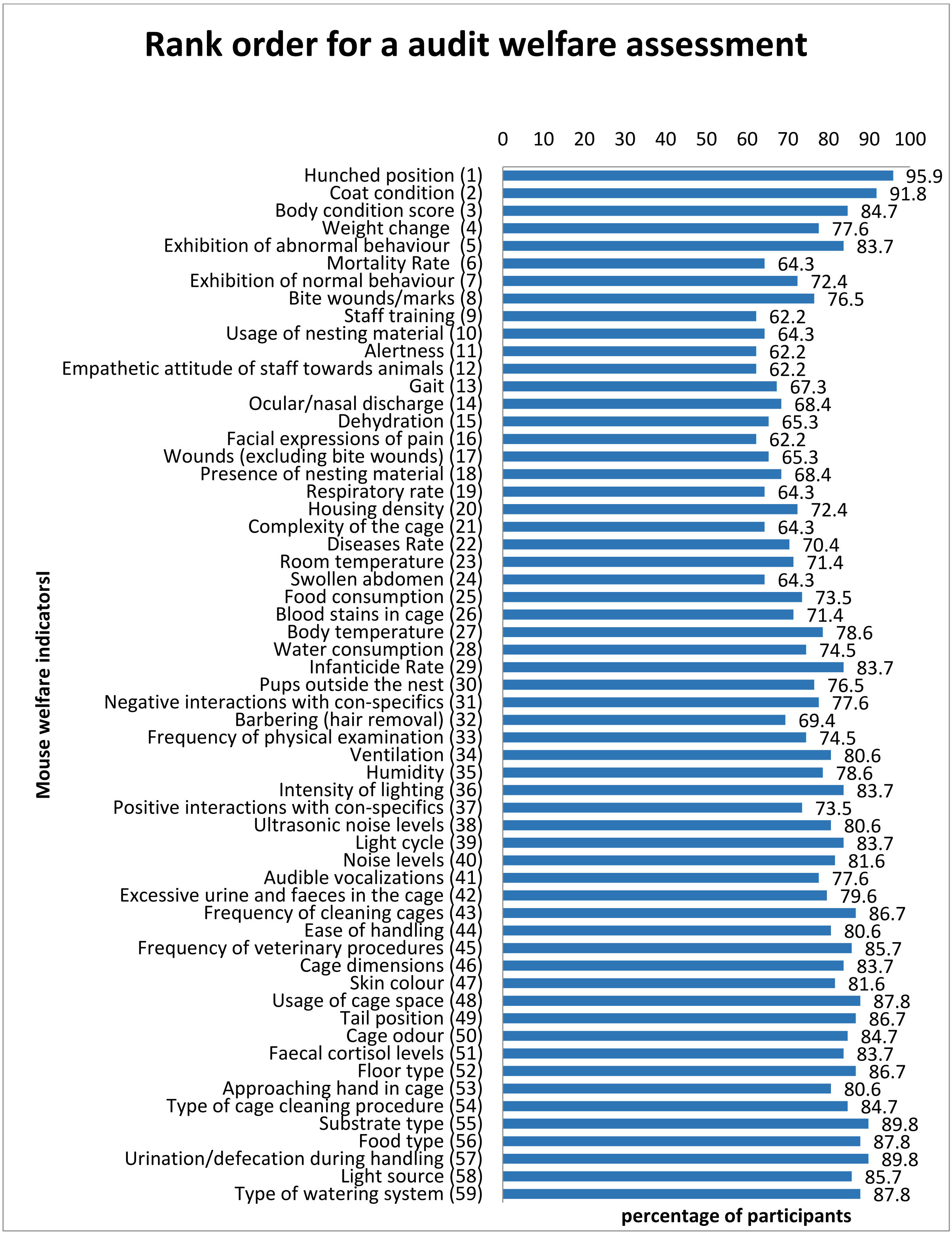


Figure 2.7. The mean rank order for mouse welfare indicators for an audit welfare assessment selected in the Delphi consultation. The percentage of participants who chose the assigned rank order +/- 2 positions are indicated to the right of the figure.

Twenty-nine per cent of the participants (n=28) agreed with the round one rank order, and seventy-one per cent of participants made only very minor changes to the indicators rank order from round one, and the order of the indicators did not change from one extreme position to another. Most of the indicators were stable in the same position (+/- 2 positions). Table 2.9 shows the percentage of participants who moved the indicators from the original rank. All the indicators from the list (27) changed only one to two positions up or down from the original position in the rank in round 1.

Indicator and rank position	Suggested new position	Percentage of participants
Respiratory rate (19)	20	13.3
Housing density (20)	21	12.2
Diseases Rate (22)	23	12.2
Room temperature (23)	24	11.2
Swollen abdomen (24)	25	11.2
Food consumption (25)	26	15.3
Blood stains in cage (26)	27	14.3
Body temperature (27)	28	18.4
Water consumption (28)	29	20.4
Infanticide Rate (29)	30	23.5
Pups outside the nest (30)	31	20.4
Negative interactions with con-specifics (31)	32	19.4
Barbering (hair removal) (32)	33	17.3
Frequency of physical examination (33)	34	10.2
Ventilation (34)	35	12.2
Ultrasonic noise levels (38)	39	13.3
Light cycle (39)	40	15.3
Noise levels (40)	41	14.3
Audible vocalizations (41)	42	12.2
Frequency of cleaning cages (43)	44	12.2
Frequency of veterinary procedures (45)	46	10.2
Cage dimensions (46)	47	10.2
Skin colour (47)	48	10.2
Usage of cage space (48)	49	10.2
Substrate type (55)	54	10.2
Food type (56)	55	12.2
Urination/defecation during handling (57)	56	10.2

Table 2.9. Indicators of the audit scenario showing the changes in the rank position for over 10% of the participants. Number in brackets represent the actual rank position suggested in the Delphi round two.

Everyday welfare assessment

Participants were asked to agree or disagree with the rank order, from the round one consultation, of the most important indicators to use in an everyday welfare assessment by technical staff (Figure 2.8). A consensus was reached with 85.7% of agreement between the participants. Overall the 85.7% of the participants were agreed on the rank order for the 59 indicators in the everyday welfare assessment scenario. There were four indicators with consensus level over 95% (*hunched posture, coat condition, food type, substrate type* and *light source)*. The lowest agreement were *Humidity, room temperature* and *gait* were the indicators with the lowest consensus level with 66%, 67% and 68% respectively.

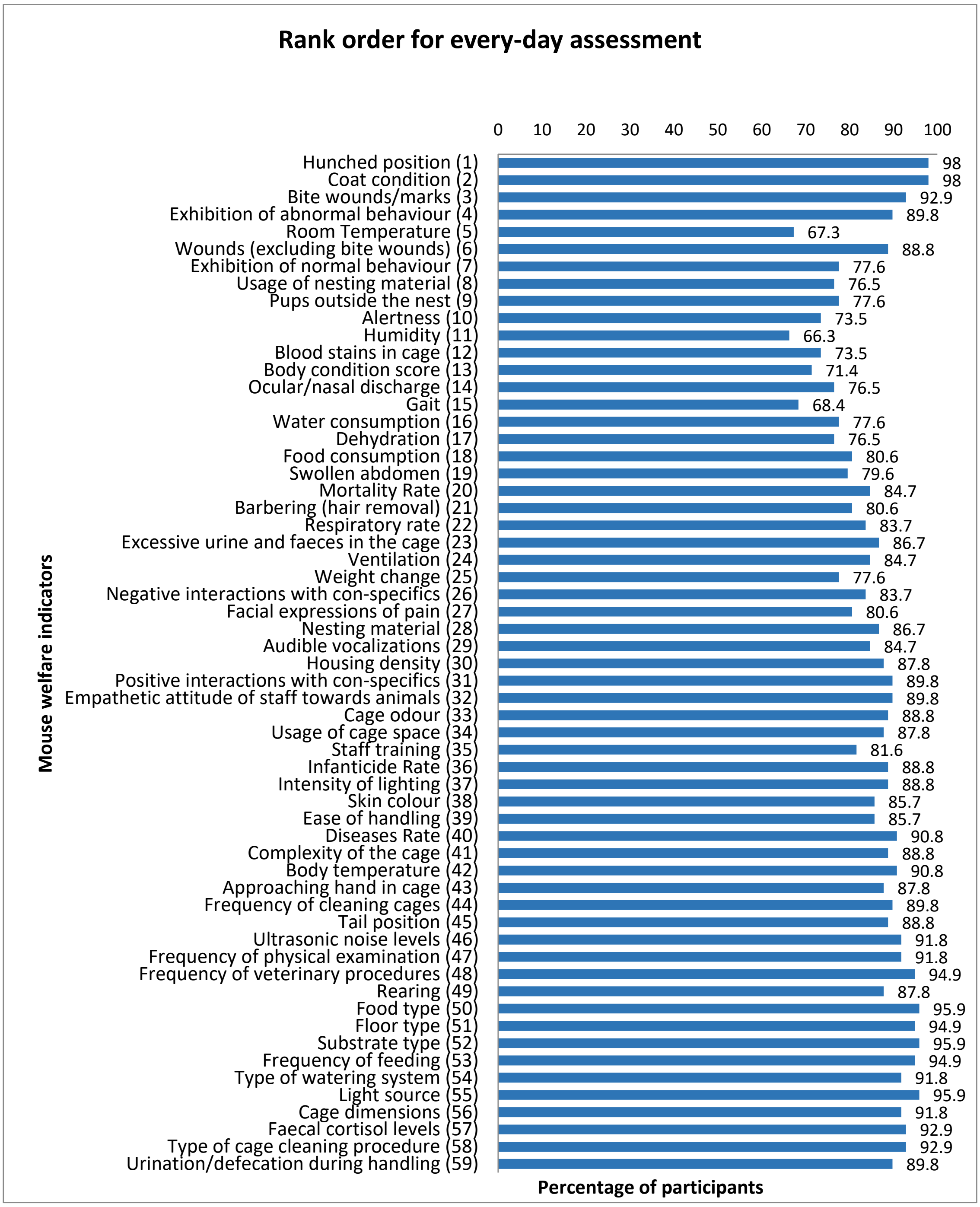


Figure 2.8. The mean rank order for the mouse welfare indicators for an everyday welfare assessment selected in the Delphi consultation. The percentage of participants who chose the assigned rank order +/- 2 positions are indicated to the right of the figure.

Thirty-eight per cent of the participants (n=37) agreed with the round one rank order, and sixty per cent of participants made only very minor changes to the indicators rank order from round one, and the order of the indicators did not change from one extreme position to another. Most of the indicators were stable in the same position (+/- 2 positions). Table 2.10 shows the percentage of participants who moved the indicators from the original rank. All the indicators from the list (26) changed by only one position down from the original position in the rank in round 2.

Indicator rank position	suggested new position	percentage of participants
Coat condition (2)	3	13.3
Bite wounds/marks (3)	4	12
Exhibition of abnormal behaviour (4)	5	17
Room Temperature (5)	6	10.2
Wounds (excluding bite wounds) (6)	7	13
Pups outside the nest (9)	10	11.2
Bloodstains in cage (12)	13	12.2
Body condition score (13)	14	10.2
Ocular/nasal discharge (14)	15	11.2
Water consumption (16)	17	10.2
Dehydration (17)	18	14.3
Food consumption (18)	19	14.3
Swollen abdomen (19)	20	13.3
Mortality Rate (20)	21	15.3
Barbering (hair removal) (21)	22	12.2
Respiratory rate (22)	23	13.3
Excessive urine and faeces in the cage (23)	24	10.2
Ventilation (24)	25	14.3
Weight change (25)	26	12.2
Negative interactions with con-specifics (26)	27	14.3
Facial expressions of pain (27)	28	12
Nesting material (28)	29	14.3
Audible vocalizations (29)	30	12.2
Housing density (30)	31	11.2
Cage odour (33)	34	12.2
Usage of cage space (34)	35	10.2

Table 2.10. Table of the indicators from the everyday scenario showing the changes in the rank position for over 10% of the participants. Number in brackets represent the actual rank position suggested in the Delphi round two.

Top ten indicators to be used in the audit assessment and the everyday welfare assessment.

From the initial list of 59 indicators (See table 2.5), the top ten selected by the participants as the most valid, reliable and practical to be used in an audit welfare and for the everyday welfare assessment are shown in Table 2.11.

Nature of the assessment	Laboratory mouse welfare indicators
Audit	Hunched posture, Coat condition, Body condition Score, Weight change, Exhibition of abnormal behaviour, Mortality rate, Exhibition of normal behaviour, Bite/wound marks, Staff training, Usage of nesting material.
Everyday assessment	Hunched posture, Coat condition, Bite/wound marks, Exhibition of abnormal behaviour, Room temperature, Wounds (excluding bite wounds), Exhibition of normal behaviour, Usage of nesting material, Pups outside the nest, Alertness.

Table 2.11. Top ten laboratory mouse indicators for the assessment of welfare based on expert's opinion.

The percentage of participants that scored these indicators as valid, practical and reliable for the audit and everyday welfare assessment is shown in Figures 2.9 and 2.10 respectively.

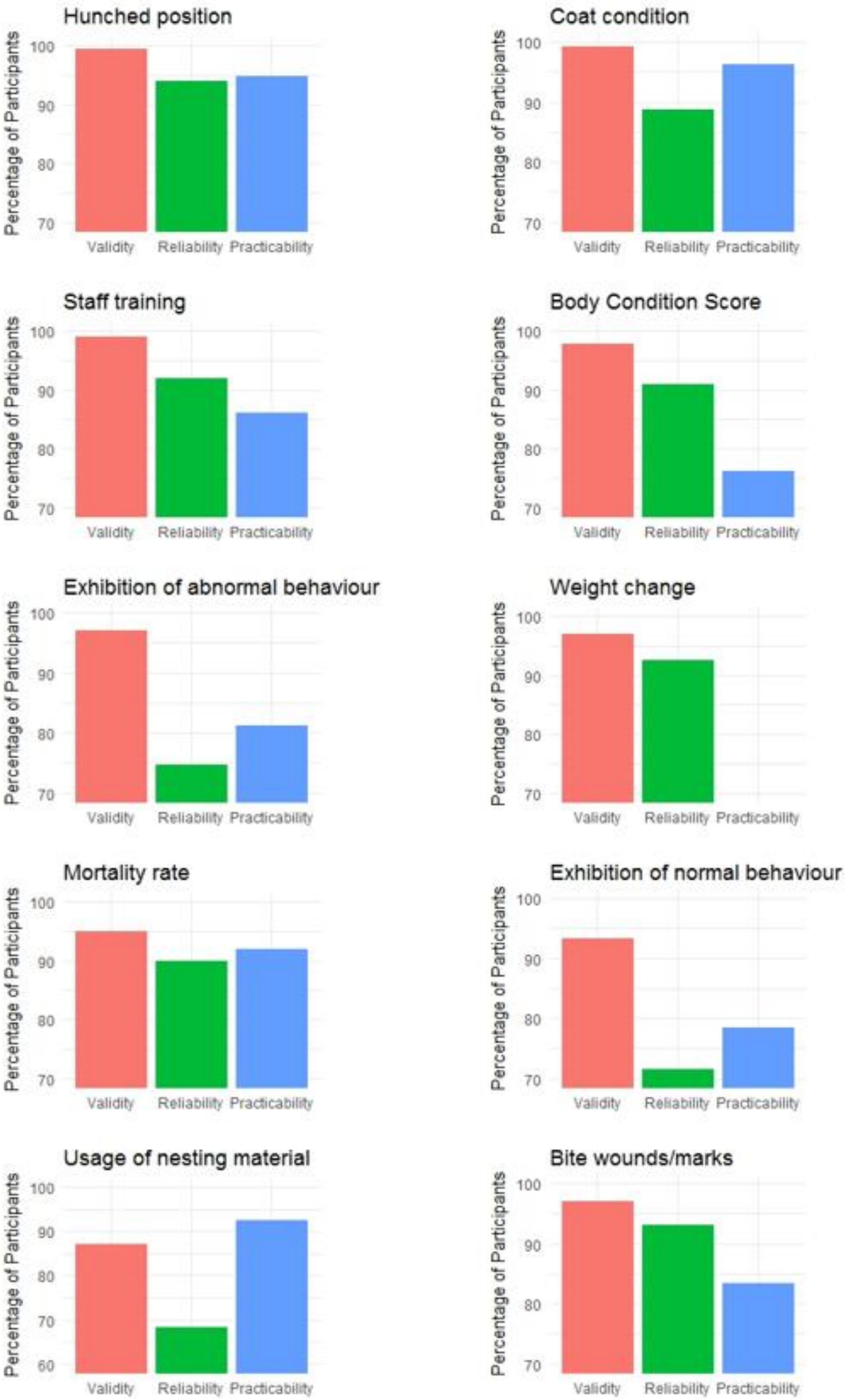


Figure 2.9. Summary of top ten indicators selected by participants to be used in an audit assessment. The x-axis represents the three items that were assessed by participants, validity (red), reliability (green) and practicability (blue). The y-axis represents the percentage of participants who scored each item as valid and very valid in the Delphi consultation.

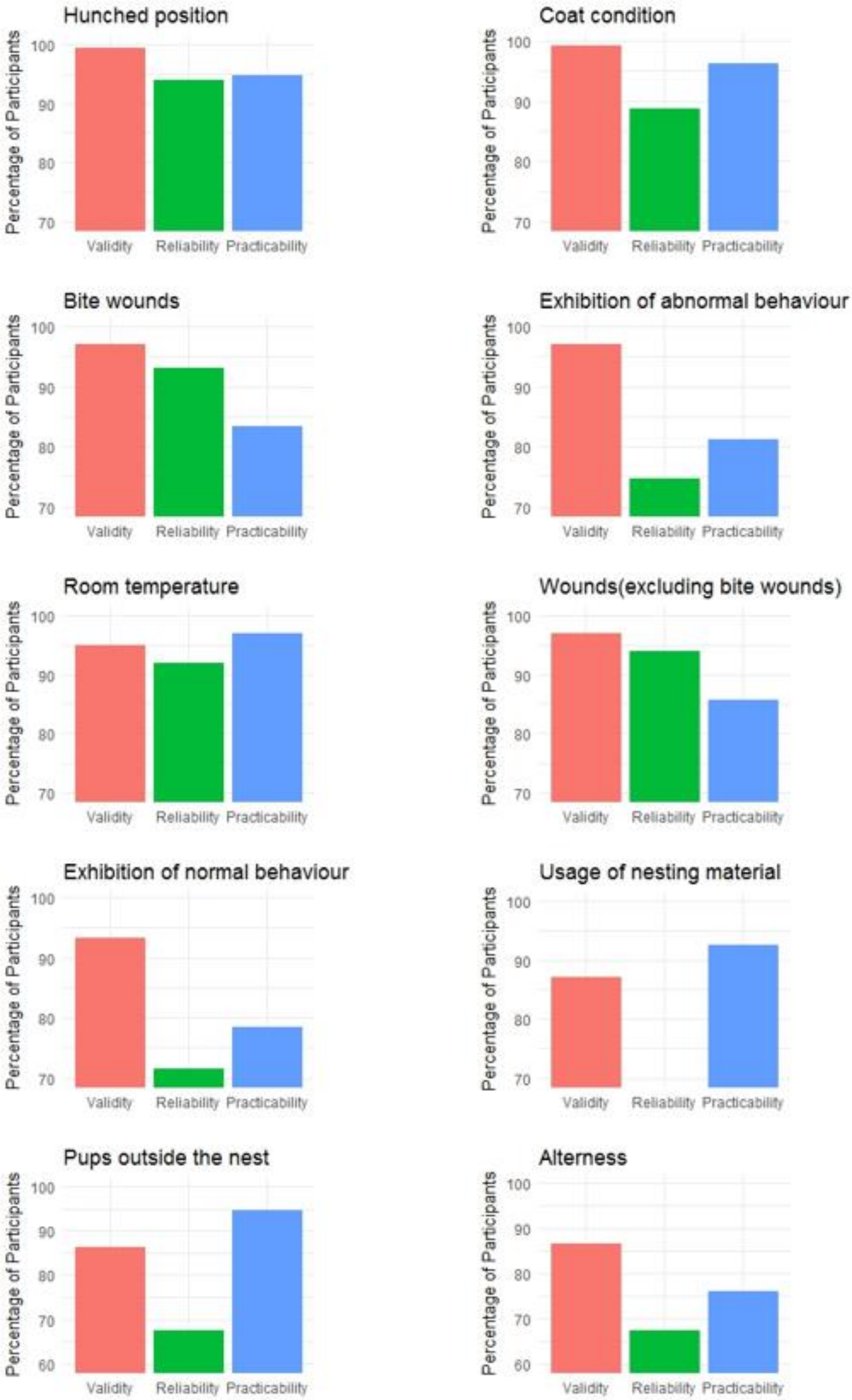


Figure 2.10. Summary of the top ten indicators selected by participants to be used in an everyday welfare assessment. The x-axis represents the three items that were assessed by participants, validity (red), reliability (green) and practicability (blue). The y-axis represents the percentage of participants who scored each item as valid or very valid in Delphi consultation.

2.6. Discussion

This study aimed to determine, through expert opinion, which laboratory mouse welfare indicators would be valid, reliable and practical to use in two scenarios; an audit welfare of a laboratory mouse facility (e.g. carried out by an independent assessor) and an everyday welfare assessment (e.g. carried out by technical staff as part of their routine duties).

Delphi methodology proved to be a valuable tool for aggregating information from laboratory mouse welfare experts across the world, allowing them to exchange opinions and come to a consensus (Whaytt *et al.*, 2003; Bennett *et al.*, 2004; Rikkonen, 2008; More *et al.*, 2010). The Delphi consultation process focussed on the rank order of 59 indicators in each specific context. A consensus was reached with an agreement of 77.2% for the most valid, practical and reliable indicators for the audit welfare assessment and 85.7% for the most valid, reliable and practical indicators for the everyday welfare assessment.

The rank order given to the participants in round one of the Delphi consultation did not have considerable changes in the results of round two. Participants changed some of the position of the indicators but only one to two places up or down from the original position from round one. The changes were predominantly after the rank position 19 in the list. The indicators where a higher proportion of the participants decided to modify the rank position were for water consumption (20.4%), infanticide rate (23.5%) and pups outside the nest (20.4%). A possible explanation of the slight movement in the rank order of these indicators may be related to their low-rank position. They might not be considered very relevant for the assessment of laboratory mouse welfare, so when organising the 59 indicators from the most to the least important, the position towards the end of the rank order could be more challenging for the participants. Another possible explanation could be related to the little evidence about the validity of these indicators as a means of assessing welfare compared with the higher rank indicators, making these indicators more susceptible to variation in expert opinion. For example, water consumption has not been validated in the literature as there is little information about its' usage as a welfare indicator other than in terms of pain assessment only (Jacobsen *et al.*, 2012).

For the audit scenario, the highest-ranked three indicators from the entire list of 59 with the highest agreement (over 84%) did not change their ranked position from round one of the Delphi consultation process, supporting participant's initial opinion about the high validity of these three indicators (i.e. hunched posture, coat condition, body condition score). In the top 10 ranked indicators, most were animal-based indicators (8 out of 10) demonstrating the high credibility (or at least the high level of confidence) that the participants had about this type of indicators. It is interesting to note that the top four indicators relate to biological function (i.e. hunched posture, coat condition, body condition score, weight change), followed by indicators relating to behaviour, social interaction and the environment (i.e. exhibition of abnormal and normal behaviour, bite/wound marks, and use of nesting material) and staff training and mortality rate as a resource-based indicators. These results further support the importance of physical and physiological indicators in welfare assessment (Van de Weerd *et al.*, 1997a). Hunched posture is considered a gold standard measure of pain which is very important for laboratory mice (Baumans *et al.*, 1994). Similar to hunched posture; coat condition, body condition score and weight change are related with animal's physiological welfare that can be affected when an animal is ill, or in pain (Olfert, 1995; Ullman-Culleré and Foltz, 1999; Paster *et al.*, 2009). These physiological measurements which constantly adapt to maintain an individual's welfare can be measured in a non-invasive manner, which might provide a high level of practicality (Barnett and Hemsworth, 1990). Behavioural indicators are also important as they considered to show the individual's adaptations to the present environmental conditions (Dawkins, 1990; Würbel *et al.*, 1996; Augustsson and Meyerson, 2004), and allow us to determine if the environment that we are providing the animals is appropriate for their welfare (Spangenberg and Keeling, 2015). Exhibition of normal and abnormal behaviour is a measurement that can assess animals ability to adjust to the environment (Spangenberg and Keeling, 2015). More recently nesting material has been using as an indirect assessment of welfare-related with pain (Van de Weerd *et al.*, 1997b; Gaskill *et al.*, 2013b) and environment temperature in laboratory mice (Rock *et al.*, 2014).

Conversely, *mortality rate* and *staff training* are the only two resource-based indicators included in this list. Staff training can have a significant impact on laboratory mouse welfare. For example, inadequate training can lead to improper handling which can cause fear affecting animal's performance and welfare (Gonyou *et al.*, 1986; Hawkins *et al.*, 2011). Mortality rate is an indicator which assesses welfare retrospectively, and it can be relevant for assessing the presence of diseases or increasing mortality rate can be evidence of problems in experimental designs

(Home Office, 2014b). Despite the small number of resource-based indicators selected, their inclusion in welfare assessment protocols is important as they include common procedures, treatments and management which can have a high welfare impact (Wells *et al.*, 2006a). This impact can be relevant especially in laboratory animals (e.g. room temperature preferences, environmental enrichment in the cages) as they can be useful for assessing how we look after the mice in the laboratory facilities in terms of environment and housing (Baumans, 2005b) contrary to other authors suggestions about the assessment of welfare using only animal-based indicators (Spangenberg and Keeling, 2015).

The indicators with the lowest percentage of agreement (62%): *staff training, alertness, empathetic attitude of staff towards animals* and *facial expressions of pain* were still highly ranked (9, 11, 12 and 16 out of 59 respectively). These indicators could be considered to have a qualitative component where the assessor gives a score based on their impression. This qualitative component might explain their high rank but low general agreement. As the participants considered these indicators important but have not yet been fully validated as some of them are relatively new (e.g. facial expression of pain – method initially published by Langford et al., in (2010)), which might have influenced their rank position. However, these indicators have only recently begun to be validated, and so this could have only had a minor impact on their rank position, e.g. the use of facial expressions for assessing pain (Leach *et al.*, 2012; Miller *et al.*, 2016; Akintola *et al.*, 2017). Therefore, their validity may be increasingly being accepted by experts.

Participants agreed with the final rank position of the 59 indicators for the everyday welfare scenario. The first two indicators (hunched posture and coat condition) showed a high level of agreement between participants with 98% agreement for the final rank. The agreement for these two indicators relates to the high perceived validity and practicability for the assessment of welfare in laboratory mice, especially for hunched posture in relation to pain and distress (Hawkins, 2002; Paster *et al.*, 2009) and coat condition in relation to anxiety and distress (Holmes *et al.*, 2002; Nollet *et al.*, 2013). These results show that they are considered valid, reliable and practical by the experts. Conversely, the indicators with the lowest percentage agreement were humidity (66.3%), room temperature (67.3%) and gait (68.4%) which were located in the rank position 11, 5 and 15 respectively. Humidity and room temperature are resource-based indicators which are usually assessed through observation of the facilities. However, as the everyday assessment is mostly performed for people who are already familiar with the facilities, these indicators might not be seen as relevant as other indicators as they are included in the codes of practice and recommendations about care and housing conditions (e.g. room temperature, humidity, ventilation) and so have to be monitored daily (AAALAC, 2011; Guillen, 2012; Home Office, 2014a). Gait, on the other hand, is an animal-based indicator related to biological health as changes in mouse gait can be associated to pain, or coordination issues (e.g. related to brain function) (Guyenet *et al.*, 2010). Gait as an indicator is easy (i.e. practical) to assess as it only requires observation of the mouse in the home cage. However, it might not be very reliable, as although it is often shown in response to pain or injury (Arras *et al.*, 2007), the natural behaviour of mice may be to hide this behaviour, as they are a prey species. Therefore, this behaviour may not be present when observing the mice in their home cage as they could feel threatened and so hide this behaviour (Kaliste, 2004). The top ten indicators selected for an everyday welfare assessment can be seen in Table 2.11. The consensus reached for all the indicators in the everyday assessment was higher (87.5%) compared to the audit assessment rank order (77.2%). A possible explanation for the difference in agreement between the scenarios could be due to variances in time, resources and personnel required in the everyday welfare assessment. The everyday assessment needs to be more practical and so should involve the selection of indicators that are easy to assess, less time consuming and that require fewer resources. In addition, the personnel who perform the assessment are usually familiar with the facilities including general environmental indicators which could have affected the selection of the indicators by making it more straightforward. The final top ten indicators selected for this assessment were mainly animal-based (9 out of 10) which are usually assessed through observation of the animals (Leach *et al.*, 2008a) giving them high levels of practicability as these indicators are assessed through observation, which can be made simultaneously (e.g. the assessment of behaviour, gait and body condition score).

Furthermore, these indicators can be more practical when assessed by an experienced staff member (e.g. technicians), which are usually the staff who perform this type of assessment (Hawkins *et al.*, 2011). The only resource-based indicator present in this final top ten list was room temperature. Despite the low percentage of agreement (67.3%) of this indicator, it was still considered a top ten measure in the everyday welfare scenario. One possible explanation of the inclusion of room temperature as the only resource-based indicator maybe is related to its practicability. As temperature can be easily accessed through unit records or direct observation of the room thermometer, therefore its' inclusion can still be practical for an everyday welfare assessment.

An important finding of this study is the differences between the final lists for each scenario. Despite an identical starting list of 59 indicators, the final, top 10 list differs between the scenarios by 4 indicators. The top ten audit welfare assessment indicators include *body condition score, weight change, mortality rate* and *staff training* which are not present in the every-day welfare assessment list. These differences could be explained in part by differences in the scenarios proposed in the questionnaire (Table 2.1) which involve a different amount of time to available, and the assessments are carried out by a type of personnel (i.e. care staff vs. auditor). An everyday welfare assessment, for example, is usually performed by technical staff, who know the facility and the individual animals to be assessed and are monitoring the welfare state of the animals in their care. However, an audit would typically be carried out by an auditor (e.g. home office inspector), who may be less familiar with the facility and animals and also has a different motivation for assessing welfare (i.e. compliance with regulations or welfare standards). In order to comply with the time limit for a unit welfare audit (4 hours), the indicators used needed to be accurate, practical and rapid to score therefore indicators such as *body condition score*, *mortality rate* and *staff training* were included by the experts. *Body condition score*, for example, provides information about mouse health status more practically than assessing body weight, where a scale and comparison of previous weight is needed (Ullman-Culleré and Foltz, 1999). Mortality rate is a resource-based indicator used as a retrospective assessment of welfare and is generally provided by unit records of the number of mice that died either unexpectedly (i.e. found dead) or as part of the normal mortality rate of a specific strain (Clough, 1982). However, this indicator is sometimes not considered as a welfare measure because it is recorded at facility level, retrospectively, thus it is not an indicator of immediate welfare (Botreau *et al.*, 2007a).

Staff training is also an important indicator of the audit welfare scenario where a longitudinal approach to welfare is considered. Although there is limited research about the direct impact of staff training on the welfare of laboratory mice, recommendations (AAALAC, 2011; Home Office, 2014a) about laboratory animal welfare consider the ability of staff to handle and observe mice as critical to reducing negative impacts on welfare as experienced and trained staff can identify problems promptly (Hubrecht *et al.*, 1993).

Alternatively, *room temperature, wounds (excluding bite wounds), pups outside the nest* and *alertness* were included in the every-day welfare assessment top ten list but not in the audit assessment. The usage of *room temperature, wounds* and *alertness* indicators in every-day welfare assessment is likely relevant as the assessment is made daily using records which can be made for room temperature (Home Office, 2014a) or observing the animals in the case of pups outside the nest, wounds and alertness (Lloyd and Wolfensohn, 1999). Due to the fact that the staff who perform this assessment are in contact with the animals every day, they are likely to be effective at noticing subtle changes when observing the animals as they are already familiar with the species, the strain, the individuals, and in many cases the procedures that have been carried out. Therefore they are more experienced in assessing these indicators, and they are of more immediate value.

It is important to underline that even though this study uses a rank order to define the level of face validity, taking into account expert's opinion, rank order is not relevant for the indicators in terms of defining their individual level of importance over other indicators (i.e. meaning that 10 is not less important than 9). The relevance of this study is the final list of indicators, taking into account the type of assessment scenario, and not the assessment of each individual indicator. As it has been stated before, it is an aggregation of different types of indicators (resource-based and animal-based) into a protocol which determines the final welfare assessment and not a single indicator alone (Rousing *et al.*, 2001; Van der Meer *et al.*, 2001a; van der Meer *et al.*, 2001b; Wells *et al.*, 2006a; Botreau *et al.*, 2007b).

Some caution should be taken in interpreting the results from this study. The scenarios used in this study involved very specific descriptions (Table 2.1) of an animal research facility, which could have affected the indicators selected as well as the reason for their selection. Due to the nature of the suggested scenarios and the specific information about the facilities (number of animals, racks, room), a specific list of indicators have been selected which may not be applicable in different circumstances. It is important to highlight that the indicators selected for this study are those that relate to the influence on welfare of housing and husbandry rather than indicators related to the experiments conducted on the animals, which were not included. However, it can form the basis for a protocol for assessing welfare in different facilities and following a specific experimental protocol by including additional indicators linked to those facilities (e.g. number of animals, racks, room) and the procedures carried out (e.g. surgeries, treatments, and behavioural tests), for example body condition score for assessing mouse condition in tumour studies (Russell *et al.*, 1959; Morton, 1998; Morton, 1999; Stokes, 2002). The specific indicators for assessing the welfare of mice enrolled in studies can be obtained from the protocols that are required to be submitted (i.e. as part of licensing), which usually include welfare assessment indicators, score sheets and endpoints before any animal research is carried out.

This study has several practical implications. It could be used as a preliminary source of face validation to select indicators for a mouse welfare assessment protocol taking into account the purpose of the assessment, i.e. a welfare audit or daily welfare check. In addition, the intra- and inter-observer reliability needs to be further assessed in different welfare scenarios and by different observers, using experimental and stock animals to be considered truly effective. It also can be concluded that when assessing stock mice, or those not yet actively involved in any research, the indicators of welfare in Table 2.11 are deemed the most valid to use, based on expert opinion. The expert opinion used in this study has confirmed the importance of biological and environmental indicators for the welfare assessment of laboratory mouse welfare and is aligned with the definition of laboratory mouse welfare proposed in Chapter one. The next step in the construction of laboratory mouse welfare protocol is the means of assessment of laboratory mouse psychological states, which will be the topic of the next two chapters.

Chapter 3. Qualitative Behavioural Assessment: free choice profiling

3.1. Introduction

As a main component of animal welfare, the assessment of psychological state or emotional component of welfare (Duncan and Petherick, 1991) is critical (Mendl and Paul, 2004; Broom, 2010). The means for measuring psychological states in animals is challenging, as the gold standard measurement in humans is a self-linguistic report, to provide information about emotional states as an underlying dimension (Bateson *et al.*, 2011; Bateson and Nettle, 2015). As this method is not possible in animals, a representation of this report as a proxy index based on inferential reasoning has been used (Bateson *et al.*, 2011). This representation is based on the expression of physiological, behavioural, cognitive and biochemical changes which arise when animals are experiencing negative or positive emotional states (Bateson *et al.*, 2011). In laboratory animals, this assessment has been carried out using separate components of the expression of emotional states including behaviours or using different behavioural assays such as cognitive bias, preference, aversion, and motivation tests (Panksepp, 1994; Cardinal *et al.*, 2002; Antonacopoulos and Pychyl, 2010; Brilot *et al.*, 2010). However, for some of these assessment tools, the measurement taken can be affected by other factors besides the underlying emotions. For example, using glucocorticoids (e.g. cortisol, corticosterone) for measuring stress in animals can be fraught with difficulties, as glucocorticoids are affected by other factors, such as circadian rhythm (Carter, 2001). One relevant method that can provide a new insight into the assessment of an animal's emotional state in the laboratory is Qualitative Behavioural Assessment (QBA). QBA takes a more qualitative approach, taking into account the observer's description of what they see and interpretation of the animal's overall demeanour (Wemelsfelder, 2007). This approach integrates the animal's subjective experience and its behaviour by focusing on the animal as a dynamic and communicative whole agent (Wemelsfelder, 1997). This animal expressivity considers categories such as curious, timid, calm, excited as the description of behavioural styles or expressions which provides more information about the animal (Wemelsfelder, 1997).

QBA contrasts with other methods that are used to assess behaviour which focus on an animal's psychological state as a single component (Wolfer *et al.*, 2004). QBA not only assesses the behaviour itself but also the sensory-emotional experience by the use of behavioural expressions used by the assessors (Wemelsfelder *et al.*, 2000; Wemelsfelder, 2007). The assessment of this sensory-emotional experience, present behind the behaviour, is necessary for assessing the psychological state of animals (Wemelsfelder *et al.*, 2000; Wemelsfelder *et al.*, 2001). For example, in the assessment of pain, behaviours such as belly pressing (Miller *et al.*, 2016) are used, but they may only assess the physiological changes in the nervous system that are involved in the perception of pain (i.e. nociception), but not the underlying emotional experience (Wallace *et al.*, 2005). As with many other emotional states, pain is defined as a sensory and emotional experience (IASP, 2008) thus to assess pain the underlying emotional experience also needs to be assessed accurately (Flecknell *et al.*, 2011). QBA includes not only the physical and behavioural component of the animal's emotional state but also the psychological by asking the observer to describe not what the animal is doing (e.g. walking, eating) but how it is performing that behaviour (e.g. calm, anxious) (Wemelsfelder, 2007).

Scientists have traditionally dismissed indicators that cannot be or have not been directly linked to the underlying emotional state as they are considered "anthropomorphic" (the attribution of human characteristics or behaviour to an animal) (Serpell, 2002). QBA however, is not defined as anthropomorphic but as a qualitative indicator which relies on the integration of perceived behavioural clues and the whole animal's demeanour. This type of assessment involves a more active role by the observer and their perception of the animals observed. This active, integrative role of the observer makes QBA more qualitative than other methods. However, there is no reason per se why a qualitative assessment method, such as QBA, could not also be developed and tested to meet scientific criteria for reliability and validity. Therefore, instead of dismissing this approach to assessing animals emotional states and the role that the observer has in welfare assessment, we should perhaps try to understand it and formalise it (Wemelsfelder, 1997). QBA thus defines emotion as an observable expressive quality of the whole dynamic of the animal being observed (Wemelsfelder, 2007). This is in contrast to other authors such as Mendl *et al.* (2010) that defined emotions in animals as a cognitive process giving an objective approach.

Previous research in farm animals has scientifically validated QBA for measuring animal emotional states by assessing its inter- and intra-observer reliability and comparing the QBA results with other objective indicators such as behavioural and physiological indicators in dairy cows (Rousing and Wemelsfelder, 2006), buffaloes (Napolitano *et al.*, 2012), sheep (Wickham *et al.*, 2012), and cattle (Stockman *et al.*, 2011). These studies used QBA in the observation of animal videos and used emotional expressions to describe emotional states. This description of emotional of states was cross-validated with other behavioural measures (e.g. social licking, fighting-(Rousing and Wemelsfelder, 2006)) and physiological measures (e.g. heart rate and core body temperature (Wickham *et al.*, 2012)). QBA has been successfully differentiated different emotional states in animals by using different behavioural expressions such as playful/sociable correlated with social licking in pigs (Wemelsfelder *et al.*, 2009) and relaxed/friendly correlated with no avoidance distance to the assessor in donkeys (Minero *et al.*, 2016). To date, there has been little or no research concerning the validity of this method for assessing the welfare of laboratory mice.

In this study, Qualitative Behavioural Assessment will be scientifically validated for the assessment of welfare in laboratory mice. The first step of validation is to assess the inter-observer reliability of QBA in mice using Free Choice Profiling (FCP) (Wemelsfelder *et al.*, 2001). FCP is a methodology widely used in food science, where observers use their descriptive vocabulary for assessing food properties (Collins, 1991). This methodology determines if the observers have similar ways of integrating all the information that they perceive into qualitative descriptors. Then, an agreement between participants can be calculated, even when different terminology is used, through General Procrustes Analysis (GPA) (Wemelsfelder, 2007). GPA calculates a consensus or 'best fit' profile between observers assessments through complex pattern matching. This consensus profile has some main dimensions (usually reduced down to 2 or 3) explaining the variation between animals (Wemelsfelder *et al.*, 2001; Fleming *et al.*, 2013; Fleming *et al.*, 2015). These dimensions represent animal expressivity, characterised in four quadrants which can be recorded in dimensional models (Wemelsfelder, 1997). These four quadrants are described by behavioural expressions such as good/calm (quadrant one), good/energetic (quadrant two), bad/calm (quadrant three), and bad/energetic (quadrant four) (Figure 3.1).

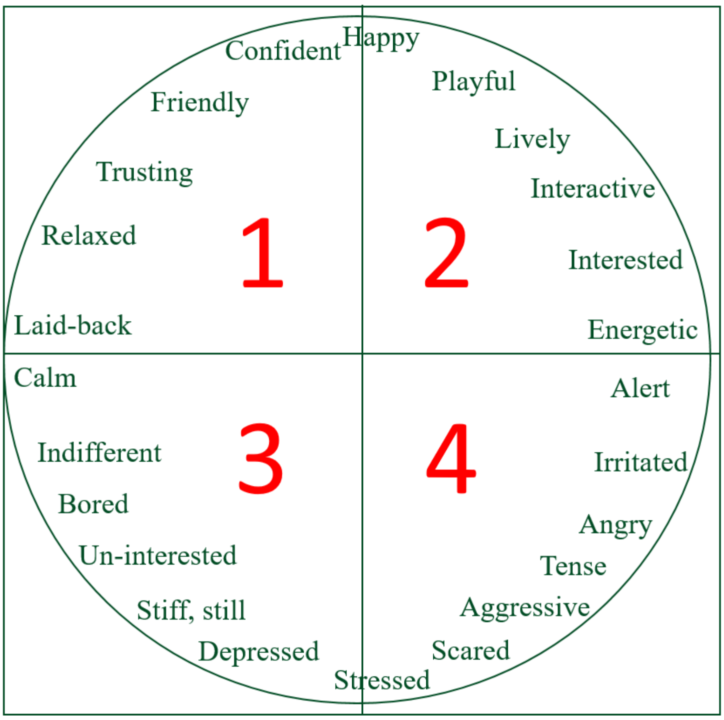


Figure 3.1. Representation of the quadrants described by the behavioural expressions in QBA using Free Choice Profiling. Numbers in red denote each quadrant which is represented by behavioural expressions. Modified from (Minero *et al.*, 2016).

The majority of variation is explained by the first dimension, with decreasing explanatory power for subsequent dimensions. Each animal receives a quantitative score on each of these dimensions. Interpretation of the consensus dimensions is made possible by identifying descriptive terms for each observer that correlate strongly with the consensus dimensions (Fleming *et al.*, 2015). This methodology has been used in other species such as pigs (Wemelsfelder *et al.*, 2001), dairy cows (Rousing and Wemelsfelder, 2006), foals (Minero *et al.*, 2009), calves (Brscic *et al.*, 2009) and buffaloes (Napolitano *et al.*, 2012).

This study aims to assess, through FCP, the inter-observer reliability of QBA for assessing laboratory mouse behavioural expression. Additionally, the observer's ability to differentiate between behavioural (behavioural test studies) and surgical videos will be assessed based on the premise that the surgical videos will likely have differences in terms of emotional states as these animals could experience pain to a certain level in comparison with then behavioural videos.

3.2. Materials and Methods

3.2.1. Ethical approval

This study involved the assessment of historical, pre-recorded videos of mice enrolled in unrelated studies. No studies were conducted to obtain the videos solely for this project. The project was conducted at Newcastle University following the registration for an unlicensed work (AWERB Project ID: 449), and in accordance with the EU Directive (2010/63/EU), ASPA (1986) and the NIH Guidelines for care and use of animals for experimental procedures.

3.2.2. Animals and videos footage

Video sequences of mice pre (baseline) and post a surgical procedure and videos of mice in behavioural studies were analysed in this study. Mice used strains including C57BL/6 and BALB/c female and males.

Surgical model: Mice were placed individually in a clear cage for 10 minutes, and pre-surgery baseline filming was carried out (Canon HD Legria HFM506) 24 hours before bile duct ligation and again at 24 hours post-surgery. All animals received pain relief (meloxicam 0.1mg/kg subcutaneously) immediately after surgery and 24 hours later, post-filming. This videos will be referred to as surgical videos in this document.

Behavioural models: The aim of the behaviour study was to assess the individual's habituation to different experimental situations. Mice were placed in the filming cages for ten minutes on ten consecutive days, initially in pairs and then individually to allow some level of habituation to the filming set up. The film was made in a quite separated room with the camera placed in front of the cage. On the final day, mice were video recorded for 10 minutes. This videos will be referred to as non-surgical in this document.

These types of video were selected as it was predicted they would show a broad spectrum of the expressive behavioural repertoire of the laboratory mouse (See figure 3.1).

A total of 54 videos were collected (23 surgical and 31 non-surgical) and were screened for inclusion in the study. The videos were selected by a researcher (ICL) by scanning all 54 videos and making a preliminary selection that was then revised by one expert in QBA and Free Choice Profiling studies. To be included, the video must represent one of the four quadrants indicated in Figure 3.1. The aim was to select videos that contain as much expressivity as possible and videos which represent the four quadrants (good/calm, good/energetic, bad/calm, and bad/energetic). The videos selected represented the whole animal expressivity or how the animal is "behaving" or carrying out certain behaviours with specific expression. This selection resulted in a final set of 20 videos, 9 of which were surgical videos (post-surgery) and 11 videos were non-surgical. One minute of video footage was selected from each video. This duration was selected based on previous QBA studies (Wemelsfelder *et al.*, 2001; Rousing and Wemelsfelder, 2006; Brscic *et al.*, 2009; Minero *et al.*, 2009; Napolitano *et al.*, 2012). The selected set of 20 videos were then integrated into a randomly ordered powerpoint presentation with a 60-second white screen between each video clip (see experimental procedures).

3.2.3. Observers

An email invitation was sent to all staff and students in two departments at Newcastle University, both of which have ties with animal science. The invitation contained the aims of the study (assess a novel indicator for the assessment of laboratory mouse welfare) and the experimental procedure. The selection criteria for the final participants were 1. Previous experience with laboratory animals (either working or studying in the past or the present) and 2. Voluntary participation in two sessions required to complete the study. Twenty observers agreed to participate in the study. The number of observers required was determined based on other QBA studies using Free Choice Profiling which suggested more than 10 participants are required for the GPA analysis (Wemelsfelder *et al.*, 2000). Participants were classified into two groups for data analysis: [1] experienced observers (who work with laboratory mice daily) (n=11) that comprised of animal technicians (n=6) and veterinarians (n=5); [2] inexperienced observers with laboratory mice (n=9) that comprised of researchers who occasionally worked with mice and other laboratory species (n=5) and MSc students studying animal welfare (n=4). This group 2, was labelled as inexperience for the data analysis but they have less experience in working with laboratory mouse.

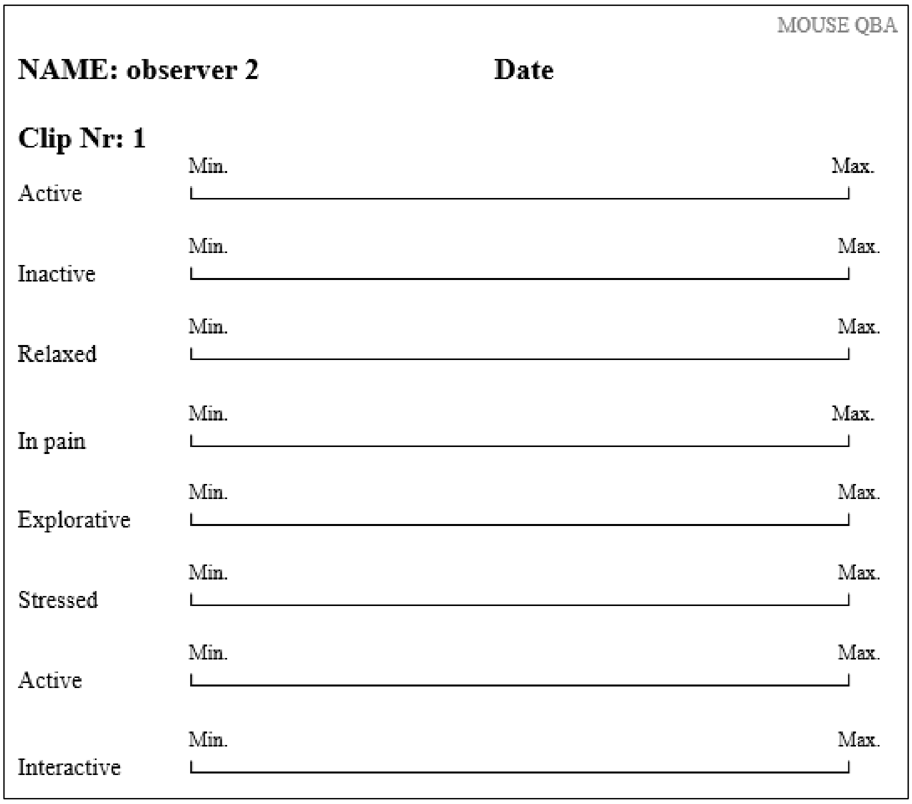
3.2.4. Experimental procedure

The Free Choice Protocol for the assessment of laboratory mouse behavioural expressions involved two phases: [1] the generation of terminology and [2] the use of participants own terminology as a quantitative measurement tool using a visual analogue scale (VAS). Participants took part in the study in a quiet, undisturbed room. Phase one took place in March 2016, phase two took place two weeks later.

In phase one participants received information about QBA, FCP and the aim of the study (Appendix A). Initially, the participants were provided with an example video sequence and were instructed to watch the videos and discuss what they have seen using behavioural expressions (e.g. calm, anxious, content, explorative). A guided discussion was carried with the group to ensure understanding of the instructions and aim of the study and the use of the terms providing examples of behavioural expressions after the example video. Following this, the main session started, and the observers were instructed to watch each 1-minute video, and then write down terms that best described the individual's behavioural expressions. They were asked to generate as many terms as they could for a given video clip. They could re-use terms for subsequent video sequences or create entirely new terms.

The score sheets for phase two were then generated by taking each participants list of terms and organising them into a score sheet (Figure 3.2), one for each video sequence.

In Phase Two, the 20 individually created score sheets were given to participants. Accompanying each of the terms they had used in phase, was a Visual Analogue Scale (VAS) that was 12.5cm long with the terms 'Minimum' at the left-end (0cm) and 'maximum' on the right-end (12.5cm) (Figure 3.2).


MOUSE QBA

NAME: observer 2 **Date**

Clip Nr: 1

Active Min. Max.

Inactive Min. Max.

Relaxed Min. Max.

In pain Min. Max.

Explorative Min. Max.

Stressed Min. Max.

Active Min. Max.

Interactive Min. Max.

Figure 3.2. Example of the score sheet for observer two video 1 in the second stage of Free Choice Profiling used for QBA. Each observer had their personal list of terms for each video linked with a Visual Analogue Scale (VAS).

Each observer was instructed to watch the video sequence again and score each term using the VAS. They were told that the VAS is a continuous line with no divisions or categories and they can interpret the scale intuitively as there is no ‘objective’ measurement rule for how to use the scale. The left point is the 0-point or minimum meaning that the characteristic is entirely absent (e.g. the animal is not, e.g., relaxed at all). The right point is ‘maximum’, which means it is entirely dominant (e.g. the animal could not be more relaxed).

3.2.5. Data analysis

Each of the 20 observers produced a set of VAS scores, for each of the 20 videos. For each term, a score was determined by measuring the distance in millimetres from the left of the scale to the point where the observer crossed the line. This data was entered into an excel sheet in a data matrix (Figure 3.3).

	A	B	C	D	E	F	G	H	I
1	**Subject$**	**Miserable**	**Depressed**	**Alert**	**Inquisitive**	**Agitated**	**Stressed**	**Apathetic**	**Distressed**
2	**Mouse1**	79	72	25	9	11	78	42	74
3	**Mouse2**	1	1	113	114	22	40	1	0
4	**Mouse3**	5	4	58	22	58	35	3	13
5	**Mouse4**	2	12	45	78	36	50	12	35
6	**Mouse5**	7	5	82	38	91	83	4	6
7	**Mouse6**	5	6	58	25	45	59	2	4
8	**Mouse7**	5	4	95	84	105	68	2	2
9	**Mouse8**	3	2	91	91	75	63	2	2
10	**Mouse9**	3	3	87	58	83	87	2	3
11	**Mouse10**	50	33	31	34	30	55	24	48
12	**Mouse11**	2	3	98	95	87	77	2	2
13	**Mouse12**	5	4	63	62	93	91	2	2
14	**Mouse13**	1	0	109	84	89	8	1	0
15	**Mouse14**	0	2	55	35	123	112	1	1
16	**Mouse15**	1	2	107	109	30	70	1	2
17	**Mouse16**	2	2	97	98	39	65	1	0
18	**Mouse17**	0	2	65	68	112	104	2	29
19	**Mouse18**	38	2	49	46	37	69	2	44
20	**Mouse19**	3	3	61	62	119	103	1	32
21	**Mouse20**	1	2	102	97	59	58	1	2

Figure 3.3. Example of the data matrix for one observer in the FCP session two. Each matrix was defined by the individual mouse videos (1 to 20) and the terms created by each individual observer (e.g. 8 terms for this observer). The scores for each represent the distance in millimetres from the left of the VAS to the point where the observer tick crosses the line (e.g. tick the VAS line at 72mm for mouse video number 1 when assessing “depressed”).

Inter-observer agreement within and between observer groups (experienced and inexperienced) was analysed using General Procrustes Analysis (GPA). The observer’s measurements of the distances between the mouse videos were compared. GPA is a pattern-matching tool which compares the differing terms observers used for a given video (Wemelsfelder *et al.*, 2001). This statistical analysis transforms each observer matrix into a multidimensional configuration (Xiong *et al.*, 2008). Using several geometric transformation phases which are standard for this statistical analysis, GPA can determine the similarity between the configurations which then produce the “consensus profile” or the best common denominator profile (Wemelsfelder *et al.*, 2000). This consensus profile is created from the mean of the transformed configurations that GPA makes to the data (Xiong *et al.*, 2008; Napolitano *et al.*, 2012). This consensus profile is quantified using “the goodness of fit” by the Procrustes Statistics (PS). A consensus profile was calculated for each of the following three observer groups; all observers, ‘experienced’ observers (veterinarians and technicians [n=11]) and ‘inexperienced’ observers (researchers and students [n=9]). A 1-tailed Student t-test (t_{99} = 76.08, p< 0.001) was used to determine if the consensus PS differed significantly from the randomised Procrustes statistics (performed by the GPA analysis) indicating that the profile was meaningful for the dimension that it represents and not a statistical artefact (Xiong *et al.*, 2008).

GPA calculates a consensus profile for all observers independently of any interpretative judgment by the experimenter. Using Principal Component Analysis (PCA) the number of dimensions of the consensus is reduced to one or more main dimensions that explain the variation between the videos. These dimensions are interpreted by correlating them with the original observer data matrices producing two-dimensional individual observer word-charts (Figure 3.4). These observer word-charts were used for the interpretation of the main dimensions as the higher a term correlates with an axis, the more weight it has as a descriptor of that axis.

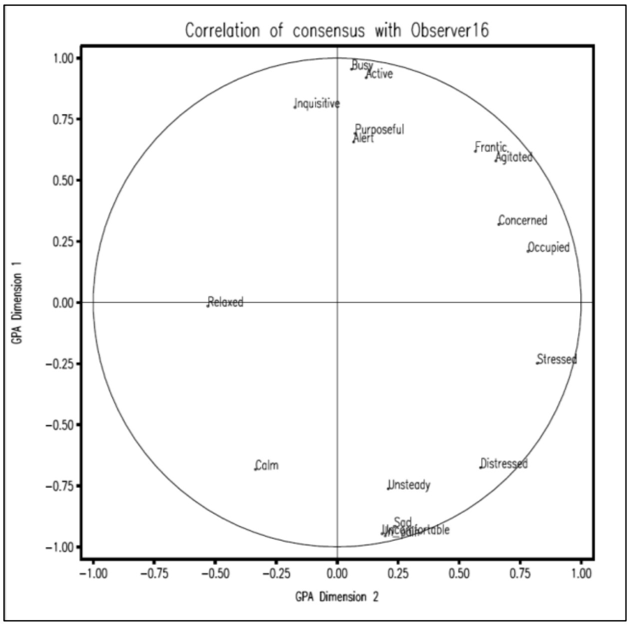


Figure 3.4. Two-dimensional correlation of consensus word chart for one observer. Dimension one is in the y-axis and dimension two is located on the x-axis. Each descriptor has been allocated into the consensus profiled taking into account the GPA scores.

Spearman correlation coefficient was used for assessing if the behavioural expressions from the observers were used in the same way in dimension one and two. The assessment of the effect that the nature of the videos (i.e. non-surgical and surgical) had on the GPA for each video was assessed using a Kruskal-Wallis test. The scores for both dimensions for group 1 (experienced), group 2 (inexperienced) and combined group (all the observers together) were used for this analysis. A Kruskal-Wallis test was used to assess if the difference in the videos (non-surgical and surgical) affected the GPA scores. A Mann-Whitney test was used as a p*ost hoc* test. The statistical analyses were performed using GenStat 16 data analysis software (GenStat 2014, VSN International, Hertfordshire, UK) and SPSS (SPSS Inc., Chicago, IL).

3.3. Results

3.3.1. Consensus profiles

Two main dimensions of the consensus profiles were identified, explaining 54.2% and 23% for combined group (all observers); 53.1% and 26.7% for Group 1 (‘experienced’ observers; technicians and veterinarians) and 55.5% and 19.4% for Group 2 (‘inexperienced’ observers; researchers and students). These two main dimensions, which were produced by the GPA analysis, explain the variation between the laboratory mouse videos.

Table 3.1 shows, for the separated and the combined group analysis, the consensus profiles which explains the variation among the observer’s assessment compared with the mean of 100 randomised profiles.

Procrustes statistics	Combined group	Group 1	Group 2
Consensus profile	73.33	75.86	74.12
Mean randomized profile ± SD[1]	41.96 ± 0.17	44.21 ± 0.60	48.02 ± 0.58
t_{99}	20*	17*	19*

*p < 0.001. [1] Mean of 100 Procrustes Statistics values obtained through 100 Generalised Procrustes Analysis of randomised data matrices.

Table 3.1. Procrustes statistics used in QBA-FCP study. The table shows the separate and combined analyses of the observer groups. Group 1 - technicians & veterinarians and group 2 - researchers & students.

The Procrustes statistical analysis showed significant inter-observer reliability with the consensus profiles, explaining a significantly higher percentage of variation between observers than the mean of the 100 randomised profiles ($p < 0.001$). The observer plots of the combined analysis (a), group 1 (b) and group 2 (c) are shown in

Figure 3.5. The combined observer's group showed some outliers; observers 5, 9, 18, 19 and 20. The Procrustes analysis carried out for group 1 and 2 also has outliers (observers 10 and 3, and 2 and 4 respectively), these outliers are different from the combined group because the GPA analysis was carried out independently for the groups.

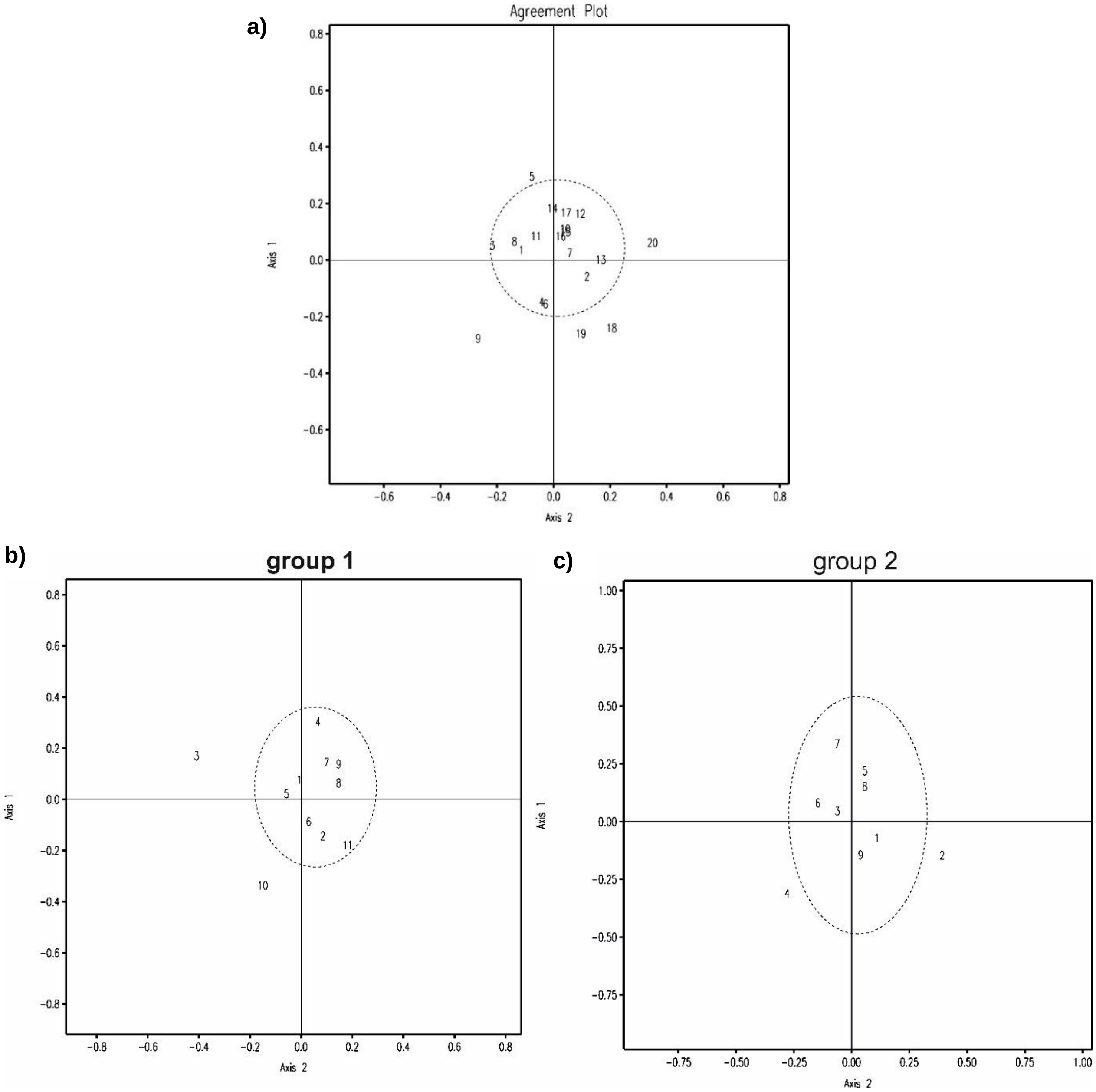


Figure 3.5. Observer plots for the three different groups (combined group (a), group 1 (b) and group 2 (c)) used for the QBA-FCP data analysis. The axes represent the Principal Coordinates analysis scaling for relative observer distance. Each number represents an individual observer. The ellipse represents the 95% confidence region of the "normal population" so the observers outside this region are considered outliers.

The observer word charts that interpreted each of the dimensions, in each group, were provided with a similar group of words with the same meaning. For example, observer 5 (Figure 3.6a) from the combined group shows that dimension one was defined from ‘active’ to ‘in pain’ and the dimension two from ‘worried’ to ‘calm/relaxed’. Observer 19 (Figure 3.6b) from group 1 shows that dimension one was defined from ‘active to dull’ and dimension two from ‘panicked’ to ‘calm’. Observer 8 (Figure 3.6c) from group 2 shows that dimension one was defined from ‘active/energetic’ to ‘dull/unwell’ and dimension two from ‘determined/panicked’ to ‘calm/comfortable’. The complete list of the 20 observer charts is in Appendix B.

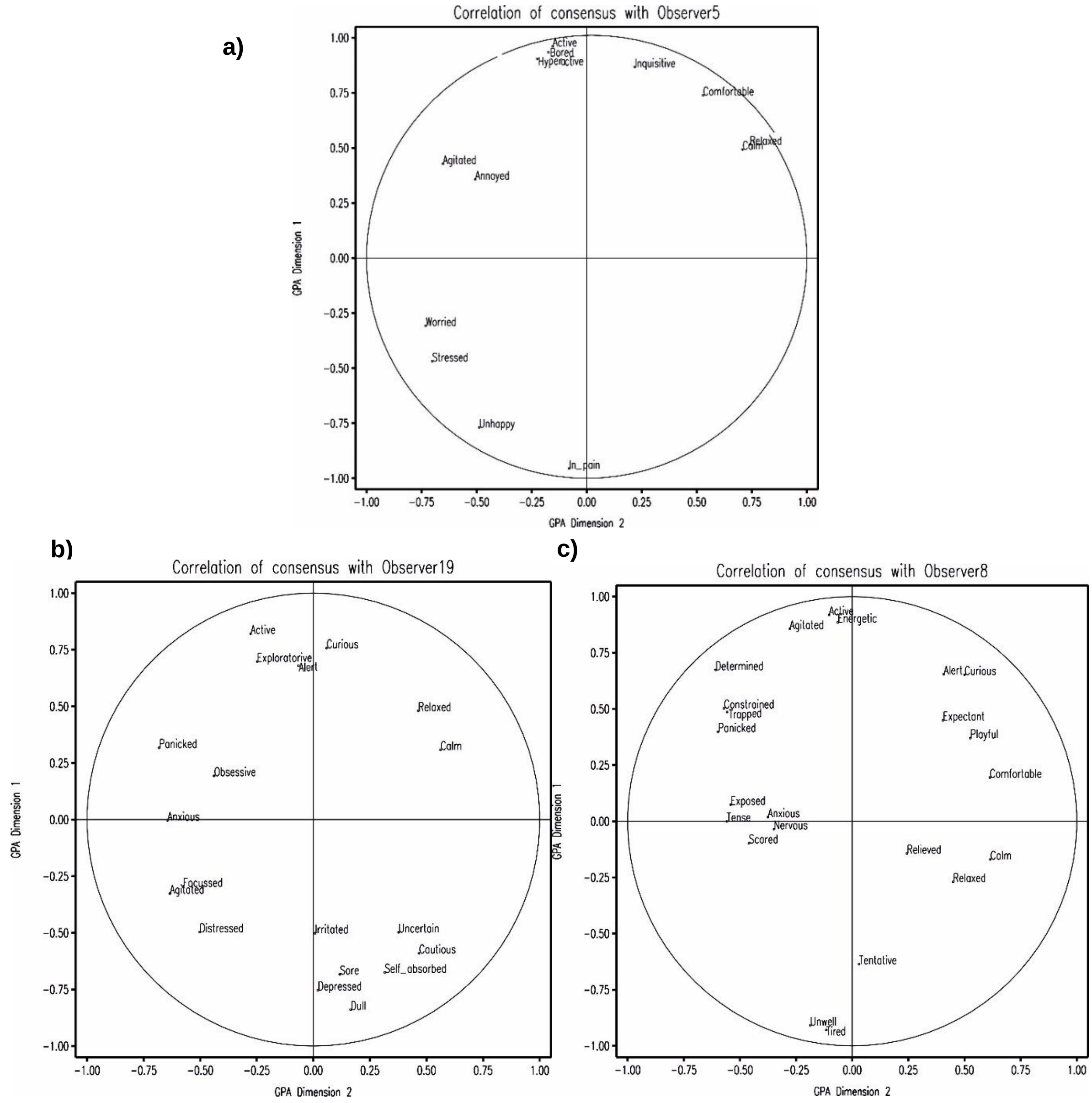


Figure 3.6. Word charts of observer 5 (a) from the combined group, 19 (b) from group 1 and observer 8 (c) from group 2. The axes represent a strength of correlation with consensus dimension 1 (y-axis) and 2 (x-axis).

Table 3.2 shows a list of the most positive and negative correlation terms used by the observers in the three groups. The terms that have a higher correlation than 0.7 for dimension 1, and higher than 0.5 for dimension 2 were extracted from the analysis and are showed in Table 3.2. The threshold of correlation selected for both dimensions was based on other studies (Wemelsfelder *et al.*, 2001; Rousing and Wemelsfelder, 2006). The threshold was set for dimension 1 taking into account the highest scores from a range between -0.9 to 9.0 and the highest scores for dimension 2 from a range between -0.7 to 0.7. The labels of the dimensions for the combined group were labelled from "alert/inquisitive" to "in pain/lethargic" for dimension 1 and from "calm/relaxed" to "agitated/stress" for dimension 2. For group 1, the dimensions were labelled from "inquisitive/alert" to "in pain/depressed" for dimension 1 and from "calm/relaxed" to "stressed/agitated" for dimension 2. For group 2, the dimensions were labelled from "agitated/alert" to "in pain/tired" for dimension one and from "relaxed/calm" to "panicked/distressed" for dimension two. This list of terms used by the observers had a distinctive pattern characterised by the high use of the same words among all groups, with high positive correlated terms such as inquisitive and calm; and the use of high negative correlated terms such as in pain and stressed.

	Positive correlations	Negative correlations
Combined group		
D 1	Active (9), Alert (6), Inquisitive (6), explorative (3), lively (3), curious (2), energetic (2), hyperactive (2), interested (2), agitated (1), aware (1), bored (1), busy (1), determined (1), Drowsy (1), excited (1), in pain (1), nervous (1), responsive (1), tired (1)	In pain (9), tired (6), depressed (4), lethargic (4), sad (4), Uncomfortable (3), unwell (3), calm (2), distress (2), dull (2), inactive (2), quiet (2), sore (2), apathetic (1), awkward (1), dejected (1), drowsy (1), exhausted (1), miserable (1), motionless (1), reluctant (1), resigned (1), self-absorbed (1), stable (1), stiff (1) strained (1), tense (1).
D 2	Calm (11), relaxed (7), happy (3), content (2), curious (2), alert (1), playful (1).	Agitated (8), stressed (8), anxious (3), determined (3), frantic (3), nervous (3), panicked (3), frighten (2), neurotic (2), scared (2), unhappy (2), aggressive (1), bothered (1), claustrophobic (1), concerned (1), constrained (1), desperate (1), irritable (1), obsessed (1), occupied (1), stirred (1), tense (1), upset (1), worried (1).
Group 1		
D 1	Inquisitive (7), active (5), alert (5), explorative (4), curious (3), Comfortable (2), excited (2), hyperactive (2), interested (2), tentative (1), aware (1), bored (1), busy (1), energetic (2), happy (1), investigative (1), lively (1), playful (1).	In pain (6), depressed (5), lethargic (3), sad (3), tired (3), cautious (2), dull (2), quiet (2), reluctant (2), sore (2), apathetic (2), awkward (2), bored (1), miserable (1), self-absorbed (1), shaky (1), stiff (1), strained (1), tense (1), unwell (1).
D 2	Calm (8), relaxed (5), content (2), happy (2) playful (1).	Stressed (7), agitated (6), anxious (4), nervous (3), aggressive (2), distressed (2),

		frantic (2), scared (2), tense (2), afraid (2), annoyed (1), bothered (1), crazy (1), desperate (1), determined (1), excited (1), focussed (1), frightened (1), hyperactive (1), mad (1), neurotic (1), panicked (1), stirred (1), upset (1), worried (1).
Group 2		
D 1	Active (5), agitated (3), alert (3), determined (2), inquisitive (2), lively (2), nervous (2), annoyed (1), busy (1), energetic (1), explorative (1), hyperactive (1), responsive (1), restless (1), unsettle (1).	Calm (3), in pain (3), tired (3), cautious (1), dejected (1), depressed (1), drowsy (1), exhausted (1), helpless (1), inactive (1), lethargic (1), motionless (1), resigned (1), sad (1), stable (1).
D 2	Relaxed (4), calm (3), curious (2), inquisitive (2), Comfortable (1), happy (1), playful (1)	Panicked (3), distressed (2), neurotic (2), obsessed (2), stressed (2), tense (2), anxious (1), constrained (1), determined (1), exposed (1), frantic (1), frightened (1), irritable (1), scared (1), trapped (1).

Table 3.2. Terms used by the observers describing each dimension (D) in each group in QBA-FCP data analysis. All terms with a correlation over 0.7 for dimension 1 and 0.5 for dimension two were included in this list. The terms are divided into positive and negative correlation with the dimension and the number of observers who used the term – denoted by a number in brackets following the terms.

3.3.2. Inter-observer reliability

Figure 3.7 shows the mouse plots for the combined analysis (a), group 1 (b) and group 2 (c). This plot was created from the GPA consensus scores of the observers for the 20 mouse videos. The position of individual mice is represented by the letter M and the number of the video.

In the combined group, the dimensions explain 77.9% of the variation among video clips (dimension 1: 54.2% and dimension 2: 23.7%). For group one, the dimensions explain 79.8% of the variation (dimension 1: 53.1% and dimension 2: 26.7%). Finally, for group 2, the dimensions explain 74.9% (dimension 1: 55.5% and dimension 2: 19.4%) of the variation. This variation in the description of the behavioural expressions between the videos can be explained by the use of the terms chosen by the observers. For example, M2 (mouse video 2) was described as 'inquisitive/alert' and 'calm' by the three groups whereas M1 was described as 'lethargic/in pain' and 'slightly calm' by the three groups. The labels of the dimensions are very similar for all three groups as well as the location of each mouse video in the plot.

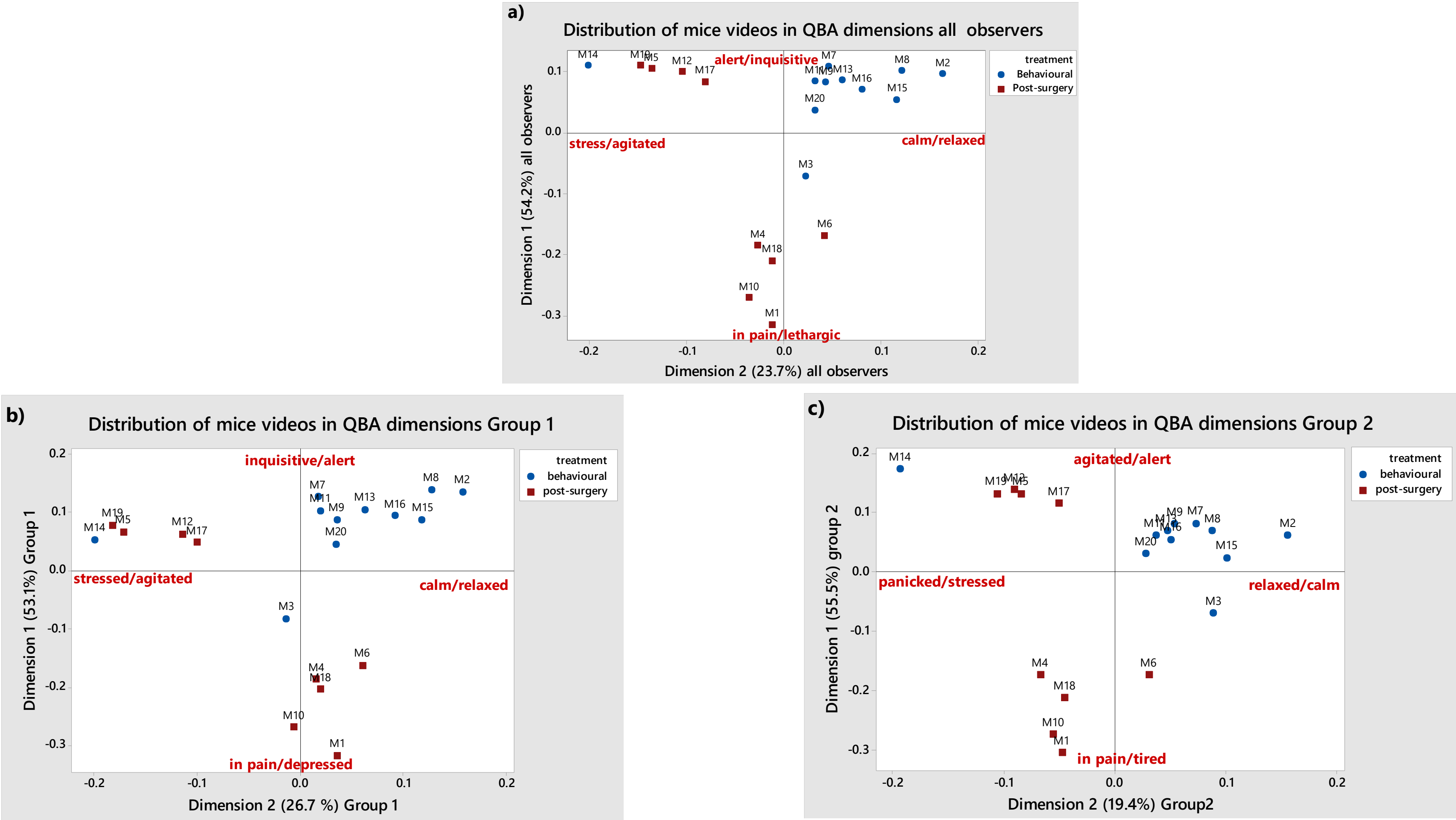


Figure 3.7. Distribution of mice videos in the QBA-FCP analysis. Figure (a) shows the mouse plot for the combined group, figure (b) for group 1 (technicians and veterinarians) and figure (c) for group 2 (researchers and students). Axes represent the scaling values for the sample (mouse videos) distance on both dimensions of the consensus profile. Each mouse video is represented by the letter M and a number. Each dimension has the percentage of the variation of the videos provided by each dimension. The treatment is also shown in the figure with a blue circle for non-surgical videos (behavioural) and red square for surgical (post-surgery) videos.

The GPA scores for each video clip were used to assess whether the groups used the dimensions similarly for assessing the videos. Spearman correlation coefficient considers a strong correlation if the coefficient is greater than 0.7, moderate if it is from 0.69 to 0.50 and weak from 0.49 to 0.0 (Bolboaca and Jäntschi, 2006). The correlation coefficient between group 1 and 2 was moderate for dimension one: Spearman rs = 0.55, $p<0.001$ and strong for dimension 2: rs = 0.78, $p<0.001$.

Further inspection of the distribution of the mouse videos in Figure 3.7, shows that the videos are distributed on the plots in a triangle shape, with videos in three out of the four quadrants: good/calm (quadrant one), good/energetic (quadrant two), and bad/energetic (quadrant four). There were fewer videos allocated in quadrant three (bad/calm).

For assessing the effect of the nature of the video sequences (surgical or non-surgical) on the GPA scores for both dimensions in the combined group and groups 1 and 2, a Kruskal-Wallis test was performed as the residuals of the GPA scores for both dimensions were not normally distributed (Table 3.3).

	D1 Combined group	**D1 group 1**	**D1 group2**	**D2 combined group**	**D2 group 1**	**D2 group2**
Kruskall-Wallis	1.669	8.562	.522	8.129	3.905	9.006
df	1	1	1	1	1	1
Exact Sig.	.196	.003	.470	.004	.048	.003

Table 3.3. Kruskal-Wallis test for the effect of the nature of the videos (no surgery or surgical) on the GPA scores for dimensions 1 (D1) and 2 (D2) in the combined group, group 1 (technicians and veterinarians) and group 2 (researchers and students).

GPA scores for dimension 1 were not significantly affected by the type of clips (no-surgical and surgical) for combined group, H (1) = 1.67, $p > 0.05$ and group 2 ('inexperienced' observers; researchers and students), H (1) = 0.52, p> 0.05. However, there was a significant effect of the type of video clip in dimension 1 for group 1 ('experienced' observers; technicians and veterinarians), H (1) = 8.56, $p < 0.05$. *Post hoc* analysis was made using Mann-Whitney test, the non-surgical videos overall had higher GPA scores compared with the surgical videos for the experienced group of observers (group 1), (U= 11, r = -0.6, p < 0.05). The GPA scores were affected by the type of clip (surgical or non-surgical) in all groups for dimension 2, H (1) = 8.13, p < 0.05 for the combined group, H (1) = 3.90, p < 0.05 for group 1, and H (1) = 0.52, p < 0.05 for group 2. *Post hoc* analysis was carried out using Mann-Whitney test, the non-surgical videos overall had higher GPA scores compared with the surgical videos for the combined group (U= 12, r=-0.6, p = 0.03), group 1 (experience observers) (U=23.5, r=0.4, p=0.46) and group 2 (inexperienced) (U=10, r=0.6, p=0.02).

3.4. Discussion

This study aimed to assess the inter-observer reliability of QBA Free Choice Profiling as a method to assess laboratory mouse behavioural expressions and to assess observer ability to differentiate between non-surgical and surgical videos using FCP. High levels of inter-observer agreement were found in the qualitative assessment of behavioural expressions of the video sequences of mouse. This result is consistent with other studies about QBA reliability in various species, e.g. pigs, dairy cows, foals and calves (Wemelsfelder *et al.*, 2001; Rousing and Wemelsfelder, 2006). Using Free Choice Profiling, two groups of observers ‘experienced’ and inexperienced’ in laboratory mice showed significant agreement in their assessment of 20 video sequences showing mice in two conditions; non-surgical and surgical, and were also able to differentiate between videos of these two conditions consistently. The surgical videos were generally assessed as more ‘lethargic’ and ‘in pain’ compared to those sequences showing mice in the no-surgical recordings, which were assessed as more ‘calm’ and ‘relaxed’.

The consensus profile showed a high level of agreement producing two principal dimensions which explained 74% to 79% of the variation among the videos (Wemelsfelder *et al.*, 2012). Despite the difference in the level of experience between the observers, they developed a similar vocabulary for describing laboratory mouse emotional states which were highly correlated with the two dimensions as demonstrated in the GPA analysis (Table 3.2). These results are similar to other QBA studies where people with different experience and professional backgrounds had a high degree of agreement, e.g. pig QBA assessment where pig farmers, veterinarians and animal protectionist all used similar terms (Wemelsfelder *et al.*, 2012). These results support the hypothesis that QBA may be a promising measurement for assessing laboratory mouse welfare, even when the assessors have varying levels of experience. In contrast to other means of assessment of laboratory mouse emotional states, QBA assesses welfare through the observation of the whole dynamic animal and its interaction with the environment and with others (Wemelsfelder, 2007). Behavioural measurements, for example, assess emotional states by breaking down the behaviours into single components (e.g. walking, resting, sleeping, exploring) but they do not give additional information about how the animal is performing those behaviours, about the underline emotion that is behind the behaviour (e.g. walking in a calm way, exploring anxiously) (Wemelsfelder, 2007). By introducing QBA in the assessment of laboratory mouse welfare, we will be able to provide scientific validity to the expressions that we use every day when assessing welfare as this methodology aims to formalise those everyday expressions scientifically. On a daily basis, in a laboratory facility, the experienced personnel who spend most of the time looking after the mice (e.g. technicians, veterinarians) have an informal language for discussing the animal's psychological state [e.g. insecure, excitable, curious when talking about animals personalities (Stevenson-Hinde and Zunz, 1978; Stevenson-Hinde *et al.*, 1980)]. They use expressions such as "not happy", "stress", "inquisitive", "the animal is not fine" when describing the general demeanour of mice. This study provides the first insight into the scientific validation of those expressions as a means for assessing laboratory mouse emotional states.

The correlation between the two groups for dimension 1 was moderate suggesting that the assessment of this dimension may not be as straightforward as the assessment of dimension 2, which had a strong correlation. A possible explanation for this may be related to the assessment of mice in pain against calm mice. Dimension one is defined as tired/agitated which suggests it is linked with energy levels. Dimension two is defined as stress/calm which suggests it is more linked with valence (the degree to which an individual's emotional experience is generally positive or negative, (Eaton and Funder, 2001)). The difficulty in the assessment of dimension one is also illustrated by the mouse plot (the location of the mouse videos in the two dimensions) where the videos were allocated in three different quadrants resembling a triangle-shaped structure. Observers were able to differentiate between stress/agitated and calm/relaxed mice, but they were not able to see a contrast between pain and calm/relaxed mice. Due to high levels of activity shown by laboratory mice, the assessment of an animal in pain vs calm might be very challenging if the observers have little or no experience/knowledge of normal mouse behaviour as a decrease in the levels of activity can be interpreted as either, calmness or pain depending of the context and the knowledge about laboratory mouse natural behaviour (Kaliste, 2004). An additional point to consider is related to the triangle shape of where the videos were allocated in the plot is the assessment of the effect of type of videos (surgical and non-surgical) on the GPA scores. For dimension one, the inexperienced group (group 2) were not affected by the nature of the videos, supporting the difficulty in assessing laboratory mouse dimension 1 when there is a lack of experience about the animals being assessed. These lack of experience can be related to the observer's knowledge of pain-specific behaviours in laboratory mouse which can affect the assessment of the videos including the differentiation between pain and calm mice as their activity levels and specific behaviours can decrease (Association of Veterinary and Research, 1989; Anil *et al.*, 2002). Another possible explanation for these results can be related to the duration of the videos. Other studies have assessed pain in laboratory mice using longer videos as the pain-related behaviours can be displayed in a more extended period of time such as over 5 minutes (Flecknell, 1986; Chesler *et al.*, 2002) or 20 minutes (Miller *et al.*, 2011). These results were evidenced by the non-significant effect found in dimension 1 for the group with less experience with laboratory mice (e.g. group 2- researchers and students).

Contrary to this result, other QBA studies have not found this triangle-shaped distribution of the videos in the plot (Napolitano *et al.*, 2012; Wemelsfelder *et al.*, 2012). However, the animals assessed were farm animals with different patterns of emotional expressions, so a straightforward comparison cannot be made. In addition, the differences in behaviour between the species are vast as they have different behavioural patterns and activities levels.

Furthermore, even within the same species, laboratory mice activity levels can be very different. The same animals after and before surgery can have high levels of activity if the pain is minimal (Association of Veterinary and Research, 1989). More research is needed to determine if this triangle-shaped location of the mouse videos in the plot obtained in this study was because of normal mouse behaviour or absence of videos which represent the four quadrants. In the selection of the videos for this study, the aim was to have videos for each of the four quadrants ([1] good/calm, [2] good/energetic, [3] bad/calm, and [4] bad/energetic). The videos were allocated in all the quadrants based on a previous selection made for the researcher, but there is a clear pattern in the localisation of the videos (triangle shape) made by the observers.

The findings of this first study using Qualitative Behavioural Assessment as a means of assessing laboratory mouse behavioural expressions provides valuable information about the qualitative judgment of the observer. The use of observer judgement as part of animal welfare assessment has been demonstrated in other species such as monkeys (Stevenson-Hinde and Zunz, 1978; Stevenson-Hinde *et* *al.*, 1980), domestic cats (Feaver *et al.*, 1986) and dogs (Walker *et al.*, 2010). This study has shown that QBA can be used for the assessment of laboratory mouse welfare as the observers agreed when using terms for describing animal's emotional states. The behavioural expressions used by the observers in this study allowed the assessment of animal welfare states using descriptors which describe mouse psychological health. These expressions described videos of surgical mice as lethargic and in pain and non-surgical mice videos as calm and relaxed. These terms are correlated with the assessment of psychological states in other species (Stevenson-Hinde and Zunz, 1978; Stevenson-Hinde *et al.*, 1980; Bateson, 2004). Observers were also able to distinguish between non-surgical and surgical animals by using negative and positive descriptive terms.

However, given the nature of the study, QBA free choice profiling is not feasible to be used as a routine welfare indicator due to its lack of practicability. The assessment of mouse behavioural expressions using Free Choice Profiling involves the creation of a set of terms for each observer making the process lengthy and time-consuming. As has been done in other species such as calves (Brscic *et al.*, 2009), cattle (Andreasen *et al.*, 2013) and sheep (Phythian *et al.*, 2013) further validation of the indicator is needed using 'fixed terms' (behavioural expressions which are preselected terms, which take into account an inclusion/exclusion criteria for QBA studies) for improved practicality in the assessment of laboratory mice welfare within an animal facility.

Chapter 4. Qualitative Behavioural Assessment: Validation of fixed behavioural expressions in laboratory mice

4.1. Introduction

In order to implement QBA as an indicator for the assessment of the welfare of laboratory mice, the development of a fixed behavioural expression list is necessary (Brscic *et al.*, 2009). The fixed behavioural expression list will allow increased practicability of the indicators by reducing the data analysis during assessment and facilitating its interpretation (Andreasen *et al.*, 2013). Chapter three involved the initial scientific validation of this tool using FCP profiling. The next step towards the validation of this method is the validation of the fixed term list. The interpretation of QBA results is based on dimensions from the statistical analysis and the identification of the expressive terms (e.g. in pain, relaxed, fearful) that best describe these components (Wemelsfelder, 2007). These terms provide information about the emotional state of the animals as they describe underlying emotions, identifying both positive and negative emotional states (Wemelsfelder, 2007). These main dimensions have been previously used for the assessment of animal emotions in QBA (Wemelsfelder, 2007) and the assessment of animal personalities (Stevenson-Hinde and Zunz, 1978; Stevenson-Hinde *et al.*, 1980). The use of fixed behavioural expressions has been validated in other species. For example, Minero *et al.* (2016), found a good correlation between the descriptors linked with positive emotional states (relaxed, at ease, happy) and behavioural measures (no avoidance from the assessor and no tail tuck) in donkeys. Furthermore, QBA has been included in official European Union welfare assessment protocols (e.g. Welfare Quality Program) and in the On-farm welfare assessment protocol for Dairy Goats (ARWIN) (Battini *et al.*, 2015). The high reliability of QBA in other species and the inclusion as an animal welfare method in a European Union welfare protocol makes this indicator a feasible emotional assessment tool for animals; therefore, research in laboratory species is required.

The FCP study detailed in Chapter 3 showed a good level of inter-observer reliability in the assessment of emotional states in laboratory mice (Spearman $r_s = 0.55$ for dimension 1 and $r_s = 0.78$ for dimension 2). The statistical analysis of the fixed behavioural expressions is carried out using Principal Component Analysis (PCA).

PCA is used to reduce the variables into statistically relevant principal components (PC) or dimensions, which summarise the animal's behavioural expressions (Muri and Stubsjøen, 2017). For example, Rutherford *et al.* (2012) showed the terms confident/curious and unsure/nervous correlated with dimension one whereas agitated/angry and calm/relaxed were associated with dimension two in a pig study. The Rutherford study aimed to assess the validity of a fixed behavioural expression list and their correlation with other welfare measures in an open field test in pigs which underwent the usage of a relaxing drug (Azaperone). Participants were able to define both dimensions by the use of fixed behavioural expressions as well as distinguish between the treatments describing the animals with Azaperone as 'curious/confident' with higher scores in dimension one (Rutherford *et al.*, 2012). The use of a standardised list of behavioural expressions which can be analysed using PCA is more feasible than the Free Choice Profiling for an assessment carried out in a laboratory animal facility as the data management and analysis is more practical for a welfare assessment.

The main concern when using QBA to assess emotional states in animals is that the indicators are based on the active role of the observer, as the use of these indicators requires that the analysis and incorporation of all the observations made, integrating the perceived details of behavioural expression, which by most scientists is considered a subjective approach (Duncan, 2005b). QBA integrates the qualitative role of the observer into the assessment of welfare which can be measured scientifically through the assessment of inter-observer and intra-observer reliability (See Chapter 3 for details). Such analysis confirms the degree to which the measurements recorded by different observers provide similar results (Muri and Stubsjøen, 2017). This type of study, where rating scores are used, is common in laboratory mice such as in nest quality scoring (Gaskill *et al.*, 2013b) and Body Condition Scoring (Ullman-Culleré and Foltz, 1999), providing more evidence of the validity and practicality of this type of assessment for the welfare of laboratory mice.

This study includes an experimental paradigm related to handling methods and anxiety in laboratory mice. Anxiety is considered as "an emotional response, typically unpleasant, which involves heightened arousal and attentiveness to the environment, and typically inhibiting action to the perception of a threat to one's well-being or one's ego (sense of self)" (DeGrazia and Rowan, 1991). As an unpleasant emotional response, anxiety can be detrimental for animal welfare and since the aim of QBA is the assessment of laboratory animal's psychological state, the use of an experimental model that includes a potential anxiety-related component could be valuable for the validation of a QBA fixed behavioural expression list as an indicator of laboratory mouse welfare. Handling is the most common procedure carried out on laboratory mice and is most commonly done by picking the mouse up by the tail (Gouveia and Hurst, 2013).The mouse is gripped by the base of the tail and picked up to be examined or transferred to another cage. Recent studies have demonstrated that mice handled by the tail show increasing levels of anxiety and stress compared to mice handled using a tube which they are allowed to get in freely (Deacon, 2006; Hurst and West, 2010) thus improving their welfare (Hurst and West, 2010; Gouveia and Hurst, 2013). In these studies, anxiety was measured using different validated methods including the elevated plus maze, open field and interactions test using the different handling method (Hurst *et al.*, 1999; Hurst and West, 2010; Gouveia and Hurst, 2013). Furthermore, handling laboratory mice using a tube instead of the tail has been shown to improve quality of the data collected in behavioural studies by increasing the response to reward (Clarkson *et al.*, 2018) linked with reduced anxiety (Hurst and West, 2010).

Experiments involving handling methods which produce anxiety in laboratory mice are relevant for this study, as they involve a behavioural expression (anxiety) thus it can be used to assess to what extent the participants (who were blinded to the handling treatment) were able to perceive a difference between the two handling methods regarding emotional expressions in laboratory mice. This study aimed to assess the validity of QBA as an indicator of emotional states in laboratory mice by using fixed behavioural expressions when observing mice that have been handled using two different methods (tail and tube). Also, the assessment of intra-observer reliability will be made to assess the extent of agreement of the participants when assessing the same videos on two separate occasions.

4.2. Materials and methods

4.2.1. Ethical Statement

Experiments were conducted at Newcastle University following the registration for unlicensed work (AWERB Project ID: 449), and in accordance with the EU Directive (2010/63/EU), ASPA (1986) and the NIH Guidelines for care and use of animals for experimental procedures. Mice were checked daily by technical staff, and no adverse effects were reported.

4.2.2. Selection of QBA fixed behavioural expressions list

A group of 4 animal behaviour and welfare experts (one veterinarian, two experts in laboratory mice and one expert in QBA) were involved in the selection of the fixed terms. The experts were recruited via email and personal communication taking into account the levels of expertise required for the selection of the fixed terms. The QBA fixed terms were selected from the Free Choice Profiling study (Chapter 3) and from previous fixed-term QBA studies. A total of 91 different terms were created for the assessors, each with a positive or negative correlation with dimension one or two. Thirty-one terms which were highly correlated on to the dimensions, as well as the most commonly used by the assessors in the study detailed in Chapter 3 (Table 3.2), were preselected. The final terms were selected according to inclusion and exclusion criteria. The inclusion criteria were the most commonly used terms for each dimension and which dimension they were representing. The exclusion criteria involved the negativity of the terms (e.g. uncomfortable, unwell, unhappy) and the least used terms in the FCP study. These negative terms were excluded based on previous fixed behavioural expressions studies as their assessment when using the VAS is more challenging (Phythian *et al.*, 2013; Grosso *et al.*, 2016).

The final list of laboratory mouse behavioural expression fixed-terms included 13 terms from the preselected list (Table 4.1) and seven terms taken from other QBA fixed terms studies (positively engaged, confident, sociable, fearful, uncertain, bored, frustrated). These additional terms were considered relevant by the four experts involved in the selection process but were not present in the initial FCP list of terms. A definition was provided for each term after general discussion and agreement between the four experts (Table 4.2).

Dimension	Positive correlations	Negative correlations
Dimension 1	Active (8) Alert (5) **Inquisitive (4)** **Energetic (2)** Interested (2) Responsive (1) Explorative (1) **Determined (1)**	**In pain (9)** **Lethargic (3)** Unwell (3) Uncomfortable (3) **Depressed (2)** Inactive (2) Reluctant (1)
Dimension 2	**Calm (11)** Relaxed (9) Curious (3) **Content (3)** **Comfortable (2)** **Playful (1)**	**Agitated (8)** Stressed (8) **Anxious (3)** Nervous (3) **Tense (3)** Frightened (2) Scared (2) Concerned (1) Exhausted (1)

Table 4.1. Preliminary list of fixed behavioural expression list extracted from the Free Choice Profiling study. The number in brackets corresponds to the number of assessors who used each term for describing the expression of the laboratory mice in the videos in the FCP. The terms in bold (n=13) were included in the final list of behavioural expressions for laboratory mice.

Term	Definition
Inquisitive	The mouse appears curious and interested in others and in exploring the environment. Willing to investigate.
Positively engaged	The mouse is carrying out activities in a focused, directed and constructive manner. The mouse appears not to be distracted by others or the environment.
Energetic	The mouse is carrying out an activity with a lot of energy and vigour, in a lively and excited manner.
Determined	The mouse is showing an active and rapid reaction to something or someone. It appears to be focused on accomplishing a specific goal or task.
Confident	The mouse is displaying assertiveness, behaving assertively with other animals and its environment in a self-assured manner.
Calm	The mouse appears peaceful and without worry. The mouse behaves in a relaxed and untroubled manner.
Content	The mouse appears happy and satisfied. Expressing happiness, with all its physiological, environmental and psychological needs met.
Comfortable	The mouse appears physically satisfied with the cage environment and looks relaxed and free from discomfort.
Playful	The mouse is engaging in lively movements purely to frolic or for fun, expressing pleasure, happiness and amusement.
In Pain	The mouse is suffering from physical discomfort leading the mouse to be reluctant to move, or to move with abnormal gait, or showing a tense,

	hunched or uncomfortable posture. The mouse looks like it is hurting or suffering and is in discomfort.
Lethargic	The mouse appears fatigued and sluggish. It has a lack of vigour and energy, showing low amounts of movement and any movement is slow and ponderous.
Depressed	The mouse appears unhappy and without hope. It is apathetic, despondent and unresponsive showing little or no response or reaction to anyone or its environment. It appears isolated.
Anxious	The mouse is uneasy, cautious and nervous
Agitated	The mouse appears to be irritable and highly reactive. An excess of physical and cognitive activity is present because of anxiety.
Sociable	The mouse actively interacts with others. It is willing to interact with others showing affiliative actions (e.g. grooming, resting in groups, sniffing)
Fearful	The mouse appears afraid or scared. It seems to be avoiding contact with others and the environment, looks to be hiding, looking for a way out or trying to escape.
Tense	The mouse looks worried and emotionally tense. Its posture might evidence physical tension.
Uncertain	The mouse appears to be insecure; its physical movement is cautious. The slowly showing alertness and insecurity. Avoidance reactions are showing with all stimuli
Bored	The mouse appears uninterested in its environment and/or cage mates. The way it moves around and orients itself appears to be unfocused and aimless, without much energy, never staying long with a particular activity or aspect of the environment.
Frustrated	The mouse appears unfulfilled with its environment and/or cage mates. It looks stressed and uneasy showing repetitive and fast movements.

Table 4.2. List of the fixed behavioural expressions for assessing the emotional states of laboratory mice using Qualitative Behavioural Assessment. This list was created based on previous Free Choice Profiling study (Chapter 3) and previous QBA fixed list studies.

The assessors were informed that this study aimed to review if QBA could be used to interpret mouse body language, based on videos showing a variety of laboratory mouse expressions in different environmental conditions. They were not given any information about the two different methods of handling. They were told that the human hand and the human hand & tube were present in the cage to stimulate a greater variety of behavioural expressions in the mice.

4.2.3. Mice handling experiments

Animal husbandry

Thirty-two female C3H mice (Charles River Laboratories, UK), aged approximately seven weeks were used in this study. Mice were housed in pairs in IVC NexGen

Cages (194mm x 181mm x 398mm, Allentown). The IVC cages had air delivery at the cage level (20 air changes per hour). Food (Special Diet Services, RM3E) and water were provided *ad-libitum*. Cages contained sawdust bedding, nesting material (Enviro-Dri® and Sizzle-Pet) and a clear Perspex home cage tube (50mm diameter, 150mm length). Mice were maintained on a 12:12 hour light/dark cycle (lights on at 07:00) and cages were cleaned out once per week. Mice were individually identifiable with a tail mark (see below). The mice had a two-week acclimation period before the start of the study. During this time, mice were handled using the tail by a trained technician once per week during cage cleaning.

After this acclimation period, the cages were randomly allocated (via random number generator) into two groups (8 cages per group). From this point onwards, animals in group 1 were always handled via the tail and animals in group 2 were always handled via a tube (Hurst and West, 2010). The study period began with a ten-day baseline period where no additional procedures were carried out other than the weekly cage cleaning. At the end of this period (day 10), the baseline behavioural filming was carried out. All handling was carried out according to their assigned method (tail or tube) including during the routine daily checks.

Interaction tests video recording

After the baseline period, mice had nine consecutive days of handling sessions (from day 11 to 20). Interaction tests and behavioural filming were carried out on days 12, 16 and 20 and an elevated plus-maze test was carried out on day 21 (Figure 4.1). The handling sessions, the interaction test, and the elevated plus-maze test were filmed using an HD video camera (Canon HD Legria HFM506).

Behavioural filming was carried out in the home cage which was moved to a quite procedure room. The nesting material and tube were removed, cameras were located at 30 cm in front of the home cage, and mice were filmed for 6 minutes on each occasion. The elevated plus-maze was performed on day 21. Mice were placed in a quiet room, and then they were placed in the elevated plus-maze individually using their specific handling method. They were filmed for 5 minutes and the time spent in each arm (open and closed) was recorded. The handling sessions were carried out daily from day 11 to day 20. The cages were placed in a quiet room, on a bench. The researcher removed the cage lid and took one mouse at the time from the cage using the specified handling method (tail and tube). For the tail method, the mouse was gripped by the base of the tail and picked up; then it was placed on the researcher's hand and held there for 60 seconds. After that period the mouse was put back in the home cage. For the tube handling the homecage tube was taken, and the mouse was gently directed into the tube. The tube was picked up and held in front of the researcher for 60 seconds, after that time the tube was put back into the home cage. All sixteen cages were used for the interaction test performed twice (before and after) the handling session on days 12, 16 and 20. The interaction test was performed in a quiet room with the home cage on a bench. This test was carried out directly in the home cage with the nesting material, and the home tube removed. The researcher placed one hand (for tail handling group), or the hand and the tube (for tube handling group) in the middle of the cage on top of the bedding for 3 minutes and video footage were recorded.

The interaction test videos showed a wide range of situations and behavioural expressions in laboratory mice, showing the willingness of the mouse to interact with the handler (Hurst and West, 2010). Each of these videos showed two mice (i.e. paired house) interacting with the handling method. Assessors who participated in the study were informed about the characteristics of the videos (two mice, home cage, the interaction between handler and mice) and that the assessment needed to be made taking into account the cage as a whole including both animals as well as the environment and the context where animals were. They were informed that individuals could influence each other in many different ways, when they move, communicate and interact together. Assessors were also informed that influence of cage mates and mouse behaviour could have its expressive quality (e.g. calm, anxious or tense) so the study aimed to assess the expressive qualities at the cage (i.e. pair) level.

Video footage

A total of 96 videos were obtained from the interaction test. From these 96 videos, the interaction session filmed after the handing session (48) were included in the preliminary selection as they were more likely to show any change of behaviour repertoire as they were filmed straight after the handling session. Ten 60-second videos from tail and tube handling were selected for this study. The inclusion criteria involved the use of the quadrants which described the behavioural expressions in QBA (see Chapter 3). Briefly, the dimensions which represent animals expressivity are recorded in dimension models which are represented in four quadrants;

good/calm (quadrant one), good/energetic (quadrant two), bad/calm (quadrant three), and bad/energetic (quadrant four). All 48, 3-minute, videos were scanned for inclusion, and the selection was made taking into account the 4 quadrants and observing where the videos show the most variable mouse behaviour expressivity. The duration of the videos used was determined to take into account previous fixed behavioural expressions studies (Phythian *et al.*, 2013; Grosso *et al.*, 2016).

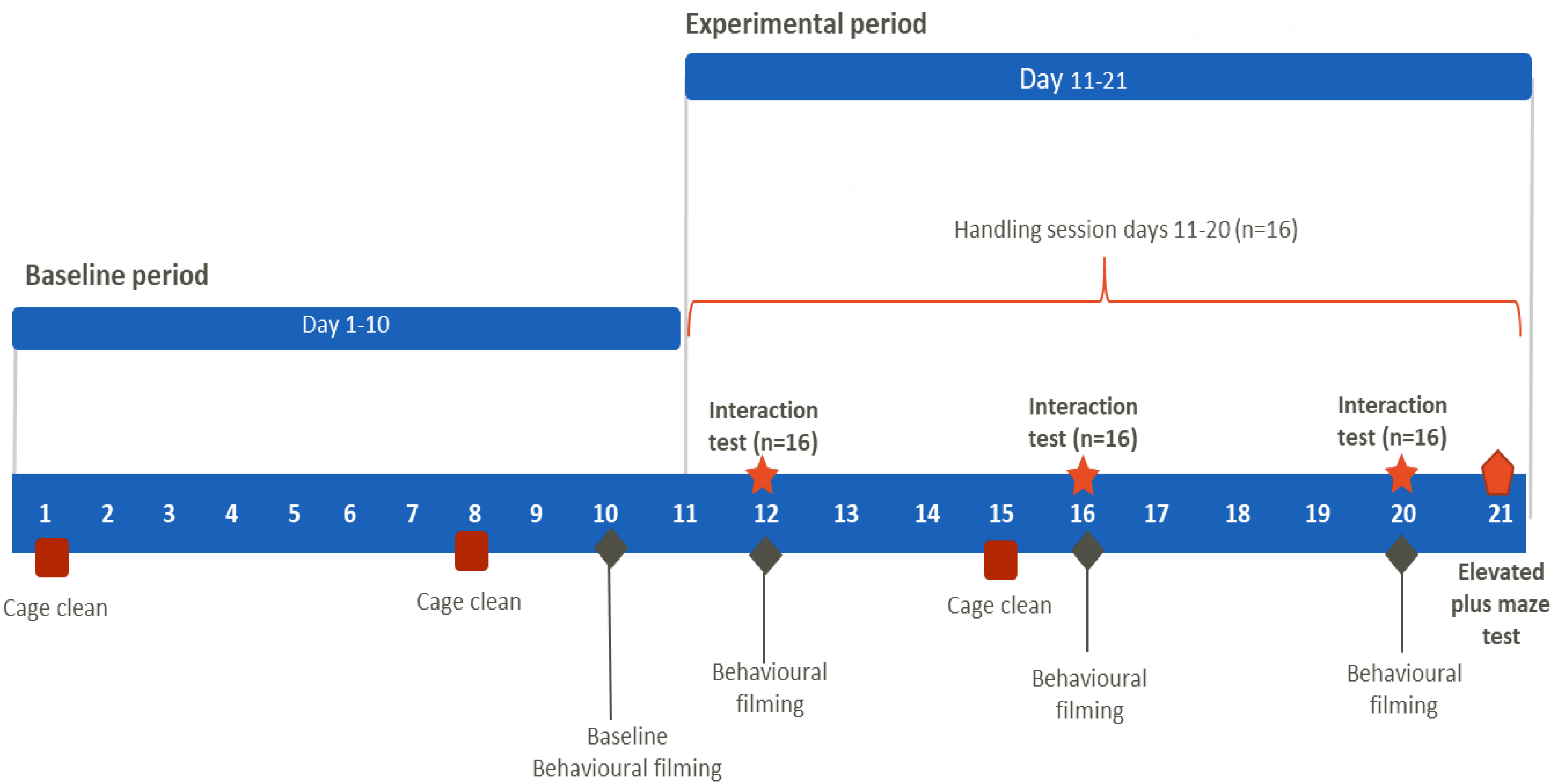


Figure 4.1. Experimental timeline for the QBA fixed terms study. The figure shows the dates of the behavioural filming, interaction tests, cage cleaning and the elevated plus-maze test. The two week acclimation period is not shown. The study was divided into two periods, baseline and experimental. On the last day of baseline period, the baseline behavioural filming was carried out (day 10). During the experimental period, the interaction tests and behavioural filming (days 12, 16, 20), and elevated plus maze tests (day 21) were carried out.

4.2.3. Qualitative Behavioural study

Assessors

A total of 19 assessors took part in this study. They were animal behaviour and welfare MSc students at Newcastle University who voluntarily agreed to participate in the study; they had general knowledge about animal behaviour and welfare in other species such as cows, dogs and pigs. The assessors did not have any previous knowledge, training or experience with laboratory mice, so did not have any prior knowledge regarding the relevance of the different handling techniques. Assessors underwent training in QBA and the fixed behavioural expressions list. The training included an introduction to QBA as an emotional state measurement tool in other species and examples of previous work. The meaning of each of the fixed behavioural expressions (Table 4.2) and their use for assessing laboratory mouse emotional states was discussed. They also were introduced to the Visual Analogue Scale (VAS) as a measurement tool. All 19 assessors took part in the first session with 9 out of the 19 assessors participating in session two.

Assessment procedures

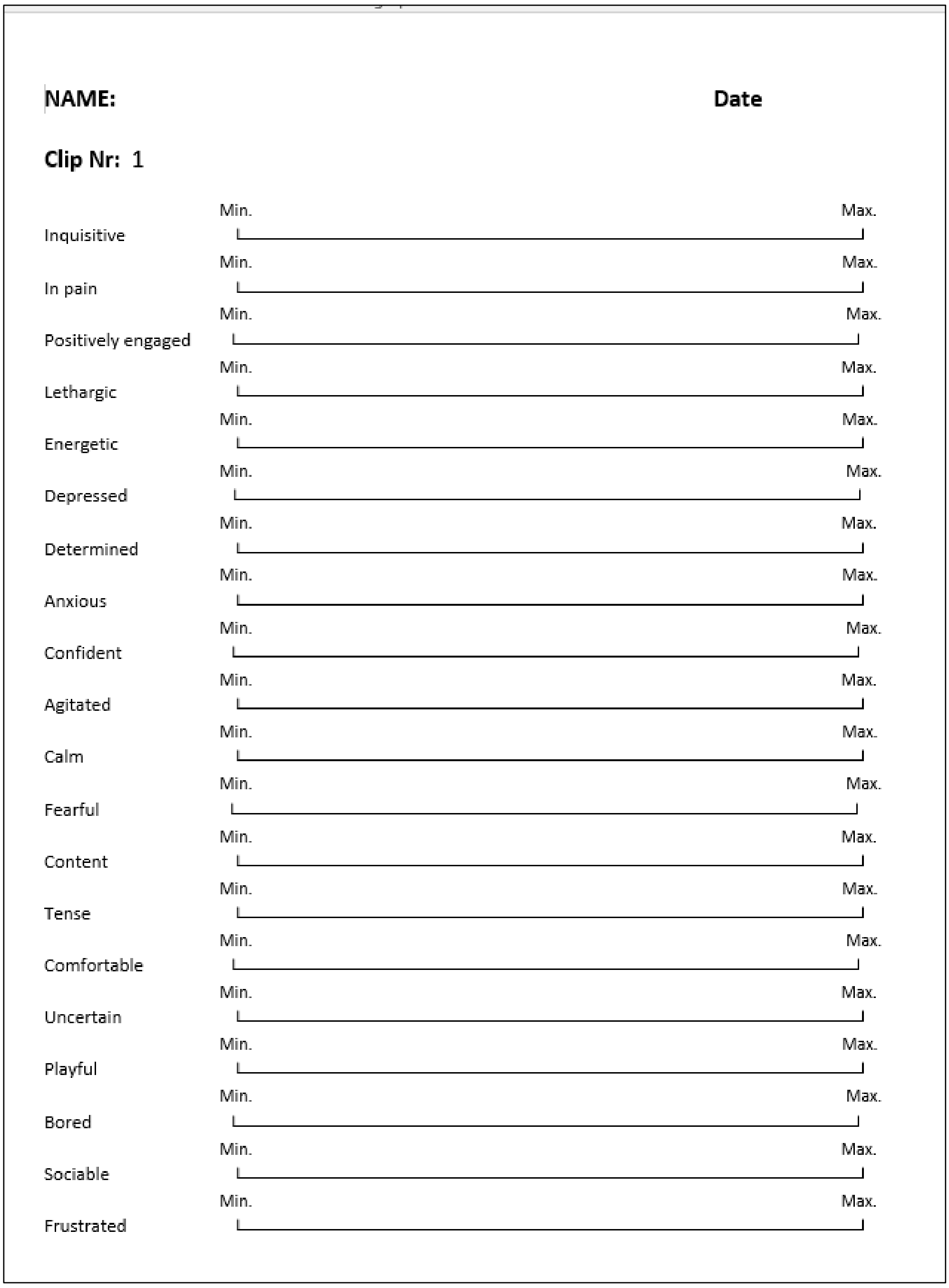
Each assessor underwent either one (n=19) or two sessions (n=9) scoring the same videos with a one-week gap between the sessions. The videos were in a different, random order in each session, and the assessors were not informed about the change of the order of the videos or that they were assessing the same group of videos. Each session included two parts; training, followed by the participants scoring the test videos.

During part one (training), each assessor received three documents:

1) Instructions for QBA
2) Twenty score sheets (one per video) which included the list of 20 fixed terms paired with a Visual Analogue Scale (Figure 4.2).
3) A list of the 20 fixed terms and their definition (Table 4.2).

Additionally, a training video explaining how to use the fixed terms was shown to participants. The assessors discussed the meaning of the terms in this session. They were informed that they must rate all 20 mice videos using one score sheet per video which contained the 20 fixed behavioural expressions with the VAS (Figure 4.2). The

VAS consisted of a continuous line with no categories or divisions so that it can be interpreted intuitively for the assessor (Brscic *et al.*, 2009). The left point of the scale (0mm point) is the minimum meaning, i.e. that the characteristic (fixed term) is absent and the right point (125mm from left) the maximum meaning for the characteristic, i.e. it is entirely present and dominant (e.g. the mouse cannot be more anxious). Assessors were instructed to avoid any discussion with each other about the terms, definitions and videos during the second part of the assessment.



NAME: **Date**

Clip Nr: 1

Inquisitive	Min.	Max.
In pain	Min.	Max.
Positively engaged	Min.	Max.
Lethargic	Min.	Max.
Energetic	Min.	Max.
Depressed	Min.	Max.
Determined	Min.	Max.
Anxious	Min.	Max.
Confident	Min.	Max.
Agitated	Min.	Max.
Calm	Min.	Max.
Fearful	Min.	Max.
Content	Min.	Max.
Tense	Min.	Max.
Comfortable	Min.	Max.
Uncertain	Min.	Max.
Playful	Min.	Max.
Bored	Min.	Max.
Sociable	Min.	Max.
Frustrated	Min.	Max.

Figure 4.2. QBA assessment sheet used for the fixed terms study. Each assessor received one sheet per mouse video. Twenty fixed terms were included. The observers received information and training about how to use these scales for scoring each term before watching the videos.

Following training (part 2), assessors watched each of the 20 videos with a one-minute white screen between each video to enable them to score each video in the VAS using the fixed behavioural expressions (part 2).

One week later, 9 assessors repeated the scoring following the same format as described above.

The VAS scores generated by each assessor, in both sessions were then calculated by measuring the distance in millimetres from the minimum point (0mm on the left) of the scale to the mark made by each assessor.

4.2.4. Data analysis

All data was collated in Microsoft Office Excel® sheet before being transferred to statistics software for analysis. The data were tested for normality and homogeneity of variance using a Shapiro-Wilk test for residuals in SPSS (v23, SPSS Inc, Chicago,USA). Where assumptions for parametric analysis were not met, non-parametric statistical methods were used.

Principal Component Analysis (PCA; correlation matrix, no rotation) was used for the analysis of session one with 19 assessors and another PCA was used for the analysis of session one and session two together for 9 assessors who repeated the scoring process (i.e. undertook 2 sessions). PCA is a mathematic algorithm which reduces the dimensionality of the original variables (fixed terms) while maintaining the variation in the data set (Jolliffe, 2011). The primary goal of a PCA analysis is to identify patterns in data, in this case, patterns in the scoring of each fixed-term scored by each assessor for each of the 20 videos. PCA aims to detect the correlation between variables and so reduce the dimensionality by identifying dimensions (principal components) that explain the variability between video clips. The use of a correlation matrix represents the analysis. The main two dimensions with eigenvalues (numbers that define the variance of the data) greater than 1.0 were used to produce the two-dimensional word chart, and the eigenvectors (or loadings) quantify the weight that each fixed term has on the two principal axes (Rencher, 2003; Temple *et al.*, 2011).

After the PCA analysis, Kendall's correlation of coefficient (*W*) test was performed to assess to what extent (inter-observer agreement) the assessors agreed in the ranking of the videos for each dimension (PC1, PC2) for session one (between 19 observers) and the intra-observer reliability within the two sessions (within 9 observers). Intra-observer reliability was calculated using Intra Class correlation (ICC). ICC is a reliability index used for correlation and agreement between measurements (Koo and Li, 2016). In this study, the two-way mixed-effects model was selected for the ICC, because all the assessors were rating all the videos (Landers, 2015). An absolute agreement was selected because the study aimed to assess the reliability of the QBA scores for multiple raters and we were interested in the absolute agreement of scores. ICC estimates and their 95% confidence intervals were calculated using SPSS, IBM Corp. SPSS (v23, SPSS Inc., Chicago, USA). In addition to the ICC, Spearman correlation was used for assessing intra-observer reliability between session 1 and session 2 for each assessor using SPSS, IBM Corp. SPSS (v23, SPSS Inc,Chicago,USA).

In order to assess if there was a treatment (handling method: tail or tube) effect on the QBA scores in session 1, a two-way, independent ANOVA was performed (depended variable: PC1 and PC2 scores, fixed factors: assessor and handling treatment). In session 2, the assessment of the effect of the sessions, treatment and assessors was carried out using a mixed model ANOVA for the nine assessors who completed both sessions, (within-subjects variable: session, between-subjects factors: handling treatment). The data analysis was carried out using Minitab® 17.1.0 (2013 Minitab Inc. license for windows) for the PCA analysis, R© (version 1.0.153© 2009-2017 RStudio, Inc.) and SPSS statistical analysis package version 24 (IBM Corp. SPSS (v23, SPSS Inc., Chicago, USA) for ICC, Kendall's correlation of coefficient (*W*), the two-way independent ANOVA and the mixed ANOVA.

4.3. Results

4.3.1. Session one (n=19): inter-observer reliability

Principal Component Analysis scores produced two main dimensions (components) which explained the 44% (PC1) and 14% (PC2) of the variance of the behavioural expressions exhibited in the videos. The distribution of the 20 fixed behavioural expressions into the two main dimensions is shown in Figure 4.3.

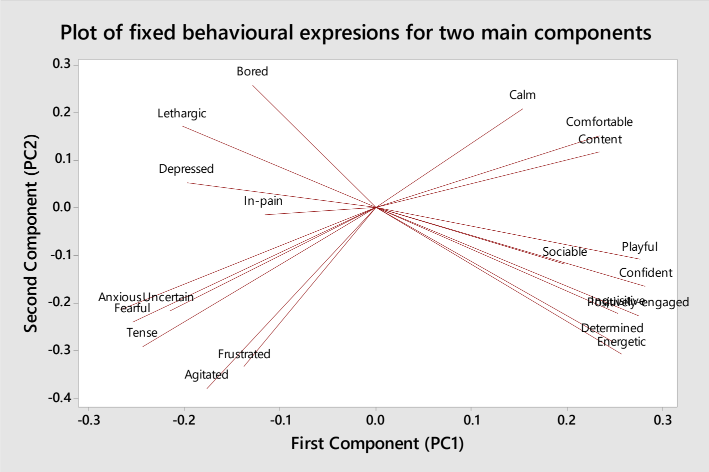


Figure 4.3. Two-dimensional loading plot of the 20 QBA fixed behavioural expressions. The figure shows the PC1 – 'First Component' (Valence) and PC2 – Second Component (Energy levels). Low scores in PC1 indicate negative valence, while high scores indicate positive valence. Low scores in PC2 indicate high energy levels, while high scores indicate low energy levels.

PC1 appears to relate to valence (from positive to negative) ranging from "confident/playful" (high scores) to "anxious/fearful" (low scores) whereas PC2 appears to relate to energy levels (from low to high) ranging from "bored/calm" (high scores) to "agitated/frustrated" (low scores). The PCA scores for all twenty fixed behavioural expressions are shown in Table 4.3.

Fixed behavioural expressions	PC1	PC2
Inquisitive	0.253	-0.224
In-pain	-0.117	-0.018
Positively-engaged	0.275	-0.229
Lethargic	-0.202	0.169
Energetic	0.257	-0.310
Depressed	-0.197	0.036
Determined	0.248	-0.284
Anxious	-0.269	-0.215
Confident	0.282	-0.168
Agitated	-0.178	-0.380
Calm	0.154	0.205
Fearful	-0.255	-0.239
Content	0.234	0.114
Tense	-0.244	-0.292
Comfortable	0.235	0.148
Uncertain	-0.217	-0.217
Playful	0.277	-0.111
Bored	-0.129	0.256
Sociable	0.198	-0.123
Frustrated	-0.138	-0.335

Table 4.3. PCA scores of the 20 fixed behavioural expressions (QBA) in the main two dimensions (PC1 - PC2) representing valence and energy levels.

Overall the level of agreement for PC1 was good, Kendall (*W)* value (0.67), p<0.001. The level of agreement for PC2 was low and not significant, Kendall *W* (0.184, NS) (Table 4.4).

	% of variation explained	Kendall's W (N=19, df = 18)
PC 1	44%	0.678
PC 2	14%	0.184

Table 4.4. PCA outcomes and inter-observer reliability for the QBA fixed terms study.

The distribution of the videos, identified by handling method is shown in Figure 4.4. Mice that were handled by tail were mostly allocated in the negative valence half (PC1 left quadrant) of the first dimension whereas mice handled by tube were allocated mostly in the positive valence half of the same principal component.

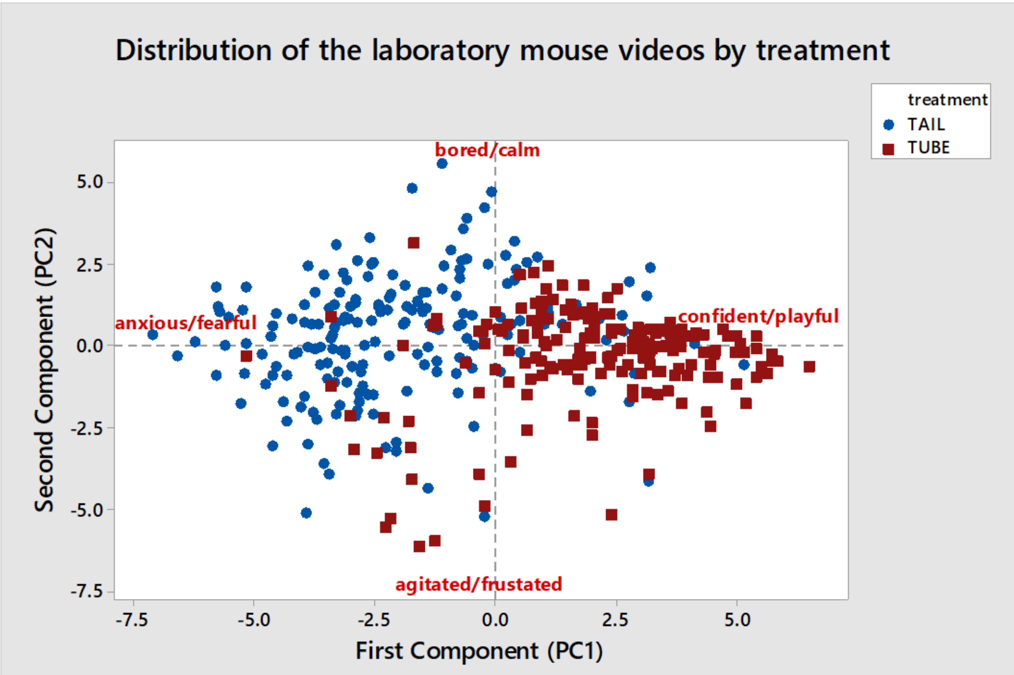


Figure 4.4. Score plot of the distribution of mouse videos (19x20 = 380), identified handling method. Each dot corresponds to one video scored by one assessor.

A two-way ANOVA was conducted that examined the effect of the assessors and the handling method on the PC1 scores (Figure 4.5). There was a significant effect of the interaction between the assessors and handling method on the PC1 scores, *F* (18, 342) = 2.832, p<0.001. The partial Eta Squared Value was calculated for assessing the percentage of variance give for each component of the interaction. The Eta squared value shows 58% of the variance in PC1 that can be attributed to the handling method in contrast with 34% for the assessor. The estimated marginal means of tail and tube handling method for each assessor shows that higher marginal means are associated with tube handling method compared to the tail handling method for each assessor (Figure 4.6).

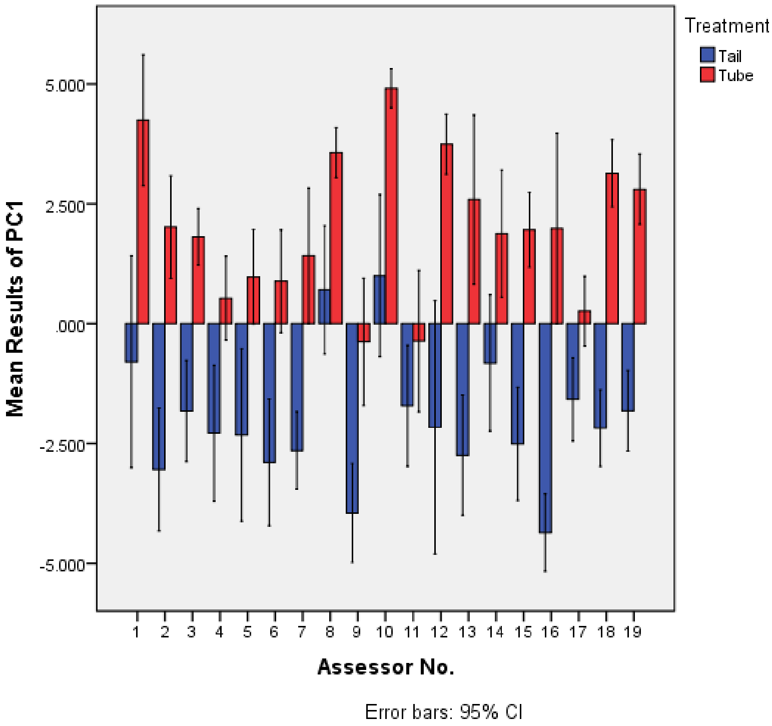


Figure 4.5. Histogram of the handling method and assessor interaction in the PC1 scores of Qualitative Behavioural Assessment. The 19 assessors are located on the x-axis, and the y-axis represents the mean rank results of PC1 values produced by the assessors. Error bars represented the Standard Error at a 95% confidence interval. The handling method is represented by the colour blue for tail and colour red for tube.

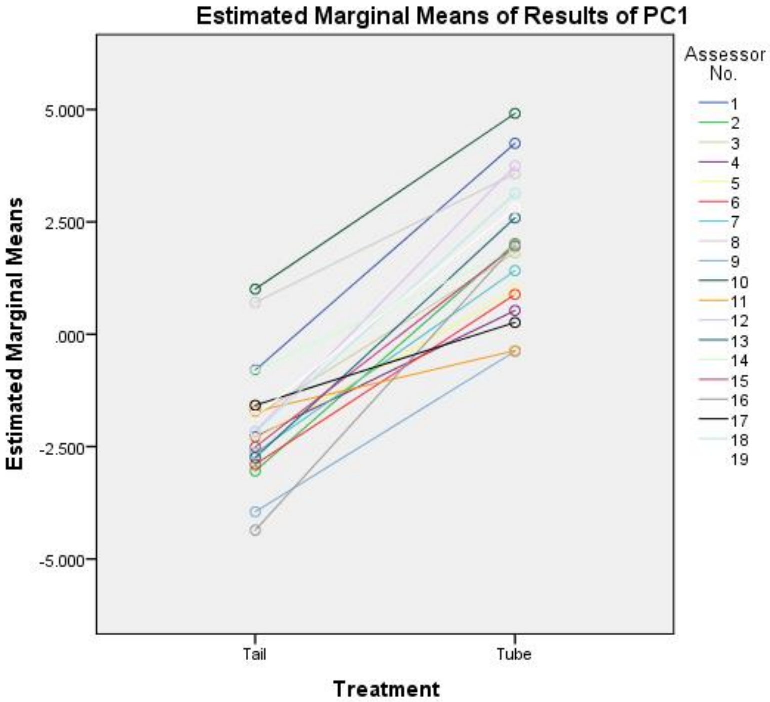


Figure 4.6. Estimated marginal means of the results of PC1 showing the effect of handling method and assessor interaction. The handling method (tail and tube) is located on the x-axis, and the estimated marginal means of the PCA scores are in the y-axes.

There was a significant effect for assessor [$F\ (18, 342) = 9.93$; $p<0.001$] indicating that assessors differed in their mean scores on PC1 (Figure 4.7A). The PC1 scores given by the 19 assessors showed a range of mean scores from negative to positive GPA scores values. A significant effect of handling method was also found in the PC1 scores [F (1, 342) = 478.65; $p< 0.001$], indicating that the handling treatments differed in their mean PC1 score (Figure 4.7B).

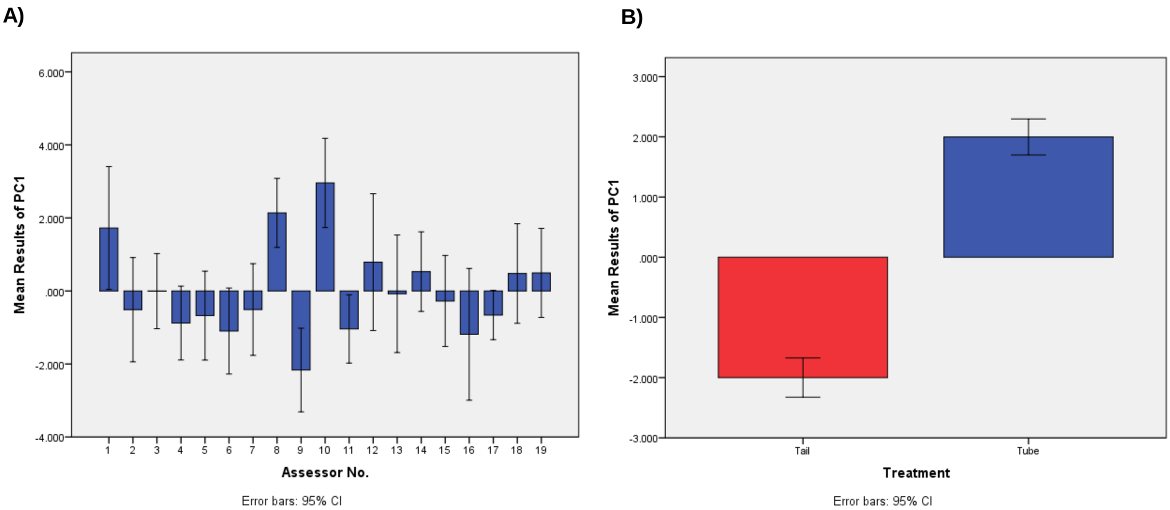


Figure 4.7. Histograms of the effect of assessor (A) and treatment [handling method] (B). A shows the assessor on the x-axis and their effect on the mean PC1 scores in the y-axis. Error bars represent the Standard Error at a 95% confidence interval. B shows the effect of the handling method (tail and tube) on the mean PC1 scores

4.3.2. Session one and two (n=9): Intra-observer reliability

Principal Component Analysis scores produced two main components (dimensions) which explained 44% (PC1) and 14% (PC2) of the variance of the fixed behavioural expressions the mice exhibited in the videos. The 20 fixed behavioural expressions were distributed in the plot shown in Figure 4.8. PC1 appears to relate to valence (from positive to negative) ranging from “confident/positively engaged” (high scores) to “fearful/anxious” (low scores) whereas PC2 appears to relate to energy levels (from low to high) ranging from “bored/calm” (high scores) to “tense/agitated” (low scores). The PCA loadings of dimension one and two for all twenty of the fixed behavioural expressions are shown in Table 4.5.

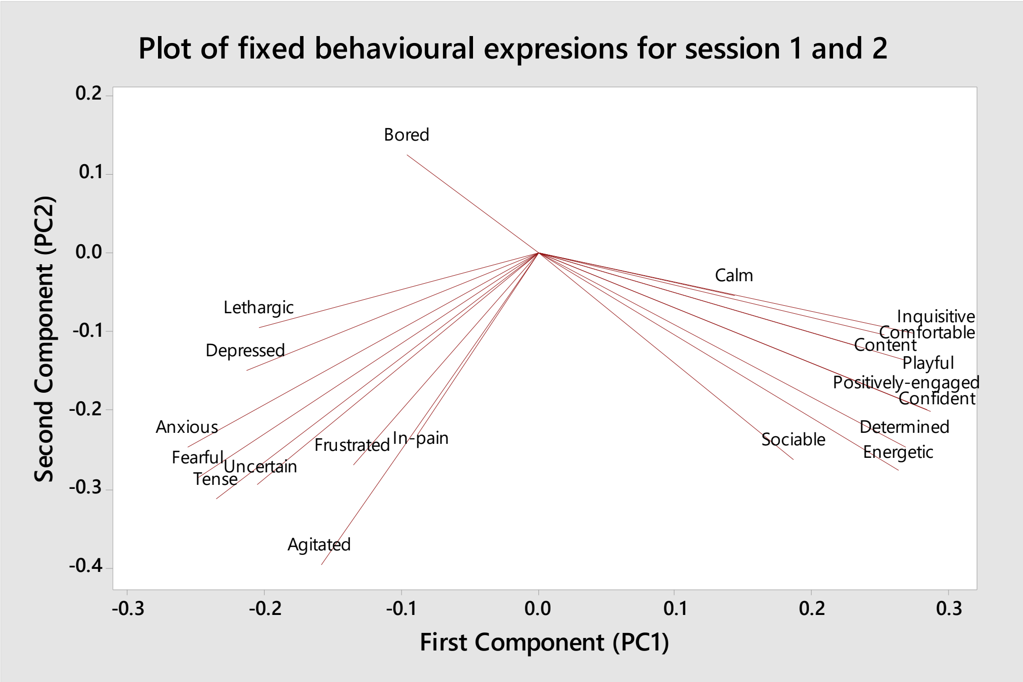


Figure 4.8. Two-dimensional loading plot for the PCA analysis of repeat QBA scores provided by 9 observers in session 1 and session 2.

Fixed Terms	PC1	PC2
Inquisitive	0.274	-0.101
In-pain	-0.088	-0.232
Positively-engaged	0.285	-0.199
Lethargic	-0.204	-0.096
Energetic	0.264	-0.277
Depressed	-0.214	-0.15
Determined	0.268	-0.247
Anxious	-0.256	-0.247
Confident	0.288	-0.201
Agitated	-0.159	-0.396
Calm	0.144	-0.053
Fearful	-0.249	-0.285
Content	0.232	-0.118
Tense	-0.236	-0.313
Comfortable	0.253	-0.105
Uncertain	-0.206	-0.295
Playful	0.267	-0.136
Bored	-0.096	0.126
Sociable	0.187	-0.263
Frustrated	-0.135	-0.269

Table 4.5. PCA loadings for each fixed behavioural expressions for dimension one and two. These loadings determine the location of each term in the dimensional plot shown in Figure 4.8.

The inter-observer reliability between the nine assessors who took part in both sessions is shown in Table 4.6. The table shows ‘good’ agreement in their scores for PC1 for both sessions and no agreement for PC2.

	Kendall’s W Session 1 (n=9, df=8)	Kendall’s W session 2 (n=9, df = 8)
PC 1	0.70	0.75
PC 2	0.12	0.15

Table 4.6. Inter-observer reliability for QBA in two different sessions.

The distribution of the videos according to session number (one and two) is shown in Figure 4.9. The 20 videos were located along with the score plot following the same distribution in both sessions.

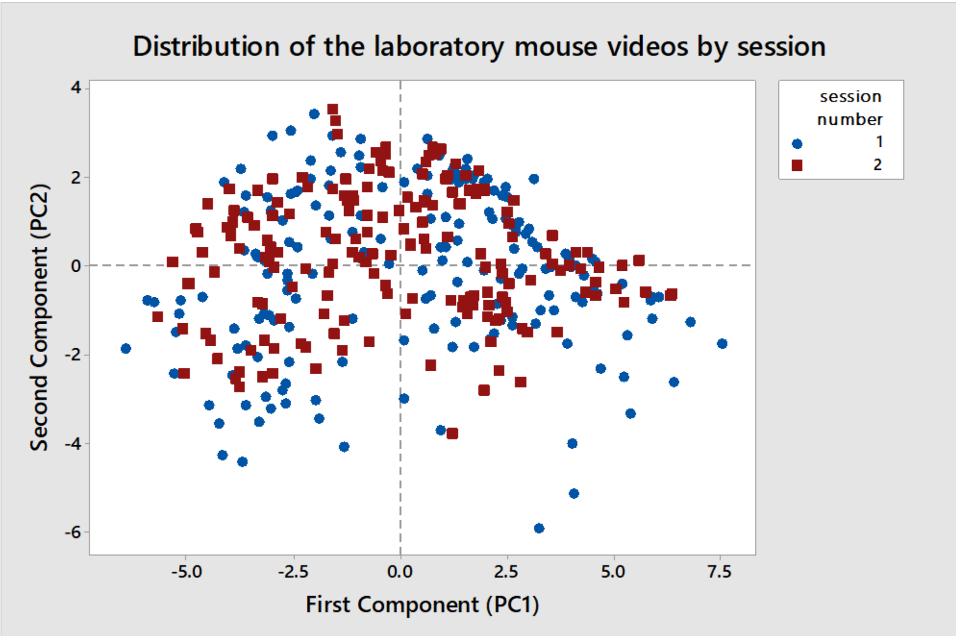


Figure 4.9. Score plot of the distribution of mouse videos in the two sessions (9x20x2 = 360). Each dot corresponds to the location of each video for each assessor (9) in two sessions.

An Intraclass correlation (ICC) was performed for assessing intra-observer reliability between the 9 observers scores for the 20 video clips in session 1 and session 2 (Table 4.7). ICC values less than 0.5 are indicative of poor reliability, values between 0.5 and 0.75 indicated moderate reliability, values between 0.75 and 0.9 indicate good reliability, and values greater than 0.90 indicate excellent reliability (Portney and Watkins, 1993; Landers, 2015; Koo and Li, 2016). Based on a mean rating (k = 9), absolute-agreement, 2-way mixed effect model is shown in Table 4.7. An excellent degree of reliability was found between the assessor’s scores of videos for PC1 in the two sessions. The average measure ICC for PC1 was 0.964 with a 95%

confidence interval from 0.933 to 0.983. The reliability between assessor scores for PC2 was low (0.147).

Principal Component	Intraclass Correlation	95% confidence interval		F test With True Value 0		
PC1		Lower Bound	Upper Bound	df1	df2	Sig
	0.964	0.933	0.983	19	323	0
PC2		Lower Bound	Upper Bound	df1	df2	Sig
	0.147	-0.054	0.147	19	323	0.093

Table 4.7. Intraclass Correlation Coefficient (ICC) using a two-way mixed model assessing absolute agreement for the PC1 and PC2 scores of the nine assessors in session 1 and session 2.

Spearman correlations were calculated for assessing the intra-observer reliability between session 1 and session 2 for each assessor (Table 4.8). The majority of assessors had a very high correlation between session 1 and session 2 in the PC1 scores (<0.5: n=1/9 0.5 to 0, 7: n=3/9, <0.7: n=5/9).

Spearman's Correlations		
Correlations n=20		Value
PC1.1a - PC1.1b	Correlation Coefficient	**.767****
	Sig. (2-tailed)	0
PC1.2a - PC1.2b	Correlation Coefficient	**.848****
	Sig. (2-tailed)	0
PC1.3a - PC1.3b	Correlation Coefficient	**.802****
	Sig. (2-tailed)	0
PC1.4a - PC1.4b	Correlation Coefficient	**.653****
	Sig. (2-tailed)	0.002
PC1.5a - PC1.5b	Correlation Coefficient	**.517***
	Sig. (2-tailed)	0.02
PC1.6a - PC1.6b	Correlation Coefficient	**.734****
	Sig. (2-tailed)	0
PC1.7a - PC1.7b	Correlation Coefficient	**.687****
	Sig. (2-tailed)	0.001
PC1.8a - PC1.8b	Correlation Coefficient	**0.388**
	Sig. (2-tailed)	0.091
PC1.9a - PC1.9b	Correlation Coefficient	**.829****
	Sig. (2-tailed)	0

**. Correlation is significant at the 0.01 level (2-tailed).
*. Correlation is significant at the 0.05 level (2-tailed).

Table 4.8. Spearman's correlation coefficient of each assessor's PC1 scores between session one and session two. A number identifies the assessor from 1 to 9, and the sessions were label "a" session 1 and "b" session 2.

To assess the effect of the assessor, handling method, session and their interactions on the PC1 scores, a mixed-design ANOVA was performed.

The effect of session x assessor interaction

There was also a significant interaction between session and assessor ANOVA F (8,168) = 3.617, p<0.001. (n=9) (Figure 4.10). Although the PC1 scores were affected by whether the scores were from session one or two, the pattern of the scoring was different for different assessors. For some assessors, the PCA 1 scores decreased from session 1 to session 2 (n=4/9), whereas for others the scores showed no or only a small change (n=3/9), and for some assessors, the scores increased from session 1 to session 2 (n=2/9).

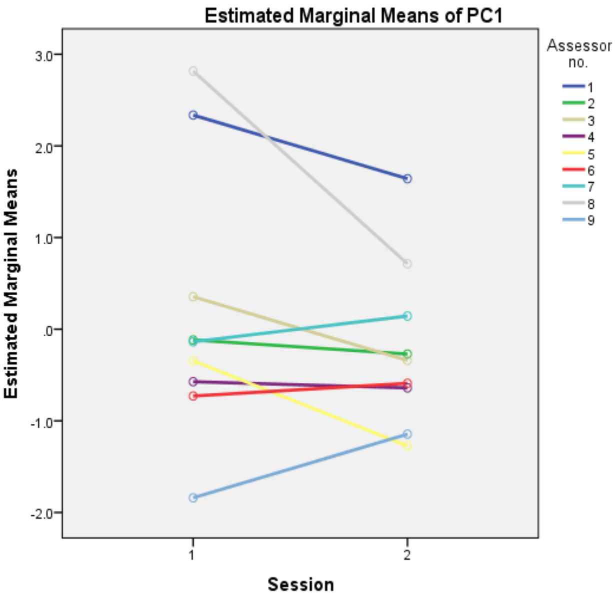


Figure 4.10. Effect of session-assessor interaction. The figure shows session number on the x-axis and the estimated marginal mean of the PCA scores on the y-axis, with assessor (1-9) being identified using different colours.

The effect of session x handling method interaction

There was no significant interaction between session one and two for both treatment (tail and tube) [*F* (1,168) = 0.156, *p*=0.693]. Therefore, the mean PC1 score for each of the two sessions did not differ significantly for the tail and tube handling groups (Figure 4.11). However, even if the interaction between handling method and session is not statistically significant, on closer inspection, figure 4.11 indicates that there is a common pattern in individual assessors scores as the mean scores for PC1 one tend to be higher for tube handling comparing with tunnel handling. In PC2 however, the scores are similar for both sessions with similar mean scores for tail and tube. Assessors were able to score the laboratory mouse videos similarly in both sessions.

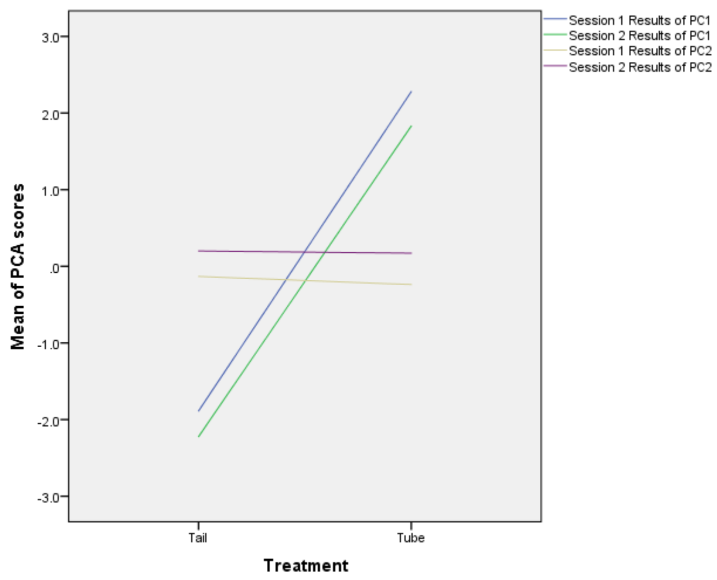


Figure 4.11. Effect of the session x handling method interaction. The x-axis represents the treatment group, and the y-axis represents the mean of the PCA scores of the laboratory mouse videos.

The effect of session x assessor x handling method interaction

There was no significant three-way interaction between session, assessor and handling method [F= (8,162) = 0.935, *p*=0.489] (Figure 4.12.). The sessions and assessor interaction are the same for the two handling groups (tail and tube).

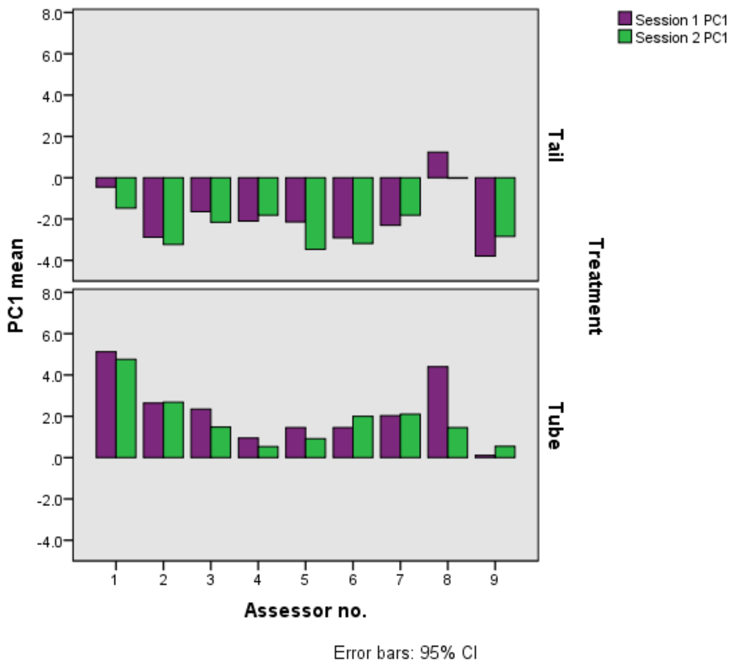


Figure 4.12. Three-way interaction of session (1 and 2), assessor and treatment (handling method) on the PC1 scores. The assessor is located on the x-axis from 1 to 9, and the PCA mean scores of the videos are located in the y-axis. Tube handling group is located at the bottom of the figure and tail handled group at the top of the figure.

The effect of session

There was a significant main effect of 'session' on the PC1 scores, [*F*, (1,162) = 7.301, *p*<0.001] (Figure 4.13). Irrespective of handling method and assessor, the PCA scores in Session 1 were higher than those in Session 2 (Figure 4.13).

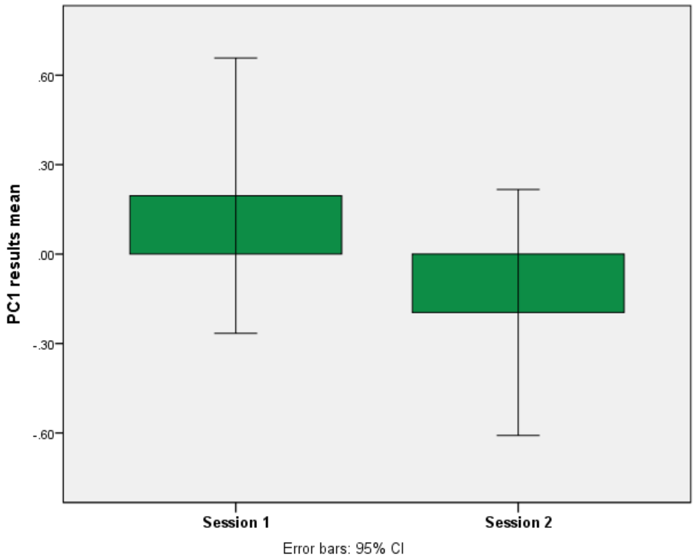


Figure 4.13. Effect of the session in the PC1 scores. Session 1 and two are located on the x-axis and the mean of the results of PC1 scores on the y-axis. Errors bars represent the standard error at a 95% confidence interval.

4.4. Discussion

This study aimed to assess the validity of QBA as an indicator of emotional expressivity in laboratory mice by using a fixed behavioural expressions list of mouse and to assess its inter- and intra-observer reliability. The results showed a high level of inter-observer reliability on the first main dimension (PC1: confident/positively engaged to fearful/anxious), but not on the second dimension (PC2: bored/calm to tense/agitated). High levels of intra-observer reliability were also shown in PC1 when a subset of assessors repeated the scoring process following a 1-week break. One of the first challenges in the assessment of QBA in laboratory mice is their "natural environment", the home cage. This home cage has a high level of restriction when compared with other species (e.g. farm animals) because of the small home cage and the constraint of resources (e.g. environmental enrichment) that we can provide in the cage which makes the assessment of their expressivity more challenging as they might not have many opportunities for showing their whole behavioural repertoire.

Nevertheless, other QBA studies have successfully assessed the fixed behavioural expressions of animals in very confined environments such as goats in a pen (Grosso *et al.*, 2016) or cattle in a transport trailer (Stockman *et al.*, 2011). However, QBA was able to identify successfully, through PCA, two main components of mouse expressive demeanour using 20 videos assessed by 19 individuals. The first dimension (PC1) seemed to represent 'Valence' and went from 'confident/playful' to 'anxious/fearful'. The second dimension (PC2) seems to represent 'energy levels' and went from 'bored/calm' to' agitated/frustrated'. Additionally, in the intra-observer reliability agreement, assessors used very similar terms for both dimensions, with descriptors going from confident/positively engaged to fearful/anxious in PC1 and from bored/calm to agitate/tense for PC2. The ability of QBA to identify the main components of mouse behavioural expressivity in the home cage shows that it may offer a valuable tool for assessing emotional states which represent the psychological states as one of the main components of laboratory mouse welfare.

The two main components of mouse fixed behavioural expressions defined by the PC scores (valence and energy levels) are very similar to the components that are widely recognised by cognitive and human research as the dynamics of emotionality (Bradley and Lang, 1994; Feldman, 1995; Barrett, 1998; Anders *et al.*, 2004; Scherer, 2005). These two components are also described in animal emotions research (Panksepp, 2004; Posner *et al.*, 2005; Mendl *et al.*, 2010) which suggests a structure for demonstrating a functional perspective of animal emotional states (Mendl *et al.*, 2010). Valence refers to "the positive and negative character of emotion or some of its aspects; subjective experiences and expressive behaviours (facial, bodily, verbal) (Joffily and Coricelli, 2013). In this study, valence is used as one of the principal components of the subjective experience of emotions as a validated tool for assessing fixed behavioural expressions in animals (Russell, 1980; Russell, 2003).

The videos were showing mice that had been handled by tail were allocated in the negative valence dimension whereas the videos showing mice that had been handled by tube were allocated in the positive valence dimension (Figure 4.4). Videos of mice that were handled by tail displayed more anxious/fearful demeanour than those handled using a tube. These results are consistent with previous studies about tail handling as an anxiety-inducing husbandry practice (Hurst and West, 2010; Gouveia and Hurst, 2013). A low agreement between assessors for the dimension related to energy has also been found in another study (Grosso *et al.*, 2016) and Chapter 3 when using FCP profiling. A potential explanation for this lack of agreement might be related to the fact that PC2 is describing the dimension of energy levels. The differentiation of energy levels in mice is a challenging task, as mice are very active and in constant motion, and it could be very difficult for an assessor without any previous knowledge (i.e. those who participated in this study) to differentiate between the energy levels and correlate them with fixed behavioural expressions (Staats, 1966; Heston, 1967; Boursot *et al.*, 1993). The PC1 scores, related to valence were more easily differentiated, and there was an agreement between assessors for this dimension. Valence (PC1) is a dimension commonly used for the assessment of emotional states in different animal welfare protocols such as on farms (Quality, 2009; Welfare Quality®, 2009c) and donkeys (Dai *et al.*, 2016) as this is the dimension that constitutes the welfare scale (Wemelsfelder *et al.*, 2009; Andreasen *et al.*, 2013; Phythian *et al.*, 2016). These results, however, need to be interpreted with caution, as the final videos that were selected for the study were only a small proportion (60 seconds) of the filmed videos (three-minute videos). The duration of the videos might affect the results, as 60-seconds might not be enough time for the assessors to determine the animal's behavioural expressions. The final selection of videos was made taking into account the variety of mouse expressivity, and they might not be representative of the two treatments used in the study. These videos were a reflection of the researcher's opinion of the fixed behavioural expressions that the mice should show between the treatments, but they are not necessarily the reflection of changes in the laboratory mouse behavioural expression because of the treatment itself. However, the fact that assessors with no experience of laboratory mice were able to agree in the difference between the videos shows that QBA could be a promising indicator for the assessment of emotional states in this study.

The Intraclass Correlation Coefficient used to assess intra-observer reliability shows that the 9 participants reliably scored the videos in the same way for PC1 between the two sessions that were one-week apart. This result of intra-observer results supports the reliability of QBA for assessing animal behavioural expressions in mice, which has been demonstrated in other species such as pigs (Wemelsfelder *et al.*, 2001; Wemelsfelder *et al.*, 2012), dairy cows (Rousing and Wemelsfelder, 2006), and buffaloes (Napolitano *et al.*, 2012) . In PC2 however, assessors were not able to either agree with each other or replicate their own scores in the assessment of PC2 of the videos. This result provides further evidence that the assessment of laboratory mouse energy levels is very challenging at least for people who do not have any experience in working with laboratory mice. However, further research is needed involving participants who are familiar with laboratory mouse behaviour to provide a better assessment of both PC1 and PC2 dimensions of QBA.

Scores for PC1 were significantly influenced by the interaction between the handling method and assessor. This effect was broken down using the partial Eta Squared Value showing a 58% of the variance in PC1 was attributed to the handling method in contrast with 34% for the assessor (Brown, 2008). One possible explanation for the effect of the assessors on the PC1 scores might be a variation in the use of the Visual Analogue Scale (VAS) for quantifying the 20 fixed emotional expressions list. This variation is a well-known effect present in the VAS scores linked with individuality in the use of the scales (Torrance *et al.*, 2001). For example, one assessor may decide to use only the second half of the scale whereas others may decide to use the whole scale. While they can still have the same judgement when assessing a term (e.g. the alertness on the mouse observed cannot be more evident), they could cross the line in the VAS in different places taking into account their individual approach for using the scale. That variation when using the VAS is consistent with other QBA studies (Fleming *et al.*, 2015; Grosso *et al.*, 2016) and also has been discussed (e.g. end-aversion bias in health measurement VAS) in human research (Torrance *et al.*, 2001; E Marian and P Shimamura, 2013). In this study, that variation was seen as all assessors agreed about the allocation of the videos and the handling treatment into different quadrants (positive and negative valence), but an assessor effect was present. The importance of the calibration of the VAS was evidenced in a prior QBA study on donkeys (Minero *et al.* (2016). In this study, Minero *et al.* (2016) calibrated the use of VAS by training participants to use the whole scale in the VAS from the minimum to the maximum points. This ‘calibration’ resulted in the removal of the assessor effect of the study (Minero *et al.*, 2016). This study shows how the assessor effect can be reduced or controlled by appropriate training about the meaning of the terms combined with the use of the VAS.

Additionally, in this experiment, all the assessors scored the same videos using the same fixed terms so providing some potential control of the assessor effect (McClelland and Judd, 1993). The PC1 scores show the ability of QBA to distinguish between the two different types of handling methods. Previous research has indicated that tail handling is more stressful for the animals than tube handling (Hurst and West, 2010; Gouveia and Hurst, 2013; Gouveia and Hurst, 2017) and this ability to use QBA to differentiate between stressful situations has also been shown in other species such as cattle (Stockman *et al.*, 2011), horses (Minero *et al.*, 2009; Hintze *et al.*, 2017), lambs (Serrapica *et al.*, 2017), sheep (Wickham *et al.*, 2012) and pigs (Rutherford *et al.*, 2012). It can be argued, that the two handling methods used might have influenced the outcomes of this study as the videos used (interaction test) had differences that were visible for the assessors (either a hand [tail handling] or hand and tube [tube handling]). This interaction test assesses the willingness of the mouse to interact with the handler and involves the handler to placing the handling method into the home cage to assess animal's behaviour (Hurst and West, 2010). This test incorporates the presence of different behaviours (i.e. sniffing, paw contact, climbing, inside handling device) to assess the willingness of the mouse to interact with the handling method (Hurst and West, 2010; Gouveia and Hurst, 2013). However, the handling method (either tail or tube) as a husbandry technique in laboratory mice is a practice that is not well known to people who do not have experience working with these animals. This new technique for handling laboratory mice (tube handling) has only recently been studied and is not yet widely in use (Hurst and West, 2010; Gouveia and Hurst, 2013; Gouveia and Hurst, 2017). Therefore, the lack of knowledge of the assessors about this handling method and the good representation of anxiety levels showed for the interaction test in other studies (Hurst *et al.*, 1999; Hurst and West, 2010) in addition to the results of the study suggest that the effect of the assessors and handling method interactions might be related to the use of the VAS.

The effect of the session on the PC1 scores also needs to be considered. The analysis showed that session, as well as the interaction of session and assessor, affected the PC1 scores. The scores for PC1 were different for session one and two. Some assessors showed considerable variation between the sessions by either decreasing or increasing regarding the direction of the variability, and this variance was more evident for some assessors than from others. Assessor 8 for example, had a considerable variability from session 1 to 2 affecting the overall PC1 results between the two sessions (Figure 4.10). This variation, however, could be explained regarding the usage of the Visual Analogue Scales. The calculation of the mixed model ANOVA for the effect of the session in the GPA scores for dimension 1 uses the mean difference between the groups. These means can be different from session one and session two as the assessors can use the VAS in different ways. Assessor 8 for example, could provide a score for the 20 videos using the first part of the scale allocating the videos in a specific order. However, in the second session, the same observer could allocate the videos in the same order but using a different section (e.g. upper part) of the scale, giving a different mean for the same videos but still allocated them in the same order. However, when comparing this with the effect of the interaction between session and handling method, there was not a significant effect. Assessors were able to identify the laboratory mouse videos by handling method and place them into different quadrants. This consistency of the differentiation between the videos selected for the two handling methods, was also seen in the session one with all the 19 assessors thus the possible explanation can be related to the ability of assessors to differentiate between behavioural expressions indicating positive or negative valence as has been shown in other species (Minero *et al.*, 2009; Napolitano *et al.*, 2012; Rutherford *et al.*, 2012; Phythian *et al.*, 2013; Minero *et al.*, 2016). The three-way interaction (session, assessor and treatment) effect was not significant for the PC1 scores. The session and assessor interaction were the same irrespective of the handling method used. This result is very relevant as it shows that assessors were able to differentiate between the handling methods even when they use the VAS in different ways, and they can repeat this differentiation of the videos observed in two different sessions.

A potential limitation of the study could be related to the lack of agreement between the assessors for PC2. This dimension does not define the welfare outcomes (final status about the welfare of the animals) of the assessment, but it usually provides useful information for welfare recommendations. In other species, the recommendations taken from PC2 are usually described regarding energy levels (i.e. lethargic/frustrated) of the animals and determine if they need to have enrichment that increases their level of activity or husbandry measures that reduce that level of activity (Grosso *et al.*, 2016). The interpretation and usefulness of the PC2 in the laboratory mouse have not yet determined although it can be used to determine the changes that need to be made for improving welfare. For example, if the outcomes of the QBA assessment suggest that the welfare is compromised (e.g. PC1 result describe the animal as lethargic/bored), the improvement of the environmental enrichment can be made for increasing activity levels (e.g. place a wheel for increasing exercise/movement). However, more research is needed in the possible application of the results of PC2 in laboratory mice.

The effect of the assessor on the scores can be considered as another limitation. The way that assessors use the VAS for the assessment can produce differences between the scores. As the study of Minero *et al.* (2016) showed, the training of the assessors about the meaning of the fixed terms as well as a calibration between assessors on the use of VAS is crucial for eliminating this effect. Finally, the expertise of the assessor in the species being assessed can be another factor that influences the final results. The majority of the studies using QBA have involved assessors who had higher levels of expertise as the knowledge of the animal's normal behaviour is considered an essential factor in this type of assessment (Rousing and Wemelsfelder, 2006; Wemelsfelder *et al.*, 2009). In this study, the assessors used were inexperienced with regard to laboratory mice. They were selected, as those with a high level of knowledge in laboratory mice would have understood the context of the tube and hand presence in the videos. This lack of knowledge about species-specific natural behaviour has not had a considerable effect on the results in the way it has been demonstrated in other studies where the assessor experience had been compared showing no relevant effect in the final results (Napolitano *et al.*, 2012; Rutherford *et al.*, 2012).

This is the first time that QBA has been used as an indicator for assessing laboratory mouse emotional states. The results of this study showed that QBA appears to be able to reliably differentiate between animal positive and negative emotional states (associated with different handling methods known to cause different levels of anxiety) which further validates its use as an animal welfare indicator of emotional health (Minero *et al.*, 2009; Fleming *et al.*, 2015). This study involves the use of a VAS for scoring fixed terms to describe animal's emotional states. Recently, there has been increasing evidence that mice should be handled by tube, rather than the traditional method of by the tail, as it decreased anxiety levels (Hurst and West, 2010; Gouveia and Hurst, 2013). QBA was a useful tool for the assessment of laboratory mouse welfare states using interaction test videos. However, in order to be incorporated into a laboratory mouse welfare protocol, its practicality needs to be assessed. The next steps towards the construction of the laboratory mouse welfare assessment protocol involved the construction of the final protocol including QBA as an indicator of psychological states, followed by a reliability and practicability assessment.

Chapter 5. Laboratory mouse welfare protocol: construction, practicability and reliability assessment

5.1. Introduction

An animal welfare protocol aims to identify when welfare is compromised, to allow the necessary arrangements to be made to resolve welfare issues and also to highlight and promote favourable welfare conditions (Mellor and Reid, 1994; Hubrecht, 2014). In order to construct an appropriate welfare protocol for laboratory mice, an appropriate definition of mouse welfare needs to be selected to determine the most suitable indicators (Hawkins *et al.*, 2011). Laboratory mouse welfare is defined in this thesis as the animal's ability to preserve their biological functioning (i.e. physiological parameters), their psychological functioning (i.e. mental well-being) (Boissy *et al.*, 2007a; Fraser, 2009; Murphy *et al.*, 2014) and the ability to meet their evolutionary needs (i.e. environment) (Dawkins, 1990; Broom, 1996; Fraser and Broom, 1997; Dawkins, 1998; Hubrecht, 2014). This definition implies that to adequately assess the welfare of animals, these three main components should be measured and that this needs to be done at the individual level (Hawkins *et al.*, 2011). However, as this definition considers the environmental conditions which animals are exposed too, such indicators provide the opportunity to assess the welfare of multiple animals at the same time (i.e. at the group level). Environmental indicators, for example, provide information about the general environment within the facility, room and/or the cage, which will be relevant to all the animals housed in those areas. In a laboratory mouse facility, we are often dealing with a large number of animals whose welfare needs to be assessed. Therefore, from a practical perspective, the assessment of every single individual is likely to be unrealistic, and so we need to assess the welfare at a group level. It is important to underline that the welfare of all animals that have undergone an experimental procedure need to be assessed as it is more likely to be compromised. A critical question, when assessing welfare at the individual level is how many animals need to be assessed as observing every animal is often not feasible (Dawkins, 1998). Therefore, many of the existing welfare assessment protocols, sample a proportion of the animals (e.g. 5% of goats on the farm from the Welfare Quality Program) from the colony and generalise the results of across this colony, for example in cattle (Welfare Quality®, 2009c), pigs (Welfare Quality®, 2009a) and poultry (Welfare Quality®, 2009b). The need to observe a subset of the colony and to generalise is an issue faced by audit assessments which tend to be carried out within a discrete time frame. However, this is not an issue for the daily assessments / welfare checks carried out by technical staff that are needed to meet regulatory requirements (Hawkins *et al.*, 2011). Such assessments are carried out at individual level, with the welfare of all the individuals within the facility being assessed.

Determining the most appropriate indicators to score (i.e. those that are reliable, valid and practical) is essential for the development of a protocol for the assessment of mouse welfare. A preliminary selection of indicators was determined in Chapter 2 using a Delphi consultation to select indicators based on expert opinion (Whaytt *et al.*, 2003). The indicators that comprise a protocol are dependent on the type of welfare assessment being carried out (i.e. daily assessment of welfare vs audit of welfare at a specific time-point) as this will influence the time and resources required to conduct the assessment (Hawkins *et al.*, 2011). The everyday protocol is designed for the assessment of welfare on a daily basis and is usually carried out by technical staff within a research facility as part of the daily individual checks required by legislation regulation (e.g. A(SP)A, EU Directive 2010/63/EU in the UK). The audit welfare assessment is likely to be an assessment made less frequently by someone other than animal care staff, aiming to assess welfare at a specific time (e.g. as part of a larger welfare audit within a facility). Moreover, this audit protocol is likely to be related to the assessment carried out by representatives of regulatory organisations (e.g. Home officer inspectors). There are limited data about the variability observed between these two types of assessment, regarding how they are different in terms of the indicators used and if there are any differences or similarities regarding the reliability and practicability of the indicators.

The assessment of psychological state (e.g. emotional state) is considered to be the most critical aspects of welfare (Boissy *et al.*, 2007a). Nevertheless, it is one of the most challenging components to assess due to the difficulty of assessing emotions in animals (Désiré *et al.*, 2002). Even in humans, who can self-report, assessing emotional states can be a challenging task (Boissy *et al.*, 2007a). In this study Qualitative Behavioural Assessment (QBA) has been included as means of assessing psychological state following preliminary validation in earlier studies (Chapters 3 and 4).

Using data from both the Delphi consultation process (Chapter 2) and QBA (Chapters 3 and 4), a final list of indicators has been defined. The objective of this study was to determine the effectiveness of these validated indicators within an overall protocol, in terms of their practicability and reliability. For indicators to be practicable, their measurement should incur a minimal cost, require minimal resources to record (e.g. specialised equipment) and be quick to score (Blank, 2004). The latter is critical for the practicability of a protocol, and it must represent a balance between the assessment taking the minimum time without affecting its reliability (Velarde and Dalmau, 2012). The development of a scoring system which includes predetermined scales (e.g. 0 normal, 1 slightly abnormal, 2 very abnormal) for each indicator and is logically ordered can also improve the practicability and must be considered when constructing the protocol (Hawkins *et al.*, 2011). A protocol with a structured scoring system and a logical order can have a critical advantage in the assessment of welfare as the records are consistent and produce more objective and useful results (Hawkins *et al.*, 2011). The resources needed, the time spent, and the ease of scoring each indicator was assessed using a Likert scale. Information in the literature about the assessment of practicability of animal welfare indicators is minimal as most of the information derives from previous studies related to the opinion of the assessors about the practicality of a specific measurement (Whaytt *et al.*, 2003). The construction of the practicability analysis was made taking into account the assessment techniques used in education were the assessment of the practicality of different learning methodologies is assessed using Likert scales. These scales consider the time spend in the assessments, the resources needed and the ease of scoring (Bushman and Schnitker, 1995).

The inter-observer reliability of a protocol is defined as the ability to produce consistent information when used by different people under identical circumstances (i.e. assessing the same animal in the same state on more than one occasion) (Blank, 2004). Intra-observer reliability was also assessed which is related to the ability of the protocol to produce consistent information when it is used by the same person at different times (Litwin and Fink, 1995). The assessment of both forms of reliability for a laboratory mouse welfare indicator is important as it is needed to demonstrate the robustness and objectivity of the protocol when applied by a wide range of observers to assess the welfare of laboratory mice in different conditions.

This study had 2 aims

1. To assess the practicability and reliability of an everyday welfare assessment and audit welfare protocols.

2. To assess the validity of the everyday protocol for identifying welfare problems in experimental animals.

5.2. Ethical statement for the assessment

This study was conducted following the registration for unlicensed work (AWERB Project ID: 449) at Newcastle University and in accordance with the Animals (Scientific Procedures) Act 1986 and the EU Directive (2010/63/EU).

The protocols consisted of observation and minimal restraint only, and there were no adverse effects reported during this study. The experimental animals assessed in this study were part of an unrelated project conducted at Newcastle University under the Animals (Scientific Procedures) Act 1986 (Project ID: A00102A74), and in accordance with the EU Directive (2010/63/EU) and was approved as licensed work by the AWERB and in accordance with the EU Directive (2010/63/EU), and ASPA (1986).

5.3. Indicators for the audit and everyday welfare protocols

5.3.1. Methods

Indicators selection criteria

Indicators selected for inclusion in every day and audit welfare protocols involved the definition of laboratory mouse welfare, the principle that the indicator was represented, and practicability and reliability assessment obtained from the Delphi consultation. The process of selection of the indicator was divided into three phases: [1] selection of an initial list of indicators, [2] pilot assessment of the individual indicators regarding practicability and reliability, [3] assessment of the overall protocol regarding practicability, reliability (audit and everyday protocols) and validity in experimental animals (every-day protocol) (Figure 5.1).

Phase 1: The initial list of indicators used for the Delphi consultation was obtained from a literature review and the scoping meeting before the formal consultation process (Chapter 2). At the end of the Delphi consultation, information about the validity, practicability and reliability of each indicator were used to create a rank order from 1 to 59. Based on the definition of animal welfare proposed in the first chapter, the indicators selected for assessing laboratory mouse welfare represent one of the components of the definition (i.e. biological, psychological or environmental). These 59 indicators were then divided into the three groups based on the aspect of the animal welfare definition to which they align: [1] biological functioning (physical health), [2] psychological state and [3] the ability to meet their evolutionary needs considering their environment (environmental indicators). Environment for laboratory animals refers to the housing and care (resources) that we provided for fulfilling the evolutionary needs that they have (Olsson and Dahlborn, 2002). This division was used to ensure that the final list of the indicators within the protocols covered all the three animal welfare components.

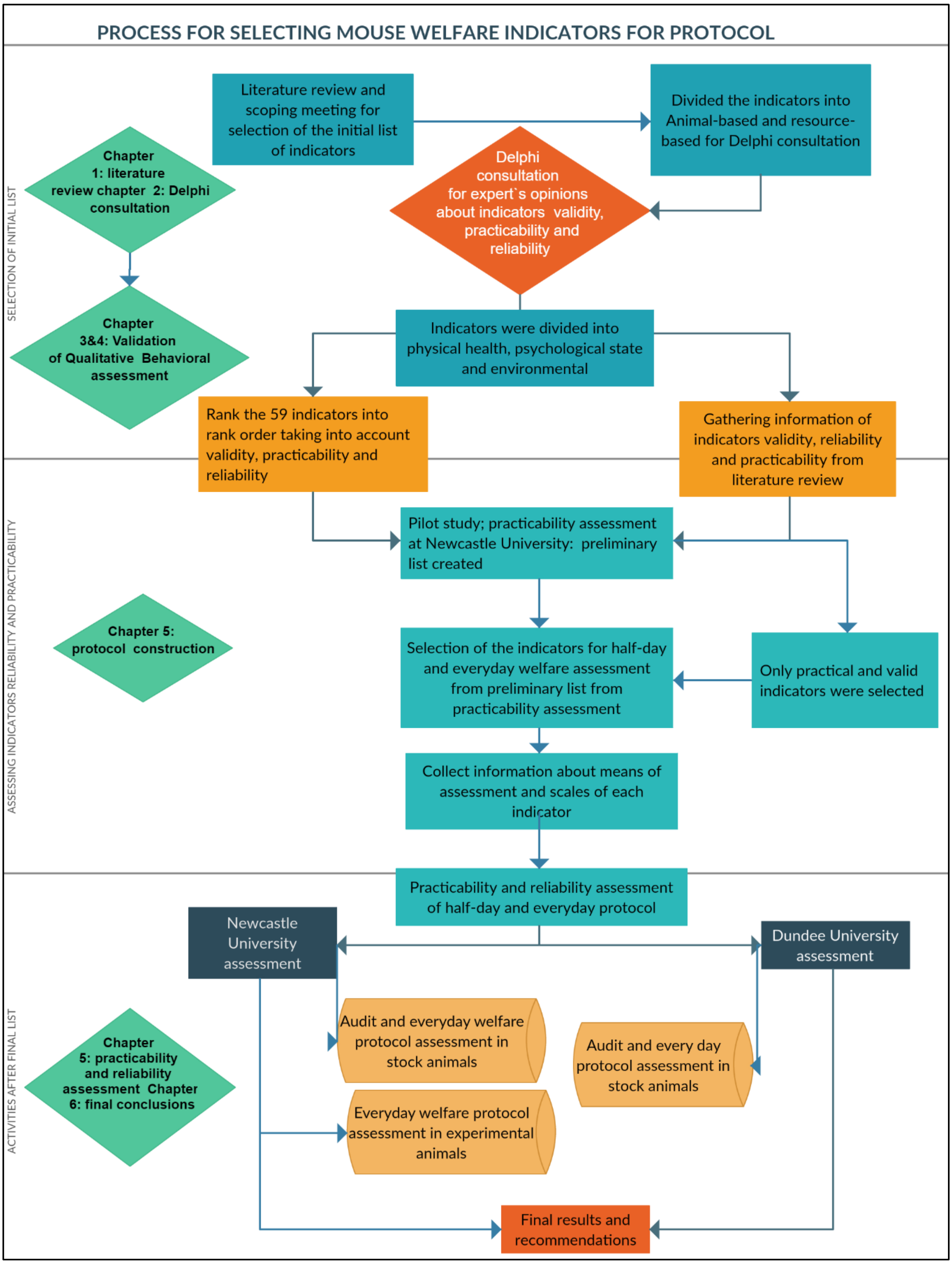


Figure 5.1. The selection process of the laboratory mouse welfare indicators to be used in an audit and an everyday welfare assessment. The process involved three main phases: initial selection of indicators, reliability and practicability assessment and protocol reliability and practicability assessment.

A literature review was conducted to gain information about the practicability and reliability of each proposed indicator, along with an appropriate means of scoring. This process reduced the number of indicators from 59 to 49. The information about these changes to the final list and the justification for the changes are shown in Table 5.1.

Previous indicators	Justification of change
Cage dimension and house density	Merged as they are dependent on one another
Wound/marks and bite wound/marks	Merged as they are assessed simultaneously
Exhibition of normal behaviour and exhibition of abnormal behaviour	Merged as they are assessed simultaneously
Positive interaction with conspecifics and negative interaction with conspecifics	Merged as they are assessed simultaneously
Approaching hand in the cage	Removed as lack of information about practicability and reliability
Rearing	Removed as is part of the exhibition of normal behaviour
Floor-type	Removed as the type of floor in the cage is standardised in the laboratory facilities
Light source	Removed as light source is standardised in the laboratory facilities
Substrate type	Removed as the type of substrate in the cage is usually standardised in laboratory facilities
Faecal cortisol levels	Removed as it lacks practicality
Ultrasonic noise levels	Removed as requires specialised equipment for the measurement
Water consumption and food consumption	Removed as requires previous information about baseline consumption

Table 5.1.Table of justification of the merging and removal of indicators which reduces the list of indicators from the 59 to 49.

Following the study carried out in Chapter 4, two additional, relatively new indicators that were not included during the Delphi consultation, were added to the list as they were considered relevant for welfare assessment: [1] handling method (Hurst and West, 2010; Gouveia and Hurst, 2013) and QBA (Wemelsfelder, 2007; Brscic *et al.*, 2009).

Therefore, a list of 51 indicators was used in a pilot study to assess the practicability of the indicators included in the protocols. Indicators were organised into three logically ordered categories (Table 5.2): facility level, cage level and individual level to increase the practicality of the assessment (Phase 2). Facility level indicators are related to the assessment of the facility or animal rooms (e.g. room temperature, humidity, staff training). Cage level indicators are related to the assessment of welfare at the cage level (e.g. cage dimensions, presence/absence of nesting material). Individual-level indicators assess welfare at the individual mouse level, (e.g. body condition score, coat condition, hunched posture).

Indicators divided into subcategories	
Facility/room:	
Disease rate	Light/dark cycle
Empathetic attitude of staff towards animals	Method of feeding
Frequency of cleaning cages	Mortality rate
Frequency of feeding	Nesting material
Frequency of physical examination	Presence/absence of bedding material
Frequency of veterinary procedures	Staff training
Handling method	Temperature
Humidity	Type of cleaning procedure
Infanticide rate	ventilation
Intensity of lighting	
Cage:	
Bloodstains in the cage	Type of bedding material
Cage odour (ammonia levels)	Usage of cage space
Complexity of the cage	Usage of nesting material
Excessive urine and faeces in the cage	Audible signs of aggression
Food type	Alertness
House density and cage dimensions	Pups outside the nest
Qualitative Behavioural assessment (QBA)	
Individual:	
Barbering (hair removal)	Easy of handling
Body Condition Score	Swollen abdomen
Body temperature	Gait
Coat condition	Hunched position
Dehydration	Ocular/nasal discharge
Tail position	Urination/defecation during handling
Respiratory pattern	Weight loss
Skin colour	Wounds/marks (including bite wounds)
Positive and negative interaction with conspecifics	Mouse grimace scale (facial expressions of pain)
Exhibition of normal and abnormal behaviour	

Table 5.2. List of indicators assessed for practicability in the pilot study.

Data analysis of the practicability assessment

In this study, the scoring for practicability was conducted in three ways: [1] time spent, [2] resources needed, and [3] easiness of scoring. Each of these measures was given 4 possible scores which can be seen in Table 5.3. The percentage practicability score was calculated using the mean score of the 3 categories divided by the maximum possible score. The mean of the scores for each category was selected as it is considered to offer the best representation of the distribution of the data (Black, 1999). This percentage score also aligns with the percentage of reliability information obtained when asked the experts in the Delphi consultation.

Category		Score
Time spent	Very time consuming	1
	Time-consuming	2
	Quick	3
	Very quick	4
Resources needed	Very specific equipment needed	1
	Specific equipment needed	2
	Equipment needed but it is in the facility	3
	No equipment needed	4
Ease of scoring	Very difficult to score	1
	Difficult to score	2
	Easy to score	3
	Very easy to score	4

Table 5.3. Table of practicability assessment score information. Each indicator was assessed using a four-point scoring system. Table based on (Bushman and Schnitker, 1995; Blank, 2004).

Practicability assessment

The total number of animals to be assessed was calculated based on the total number of stock animals present at the facility on the day of the assessment. The assessment was made at Newcastle University which had a total of 4500 mice from which approximately 10% were stock animals when the assessment was made (information provided by the facility). The sample size was determined based on previous studies about animal welfare protocols were approximately 5% to10% of the animals are used for the assessment (Welfare Quality®, 2009a; Battini *et al.*, 2015). A total of 8 cages were assessed (sample size for the protocols section 5.4.1). Twenty-nine mice (12 female and 17 males) C57BL/6J (Charles River Laboratories, UK) aged approximately eight weeks were used for this practicability assessment. Mice were housed in pairs or groups of 4 in eight cages. They were housed in IVC cages (NexGen, Allentown). The home cages had sawdust bedding, nesting material (Enviro-Dri® and Sizzle-Pet) and were cleaned out once per week. Animals had *ad-libitum* access to food (Special Diet Services, RM3E diet) and water. The practicability assessment evaluated the preliminary list of 51 indicators.

The practicability assessment was carried out once for each individual indicator by one researcher (ICL), assessing each animal cage. Data were recorded on a four-point scoring system for each of 3 categories. The categories comprised the time spent assessing the indicator, the resources needed and the easiness of scoring (Table 5.3). The time spent category consisted of 4 levels, very time consuming (over 60 seconds), time-consuming (from 30- 60 seconds), quick (from 15-30 seconds), and very quick (less than 15 seconds). The resources needed involved 4 levels, very specific equipment needed (test kit usually required per individual animal), specific equipment needed (test kit needed, but can be used for several animals), equipment needed but is routinely available in the facility (e.g. thermometer, humidimeter) and no equipment needed. Ease of scoring had 4 levels: very difficult to score (e.g. requires a long period of observation and calculations, i.e. over 5 minutes), difficult to score (requires a long period of observation from 2 to 5 minutes), easy to score (observation for a short period from 5 seconds to 2 minutes), very easy to score (usually done in conjunction with another measurement, less than 5 seconds).

5.3.2. Results

Practicability assessment

The mean time spent recording each indicator is shown in Table 5.4. The mean total time to complete the assessment was also calculated.

Cage indicators	Mean (Minutes)	95% CI of the Difference	
		Lower	Upper
Housing density and cage dimensions	0.21	0.13	0.30
Use of cage space	0.19	-0.09	0.47
Complexity of the cage	0.13	0.05	0.20
Pups outside the nest	0.40	0.15	0.65
Use of nesting material	0.38	0.27	0.48
Exhibition of normal and abnormal behaviour	0.38	0.19	0.56
Qualitative Behavioural assessment (QBA)	0.50	0.42	0.58
Excessive urine and faeces in the cage	0.08	0.01	0.15
Blood stains in the cage	0.14	0.08	0.20
Average time	**0.27**	0.13	0.39
Total Time	**2.41**		
Individual indicators	Mean (Minutes)	95% CI of the Difference	
		Lower	Upper
Mouse grimace scale	0.50	0.44	0.56
Tail position	0.14	0.04	0.24
Skin colour	0.15	0.09	0.21
Gait	0.11	0.03	0.20
Hunched position	0.25	0.16	0.34
Ocular/nasal discharge	0.16	0.10	0.23
Wounds/marks (including bite wounds)	0.11	-0.03	0.26
Easy of handling	0.20	0.12	0.28
Body Condition Score	0.46	0.39	0.54
Body temperature	0.51	0.41	0.62
Weight loss	0.53	0.40	0.65
Dehydration	0.15	0.05	0.25
Swollen abdomen	0.10	0.04	0.16
Average time	**0.26**	0.17	0.35
Total Time	**3.37**		

[1] indicators with 0.00 time spent were assessed at the same time as other indicators. Thus they did not require additional time.

Table 5.4. Table of the mean time spent per indicator during the practicability assessment (n=8 cages).

The remaining data collected during the practicability assessment of all of the indicators are shown in Table 5.5. The indicators that showed high levels of practicability (100%) were mainly related to the environment. However, some cage level indicators, e.g. use of nesting material and alertness also showed high practicability (100%). Six individual-level indicators (alertness, bloodstain in the cage, coat condition, hunched posture, urination/defecation during handling, wounds/marks- including bite wounds) also had 100% or practicability.

Indicator	**Time spent**	**Resources needed**	**Ease of scoring**	**Mean score**	**Practicability**
Alertness	4	4	4	4	100.00%
Audible signs of aggression	4	4	2	3.33	83.30%
Barbering (hair removal)	3	4	4	3.67	91.70%
Blood stains in the cage	4	4	4	4	100.00%
Body Condition Score	3	4	3	3.33	83.30%
Body temperature	1	1	1	1	25.00%
Cage odour (ammonia levels)	2	1	1	1.33	33.30%
Coat condition	4	4	4	4	100.00%
Complexity of the cage	4	4	4	4	100.00%
Dehydration	2	4	2	2.67	66.70%
Disease rate	1	3	1	1.67	41.70%
Easy of handling	3	4	4	3.67	91.70%
Empathetic attitude of staff towards animals	1	4	1	2	50.00%
Excessive urine and faeces in the cage	4	4	4	4	100.00%
Exhibition of normal and abnormal behaviour	1	3	1	1.67	41.70%
Food type	3	4	1	2.67	66.70%
Frequency of cleaning cages	4	4	4	4	100.00%
Frequency of feeding	3	4	1	2.67	66.70%
Frequency of physical examination	3	4	1	2.67	66.70%
Frequency of veterinary procedures	3	4	1	2.67	66.70%
Gait	3	4	3	3.33	83.30%
Handling method	3	4	4	3.67	91.70%
Housing density and cage dimensions	2	2	3	2.33	58.30%
Humidity	4	2	4	3.33	83.30%
Hunched posture	4	4	4	4	100.00%
Infanticide rate	1	3	1	1.67	41.70%
Intensity of lighting	2	1	4	2.33	58.30%
Light/dark cycle	4	4	4	4	100.00%
Method of feeding	3	4	4	3.67	91.70%
Mortality rate	1	3	1	1.67	41.70%
Mouse grimace scale	3	4	4	3.67	91.70%
Nesting material	4	4	4	4	100.00%

Ocular/nasal discharge	3	4	4	3.67	91.70%
Positive and negative interaction with conspecifics	1	3	1	1.67	41.70%
Presence/absence of bedding material	4	4	4	4	100.00%
Pups outside the nest	4	4	4	4	100.00%
Qualitative Behavioural Assessment (QBA)	4	4	2	3.33	83.30%
Respiratory pattern	2	4	2	2.67	66.70%
Skin colour	2	4	4	3.33	83.30%
Staff training	4	4	2	3.33	83.30%
Swollen abdomen	3	4	3	3.33	83.30%
Tail position	3	4	3	3.33	83.30%
Temperature-humidity-ventilation	4	2	4	3.33	83.30%
Type of bedding material	4	4	4	4	100.00%
Type of cleaning procedure	4	4	1	3	75.00%
Urination/defecation during handling	4	4	4	4	100.00%
Use of cage space	4	4	3	3.67	91.70%
Use of nesting material	4	4	4	4	100.00%
ventilation	4	2	4	3.33	83.30%
Weight loss	1	3	4	2.67	66.70%
Wounds/marks (including bite wounds)	4	4	4	4	100.00%

Table 5.5. Practicability assessment of all of the indicators using the mean score. The final practicability percentage was provided from the mean score of the 4 categories.

The information from the practicability assessment along with the information on validity and reliability from the Delphi study (Chapter 2) was used to select the indicators to move forward with the construction of the audit and everyday welfare assessment protocol. The selection process consisted of three stages: 1. classification of the indicators into the three components of the animal welfare definition: biological functioning (physiological parameters), psychological functioning (mental well-being) and the ability to meet their evolutionary needs (environment), 2. Removal of all indicators where validity and reliability levels in the Delphi study (Chapter 2) were below 70% consensus, 3. Removal of indicators where the practicability level was below 70%. There is no information in the literature about the acceptable level of practicability for animal welfare indicators. Thus the level of 70% was set up for this assessment based on scales used in education (Buchanan-Smith *et al.*, 2005). The final list consisted of 20 indicators for the audit welfare assessment and 16 for the everyday welfare protocol. For each protocol, a document was created which included detailed information describing the indicator and the scoring method (i.e. the scale used for the scoring) (Appendix D and F).

List of indicators for the audit unit welfare assessment (Phase 1)

This protocol included 20 indicators; 8 biological functioning, 4 psychological state and 8 evolutionary needs. Table 5.6 shows the final list of indicators which were divided into three separate scoring sheets for the assessment: [1] facility/room level (Figure 5.2), [2] cage level (Figure 5.3) and [3] animal level (Figure 5.4). The assessors had one sheet per facility/room, one sheet per cage and one sheet per animal. The cage ID number and the mouse individual's identification were included on the sheet. For the cage and the individual assessment sheet, assessors were instructed to observe the cage and animals for 5 minutes taking into account the average time spent on the indicators during the pilot assessment (Table 5.4). Qualitative Behavioural Assessment was included in the cage level assessment sheet as it needs to be performed in the home cage without interference. The QBA sheet consisted of the 20 fixed behavioural expressions (from Chapter 4), with a Visual Analogue Scale for the scoring for each expression (Figure 5.5).

Welfare principle	Indicator	Validity[1]	Reliability[1]	Practicality from pilot	Audit rank[2]
Biological functioning	Hunched posture	99.3	94	100.0	1
	Coat condition	99.3	88.8	100.0	2
	Body Condition Score	97.8	91	83.3	3
	Wounds/marks (including bite wounds)	97	93.6	100.0	13
	Blood stains in the cage	94.8	100.0	100.0	26
	Gait	94.1	75.4	83.3	13
	Swollen abdomen	97.7	83.3	83.3	24
	Ocular/nasal discharge	94.7	85.6	91.7	14
Psychological state	Usage of nesting material	87.2	68.4	100.0	10
	Alertness	86.6	67.4	100.0	11
	Barbering (hair removal)	73.3	74.6	91.7	32
	QBA [3]			83.3	
Evolutionary needs (Environment)	Staff training	99	86	83.3	9
	Complexity of the cage	84	76	100.0	21
	Temperature	95	92	83.3	23
	Ventilation	93	87	83.3	34
	Humidity	90	85	83.3	35
	Type of cleaning procedure	70	75.0	83.3	54
	Frequency of cleaning cages	88	100.0	83.3	43
	Handling method[3]			91.7	

[1]validity and reliability are shown in percentage values taken from the Delphi consultation. [2] audit rank value from Delphi consultation survey. [3] QBA and handling method were added after the Delphi consultation, so no information about validity, practicability or reliability was available.

Table 5.6. The final list of 20 indicators used for the audit welfare assessment.

Each score sheet had options to select from taking into account the observation made in the assessment. The assessor selected one of the options available for each indicator which was converted to the traffic light system score based on the information of the scale given in the protocols (Appendix D-F).

FACILITY ASSESSMENT

DATE:
FACILITY:
START TIME: FINISH TIME:

Welfare indicator	Scale	SCORE
Temperature	Current room temperature	
Humidity	Current room humidity	
Ventilation	Current room/cage ventilation	
Handling method	Proportion of animal handling by tail	
	Proportion of animals handling by cup	
	Proportion f animals handling by tunnel	
Frequency of cleaning cages	Frequency of cleaning cages	
	Number of air changes per hour in the cages	
Type of cleaning procedure	Mice are transferred to a clean cage with clean sawdust and the nesting material from the dirty cage	
	Mice are transferred to a clean cage with clean sawdust, dirty sawdust for the dirty cage and clean nesting material	
	Mice are transferred to a clean cage with clean sawdust and clean nesting material.	
Staff training	Mainly all the staff have PDS related to the care of laboratory animals (more than 51%)	
	There is a proportion (30 % to50%) of the staff that have PDS	
	Staff do not have PDS (less than 29%)	

Figure 5.2. Facility-level assessment score sheet used for the audit welfare assessment. There is an explanation of the scale and a score section for each indicator.

Cage assessment

START TIME: FINISH TIME:
CAGE ID:

Welfare indicator	Scale	score
Complexity of the cage	Additional elements or structures present	
	Only nesting material as additional element	
	No additional material or elements	
Alertness	Mouse alert when cage lid is open	
	It might respond to the observer but not immediately.	
	Mouse does not show any interest in the environment or the observer.	
Usage of nesting material	Complete or incomplete dome	
	A cup-shaped or flat nest	
	Untouched or scattered nesting material throughout the cage	
Qualitative Behavioural assessment (QBA)	See next page	
Blood stains in the cage	Blood stains absent in the home cage	
	Blood stains present in the home cage	

Figure 5.3. Cage level assessment score sheet used for the audit welfare assessment.

INDIVIDUAL ASSESSMENT

START TIME: FINISH TIME:

CAGE ID

MOUSE ID

Indicator	scale	Score
Hunched posture	No hunched posture present	
	Walks slowly in a hunched posture	
	Hunched posture with no movement	
Coat condition	Shiny, clean and well-groomed coat	
	Coat clean but ungroomed	
	Ruffled and untidy coat, it can be greasy and stick together	
Gait	Body weight supported on all limbs	
	Mouse might limp while walking	
	Mouse has difficulty moving forward or is reluctant to walk	
Ocular/nasal discharge	No ocular/nasal discharge	
	Water-like or mucus-like ocular/nasal discharge	
Wounds/marks (including bite wounds)	No wounds	
	Superficial wounds	
	Extensive and deep wounds	
Body Condition Score	The mouse is well-conditioned	
	The mouse in either under-conditioned or overweight	
	The mouse is either emaciated or obese.	
Barbering (hair removal)	No barbering	
	Barbering on different parts of the body	
Swollen abdomen	Abdomen is not swollen	
	Abdomen moderately swollen	
	Abdomen is very swollen	

COMMENTS:

Figure 5.4. Individual-level assessment sheet used for the audit animal welfare protocol.

Qualitative behavioural assessment

CAGE ID:

DATE:

Inquisitive	Min.	Max.
In pain	Min.	Max.
Positively Engaged	Min.	Max.
Lethargic	Min.	Max.
Energetic	Min.	Max.
Depressed	Min.	Max.
Determined	Min.	Max.
Anxious	Min.	Max.
Confident	Min.	Max.
Agitated	Min.	Max.
Calm	Min.	Max.
Fearful	Min.	Max.
Content	Min.	Max.
Tense	Min.	Max.
Comfortable	Min.	Max.
Uncertain	Min.	Max.
Playful	Min.	Max.
Bored	Min.	Max.
Sociable	Min.	Max.
Frustrated	Min.	Max.

Figure 5.5. Qualitative Behavioural Assessment sheet for use in audit welfare assessment.

List of indicators for the everyday welfare assessment

This protocol included 16 indicators; 8 biological functioning, 4 psychological state and 4 evolutionary needs. The everyday welfare protocol was also divided into three levels of assessment; facility/room, cage and animal level. The protocol also had individual score sheets for each level of assessment, facility/room, and cage (Figure 5.6), and animal levels (Figure 5.7) and QBA fixed terms (Figure 5.5). All the score sheets are shown in Appendix E. The final indicators included in this protocol are shown in Table 5.7.

FACILITY/CAGE ASSESSMENT:

FACILITY:

CAGE ID:

Indicator	Scale	Date				
		02/06/18	03/06/18	04/06/18	05/06/18	06/06/18
Temperature	Current room temperature					
Humidity	Current room humidity					
Ventilation	Current room/cage ventilation					
Complexity of the cage	Additional elements or structures present					
	Only nesting material as additional element					
	No additional material or elements					
Alertness	Mouse alert when cage lid is open					
	It might respond to the observer but not immediately.					
	Mouse does not show any interest in the environment or the observer.					
Usage of nesting material	Complete or incomplete dome					
	A cup-shaped or flat nest					
	Untouched or scattered nesting material throughout the cage					
Qualitative Behavioural assessment (QBA)	See next page					
Blood stains in the cage	Blood stains absent in the home cage					
	Blood stains present in the home cage					

Figure 5.6. Facility and cage assessment sheet used for the everyday protocol assessment. Each indicator had their scale or completion information and the dates when the assessment was carried out.

INDIVIDUAL ASSESSMENT

CAGE ID

MOUSE ID

Indicator	Scale	Date				
		02/06/18	03/06/18	04/06/18	05/06/18	06/06/18
Hunched position	No hunched posture present					
	Walks slowly in a hunched posture					
	Hunched posture with no movement					
Coat condition	Shiny, clean and well-groomed coat					
	Coat clean but ungroomed					
	Ruffled and untidy coat, it can be greasy and stick together					
Gait	Body weight supported on all limbs					
	Mouse might limp while walking					
	Mouse has difficulty moving forward or is reluctant to walk					
Ocular/nasal discharge	No ocular/nasal discharge					
	Water-like or mucus-like ocular/nasal discharge					
Wounds/marks (including bite wounds)	No wounds					
	Superficial wounds					
	Extensive and deep wounds					
Body Condition Score	The mouse is well-conditioned					
	The mouse in either under-conditioned or overweight					
	The mouse is either emaciated or obese.					
Barbering (hair removal)	No barbering					
	Barbering on different parts of the body					
Swollen abdomen	Abdomen is not swollen					
	Abdomen moderately swollen					
	Abdomen is very swollen					

COMMENTS:

Figure 5.7.Scoresheet for the individual level indicators used in the everyday protocol.

Welfare principle	Indicator	Validity[1]	Reliability[1]	Practicality from pilot[4]	Everyday rank position[2]
Biological functioning	Hunched position	99.3	94	100.0	1
	Coat condition	99.3	88.8	100.0	2
	Wounds/marks (including bite wounds)	97	93.6	100.0	5
	Blood stains in the cage	94.8	74.4	100.0	12
	Body Condition Score	97.8	91	83.3	13
	Ocular/nasal discharge	94.7	85.6	91.7	14
	Gait	94.1	75.4	83.3	15
	Swollen abdomen	97.7	71.6	83.3	19
Psychological functioning	Usage of nesting material	87.2	68.4	100.0	8
	Alertness	86.6	67.4	100.0	10
	Barbering (hair removal)	73.3	74.6	91.7	21
	QBA [3]			83.3	
Evolutionary needs	Temperature	95	92	83.3	5
	Humidity	90	85	83.3	11
	Ventilation	93	87	83.3	24
	Complexity of the cage	84	76	100.0	24

[1]validity and reliability are shown in percentage values taken from the Delphi consultation. [2] The everyday rank value is taken from the Delphi consultation survey. [3] QBA was added after the Delphi consultation, so no information about validity, practicability or reliability was available.[4] practicability was taken from the pilot thus is the same for both audit and everyday assessment.

Table 5.7. The final list of 16 indicators used for the everyday welfare assessment.

5.4. Reliability assessment of the audit welfare protocol

5.4.1. Methods

Protocols scoring system

A three-level traffic light scoring system was developed for each individual indicator, one for each welfare principle and finally one for the overall welfare assessment, i.e. at a facility level (Figure 5.8). This system assigns a colour-coded category (green, amber or red) to each section of the system defined by recommendations taken from the Code of Practice for the Animals used in Scientific Procedures (Home Office, 2014a) where they exist and the document related to the reduction of animals in scientific research (Home Office, 2014b). When a specific recommendation was not available, existing literature concerning the measurement levels and standards for that indicator was used to provide the scores. This three-level traffic light system has been used in the development of other animal welfare protocols (Leach *et al.*, 2006; Council, 2009b; Sørensen and Fraser, 2010). Green was used when the indicator level is above the recommendations, amber when the indicator was within the recommended levels and red when the indicator was below the recommendations. For example, for an indicator that has recommended levels based on literature such as coat condition, ‘green’ was used when the coat is shiny and well-groomed; ‘amber’, when the coat was dull but groomed, and ‘red’ when the coat was ruffled, untidy and greasy as suggested by Paster *et al.* (2009).

Figure 5.8 shows that for individual indicators, the traffic light system was simply used to report whether they were above recommendations (‘green’), within recommendations (‘amber’) or below recommendations (‘red’). At the welfare principle level categorisation (i.e. green, amber or red) was based upon the scores of the individual indicators that comprise. Each principle level received the traffic light system corresponding to the highest percentage of the individual indicators that they are representing. If the indicators had equal percentage scores, then a precautionary principle was applied, and the principle received the lowest score. As the three principals were equally important for the assessment, the final score system for the assessment of animal welfare was provided in terms of the three principles.

The traffic score system was also adapted for the Qualitative Behavioural Assessment indicators. The adaption was made taking into account how others integrated this indicator into the welfare assessment (Welfare Quality®, 2009c; Battini *et al.*, 2015). In the QBA output, the Principal Component Analysis (PCA) plot was generated. Cage was located in the plot which is divided into four quadrants: good/calm (quadrant one), good/energetic (quadrant two), bad/calm (quadrant three), and bad/energetic (quadrant four). A ‘green’ score was given was the PCA plot showed the cages located in either the positive arousal/positive emotional valence quadrant or in the negative arousal/positive emotional valence quadrant. An ‘amber’ score was given were in PCA plot the cages were located in positive arousal/negative emotional valance quadrant considering other protocols scoring scheme (e.g. welfare quality). A ‘red’ score was given when the PCA plot showed the cages located in negative arousal/negative emotional valance quadrant.

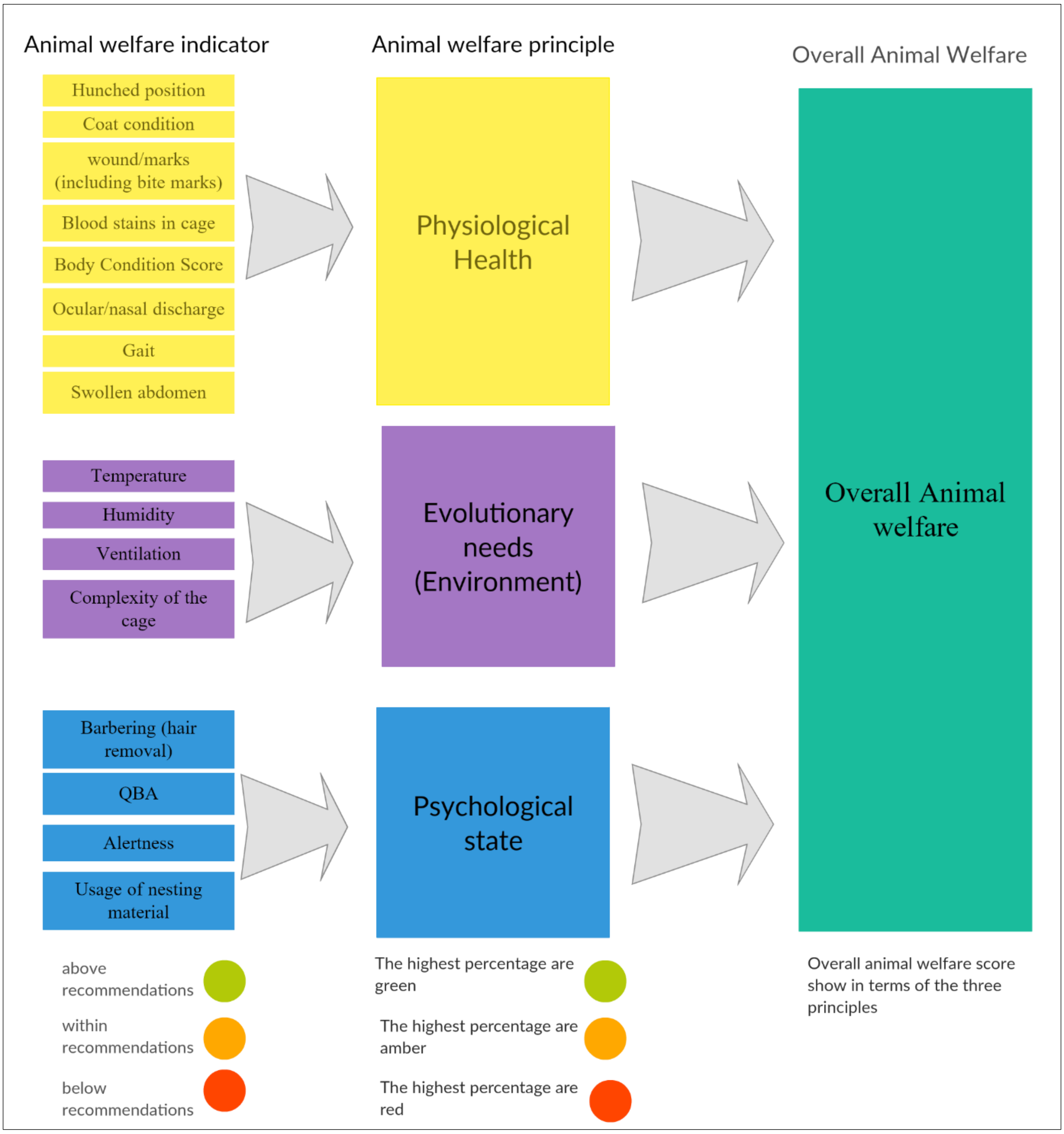


Figure 5.8. Traffic light system used for reporting the results of the laboratory mouse welfare assessment. This figure represents the everyday welfare protocol where each animal welfare indicator represents a welfare principle.

Sample size for the protocols

The total number of cages/animals required for assessment was dependent on the total number of stock animals in each facility. Literature suggests that the total number of animals that need to be assessed can be from 25% of the population for dairy cows (Welfare Quality®, 2009c) to 5% of the population of dairy goats (Battini *et al.*, 2015). There is no specific information for laboratory mice, and because the population of this animal is usually over 1000 animals per unit, therefore, the sample size was defined from 5% to 10% of the population of stock animals within the laboratory facility. Stock cages were selected randomly from within the animal room.

Laboratory facilities and animals assessed

The protocol was used for the assessment of welfare of stock laboratory mice at Newcastle and Dundee Universities. In both locations, the animals were housed in IVC cages, with environmental conditions, e.g. humidity, temperature controlled at the room level. The facility at Newcastle, where the assessment of the audit protocol was conducted, held an average of 6000 mice, of which about 7% (i.e. ~420 mice) were considered stock animals. Dundee University had an average of 4000 mice, of which approximately 6% (i.e. ~ 240 mice) were stock animals.
At Newcastle University, the protocol was completed by two veterinarians and one researcher all with over 5 years of experience of working laboratory mice. At Dundee University, the protocol was performed by one technician with over 15 years of experience and one veterinarian with over three years of experience working with laboratory mice. Both studies involved a one-time assessment which lasted one to two hours within each facility (duration dependent on the number of stock animals present in the unit at the moment of the assessment). The protocol was sent to the facilities one week in advance to allow familiarisation.

On the day of the assessment, additional information about QBA was provided to ensure that participants were familiar with this indicator. Assessors received information about the use of QBA as an animal welfare indicator in other species, how it can be applied to mice and information about the previous studies in this thesis (Chapters 3 and 4). Following this short training session, assessors were asked not to discuss their scores with each other during the assessment. They were informed that the aim of the study was not to assess the welfare of the animals, but the practicability and reliability of the protocol. At Newcastle University, 32 stock female C57BL/6 mice (Charles River Laboratories, UK) divided across 10 cages were assessed. The mice were housed in pairs or groups of three or four in IVC NexGen cages (Allentown). Home cages had sawdust bedding and nesting material (Enviro-Dri® and Sizzle-Pet), a cardboard tube and a chew stick. Animals had *ad-libitum* access to water and food (Special Diet Services, RM3E diet) and sunflower seeds as an additional environmental enrichment. At Dundee University, 40 BALB/c mice (10 males and 30 females: Charles River Laboratories, UK) divided across 10 cages were assessed. The mice were housed in pairs or groups of four in IVC cages (Arrowmight) with bedding and nesting material (LBS biotechnology) and *ad-libitum* access to water and food (SDS Mouse breeder and growing, LBS biotechnology).

The assessment was performed in both facilities in the morning, after the daily checks, after the facility/cage indicators were assessed, the cages were removed one by one and placed in a cage changing station for the assessment. The assessment of the cage level indicators did not involve any handling of the animals. After the cage level assessment, each animal within the cage was assessed individually through observation and without handling the animals.

5.4.2. Data Analysis

The inter-observer agreement was assessed using descriptive statistics and Attribute Agreement Analysis. Attribute Agreement Analysis is useful for determining areas of non-agreement and calibration for a high level of agreement and is used for discrete datasets (Quinn *et al.*, 2008; Aksoy and Orbak, 2009). The inter-observer agreement for QBA was carried out using Principal Component Analysis (PCA) (See Chapter 4 for further details). Briefly, the information on the scores given for each fixed term on the QBA cage assessment sheet (Figure 30) was analysed using PCA which uses an algorithm for reducing the dimensionality of the fixed terms (Jolliffe, 2011). This data analysis identifies patterns in the data created when the assessor scores the same cages using the same terms. PCA identifies then the common patterns in the data producing principal component (dimensions) which explain the variability between the assessors (Rencher, 2003; Temple *et al.*, 2011). Inter-observer reliability of the QBA observations was assessed using Kendall Correlation Coefficient. This test aimed to determine to what extent the assessors agreed with each other on the use of the dimensions when assessing the cages using the fixed terms during the audit protocol. The data analysis was carried out using Minitab® 17.1.0 (2013 Minitab Inc. license for windows) for the PCA analysis and the Attribute Agreement Analysis, R© (version 1.0.153 – © 2009-2017 RStudio, Inc.) and SPSS statistical analysis package version 24 (IBM Corp. SPSS (v23, SPSS Inc., Chicago, USA) for Kendall’s correlation of coefficient (*W*).

5.4.3. Results

Audit - Inter-observer agreement at Newcastle University

The total time spent assessing the ten cages (32 animals) was 68 minutes with an average of 6 minutes per cage. The scores for the environmental indicators at facility level showed a high level of agreement between the three observers. The scores for each environmental indicator given by the participants are shown in Table 5.8. All observers agreed with the scores, which is further illustrated by the complete lack of variability (Standard Deviation). Handling method was scored as ‘red’ (score of 3), and temperature was scored as ‘amber’ (score of 2). The rest of the indicators (humidity, ventilation, frequency of cleaning cages, type of cleaning procedure and staff training) received a score of ‘green’ (score of 1).

Indicator (n=3)	Minimum	Maximum	Mean	Std. Deviation
Temperature	2	2	2.00	.000
Humidity	1	1	1.00	.000
Ventilation	1	1	1.00	.000
handling method	3	3	3.00	.000
frequency of cleaning cages	1	1	1.00	.000
type of cleaning procedure	1	1	1.00	.000
staff training	1	1	1.00	.000

Table 5.8. Descriptive statistics for the environmental indicators used in the audit welfare assessment at Newcastle University. The scores using the traffic light system are represented with a numerical code (1= green, 2= amber and 3= red).

The descriptive statistics for the cage level indicators are shown in Table 5.9. All observers agreed on the scores of the complexity of the cage and bloodstains in the cage with a 100% agreement between assessors shown in the Attribute Agreement Analysis (Table 5.10). These indicators were scored ‘green’ (score of 1). Different cages had different scores given by the assessors for alertness and use of nesting material, with the Attributed Agreement analysis showing 70% agreement for alertness scored as ‘green’ (1) and 60% for usage of nesting material scored as ‘amber’ (2). The total average agreement of the cage level indicators was 88.3% and therefore given a ‘green’ score (1).

Descriptive Statistics					
Indicator (N=10)	Range	Minimum	Maximum	Mean	Std. Deviation
Complexity of the cage	0	1	1	1.00	.000
Alertness	1	1	2	1.10	.305
Usage of nesting material	1	1	2	1.63	.490
Blood stains in cage	0	1	1	1.00	.000

Table 5.9. Descriptive statistics for scores of the cage level indicators used in the audit assessment at Newcastle University. The scores, using the traffic light system, are represented with a numerical code (1= green, 2= amber and 3= red).

Attribute Agreement Analysis					
Indicator	Inspected	Matched	Percent	95% CI	
Complexity of the cage	10	10	100%	74.11	100
Alertness	10	7	70%	34.75	93.33
Usage of nesting material	10	6	60%	26.24	87.84
Blood stains in cage	10	10	100%	74.11	100

Table 5.10. Attribute agreement analysis for the scores of the cage level indicators used as part of the audit welfare protocol.

From the analysis of the QBA, the two main dimensions explain 58% of the variation in the cages observed (PC1: 44% and PC2: 14%). Dimensions were labelled from confident/positive engaged to anxious/frustrated for Dimension 1 and from playful/determined to uncertain/tense for Dimension 2 (Figure 5.9).

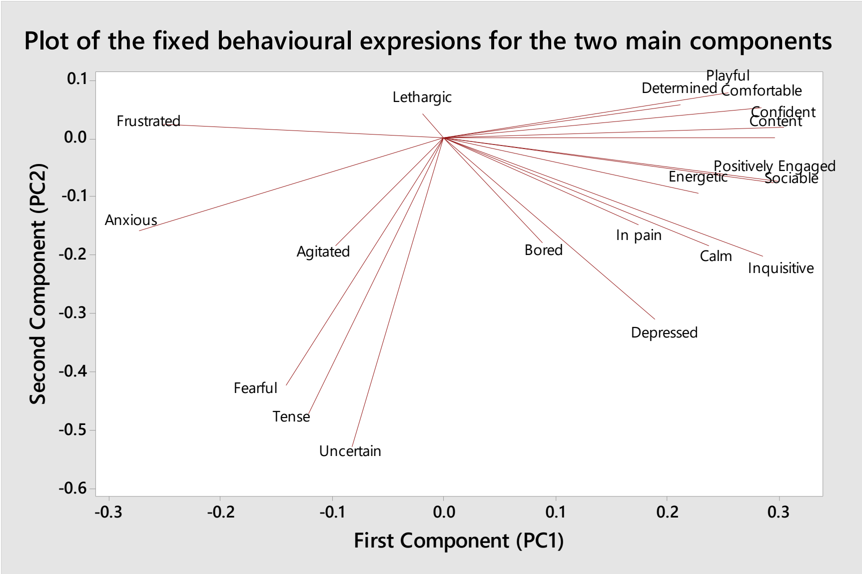


Figure 5.9. Distribution of the 20 fixed behavioural expressions used as part of the audit welfare protocol at Newcastle University. The x-axis represents Dimension 1, and the y-axis represents dimension 2.

The Kendall Correlation Coefficient showed good levels of agreement for the QBA terms between assessors, in this audit welfare assessment protocol. The assessors showed a good level of agreement for PC1, *W*= 0.790 (n = 10, df = 2), and for PC2, *W* = 0.73, (n=10, df=2). Closer inspection of the plot shows that the assessors allocated the cages consistently into the same quadrants. Assessor one allocated the cages in the quadrant labelled as playful/determine and anxious frustrated given a score of 'red' (score of 3). Assessor two allocated the cages in the confident/positively engaged and so scored 'amber' (score of 2) and assessor three in the anxious frustrated and confident/positive engaged and so scored 'amber' (score of 2) (Figure 5.10).

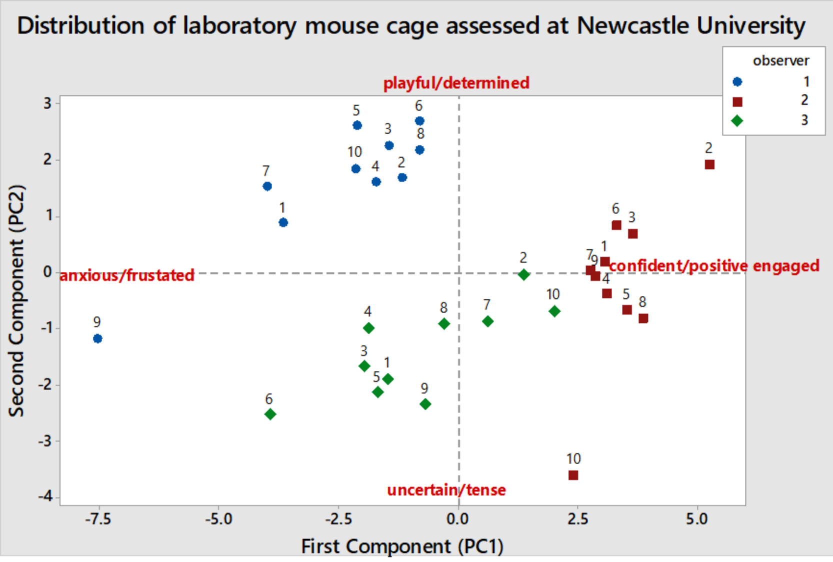


Figure 5.10. Plot distribution of the ten cages assessed in by three observers in the half-fay welfare protocol at Newcastle University. The x-axis dimension 1, and the y-axis represents dimension 2. Each dot represents one cage scored by one assessor. The red legends are the behavioural expressions that represent each dimension.

All observers agreed with the animal level indicators scores (i.e.100% agreement). All 32 were scored as 'green' (1) in all the eight animal level indicators within the 95% CI (91.06, 100).

The summary of audit welfare assessment protocol at Newcastle University is shown in Figure 5.11 including the traffic light system as a final score.

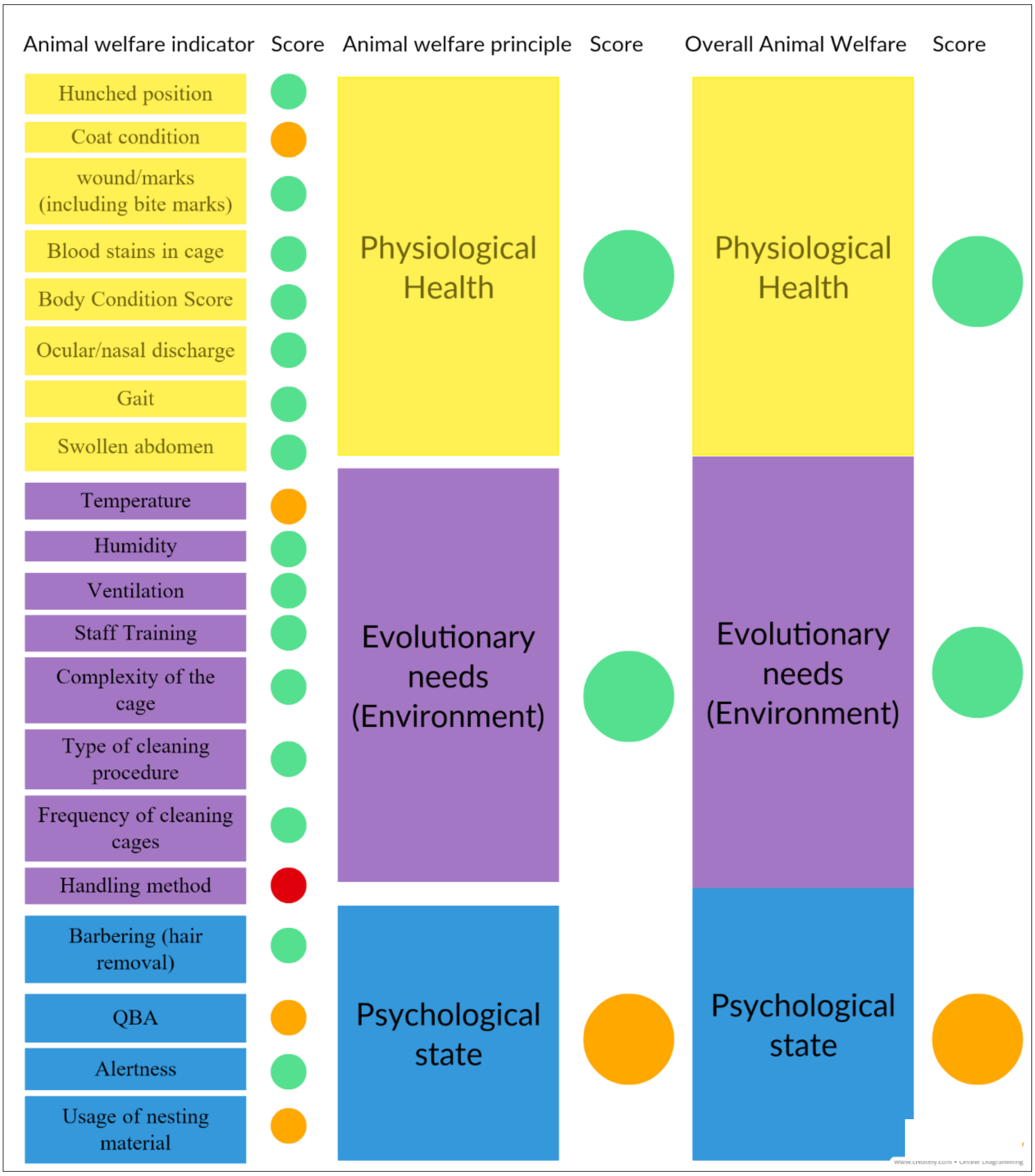


Figure 5.11. Summary of audit assessment protocol used at Newcastle University including the traffic light system score.

Inter-observer agreement at Dundee University – audit

The total time spent assessing the ten cages (40 animals), was 53 minutes with an average of 5 minutes per cage. For the facility/room level indicators, the two assessors gave identical scores for humidity, ventilation, and the frequency of cleaning cages and type of cleaning procedure, which were all ‘green’ (1), and temperature, handling method and staff training were all ‘amber’ (2) (Table 5.11).

Descriptive Statistics				
N= 10	Minimum	Maximum	Mean	Std. Deviation
Temperature	2	2	2.00	.000
Humidity	1	1	1.00	.000
Ventilation	1	1	1.00	.000
handling method	2	2	2.00	.000
frequency of cleaning cages	1	1	1.00	.000
type of cleaning procedure	1	1	1.00	.000
staff training	2	2	2.00	.000

Table 5.11. Descriptive statistics for the environmental indicators used in the audit welfare assessment at Dundee University. The scores using the traffic light system are represented with a numerical code (1= green, 2= amber and 3= red).

There was a high level of agreement for the cage level indicators (Table 5.12), with the Attribute Agreement Analysis shown in Table 5.13. Complexity of the cage and bloodstains on the cage had 100% of agreement scoring ‘green’ (1). Alertness had 90% agreement with ‘amber’ scores (2), and usage of nesting material had 60% agreement with ‘amber’ scores (2).

Descriptive Statistics				
N= 10	Minimum	Maximum	Mean	Std. Deviation
Complexity of the cage	1	1	1.00	.000
Alertness	1	2	1.05	.224
Use of nesting material	1	2	1.50	.513
Blood stains in cage	1	1	1.00	.000

Table 5.12. Descriptive statistics for the cage level indicators used in the audit welfare protocol at Dundee University. The scores using the traffic light system are represented with a numerical code (1= green, 2= amber and 3= red).

Attribute Agreement Analysis					
Indicator	Inspected	Matched	Percent	95% CI	
Complexity of the cage	10	10	100%	74.11	100
Alertness	10	9	90%	55.50	99.75
Usage of nesting material	10	6	60%	26.24	87.84
Blood stains in cage	10	10	100%	74.11	100

Table 5.13. Attribute agreement analysis performed for the four cage level indicators in the audit welfare assessment at Dundee University.

From the analysis of the QBA analysis, the two main dimensions explained 61% of the variation of the cages observed; Dimension 1 (48%) was labelled from social/positively engaged to agitated/uncertain, and dimension 2 (13%) was labelled from playful to depressed/bored (Figure 5.12).

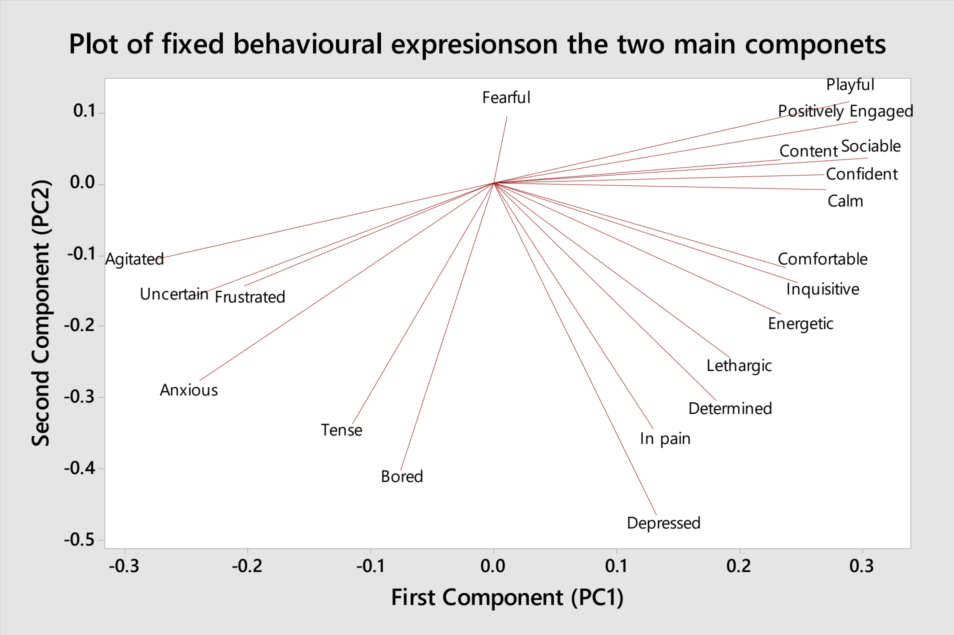


Figure 5.12. Distribution of the 20 fixed behavioural expressions used in the Qualitative Behavioural Assessment as part of the audit animal welfare protocol at Dundee University. The x-axis represents dimension 1, and Y-axis represents dimension 2.

Kendall Coefficient of Concordance showed moderate agreement for PC1 for both assessors (Kendal's *W*= 0.53), and no agreement for PC2 (Kendal's *W*= 0.0). The plot of the location of the cages in the dimensions (Figure 5.13) showed that the two assessors located the cages in two different quadrants. Assessor one distributed the cages in the left quadrant described them as agitated/uncertain and playful, which is scored as 'red' (score of 3). Assessor two located the cages in the right quadrant described as social/positive engaged which was scored as 'green' (score of 1).

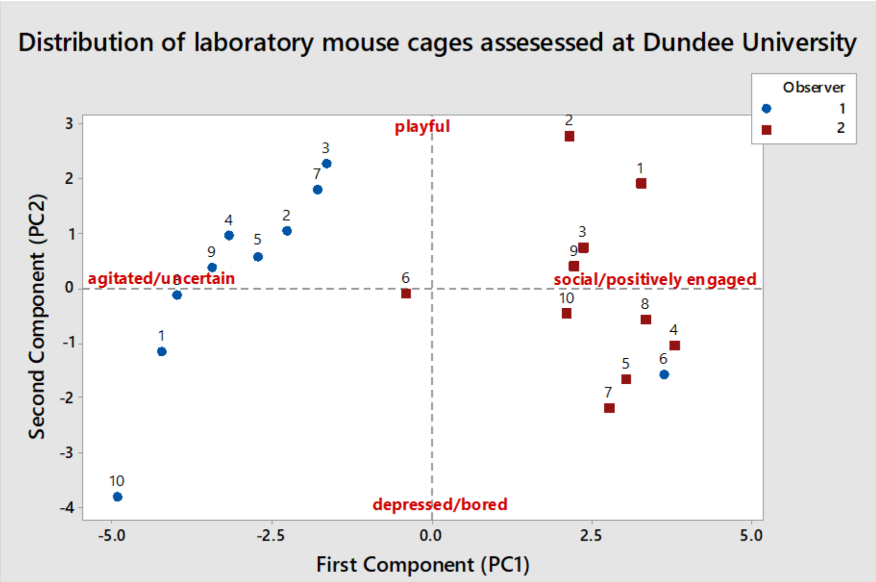


Figure 5.13. Plot distribution of the ten cages assessed by two observers in the audit welfare assessment at Dundee University. The x-axis represents dimension 1, and the y-axis dimension 2. Each point represents one cage assessed by one observer. The red legends are the behavioural expressions that represent each dimension.

All observers agreed regarding the animal level indicators scores (i.e.100% agreement), with all indicators scored a ‘green’ (score of 1) for all the 40 mice assessed with 95% CI (74.11, 100).

The summary of the results from the audit protocol at Dundee University including the final traffic light system results is showed in Figure 5.14

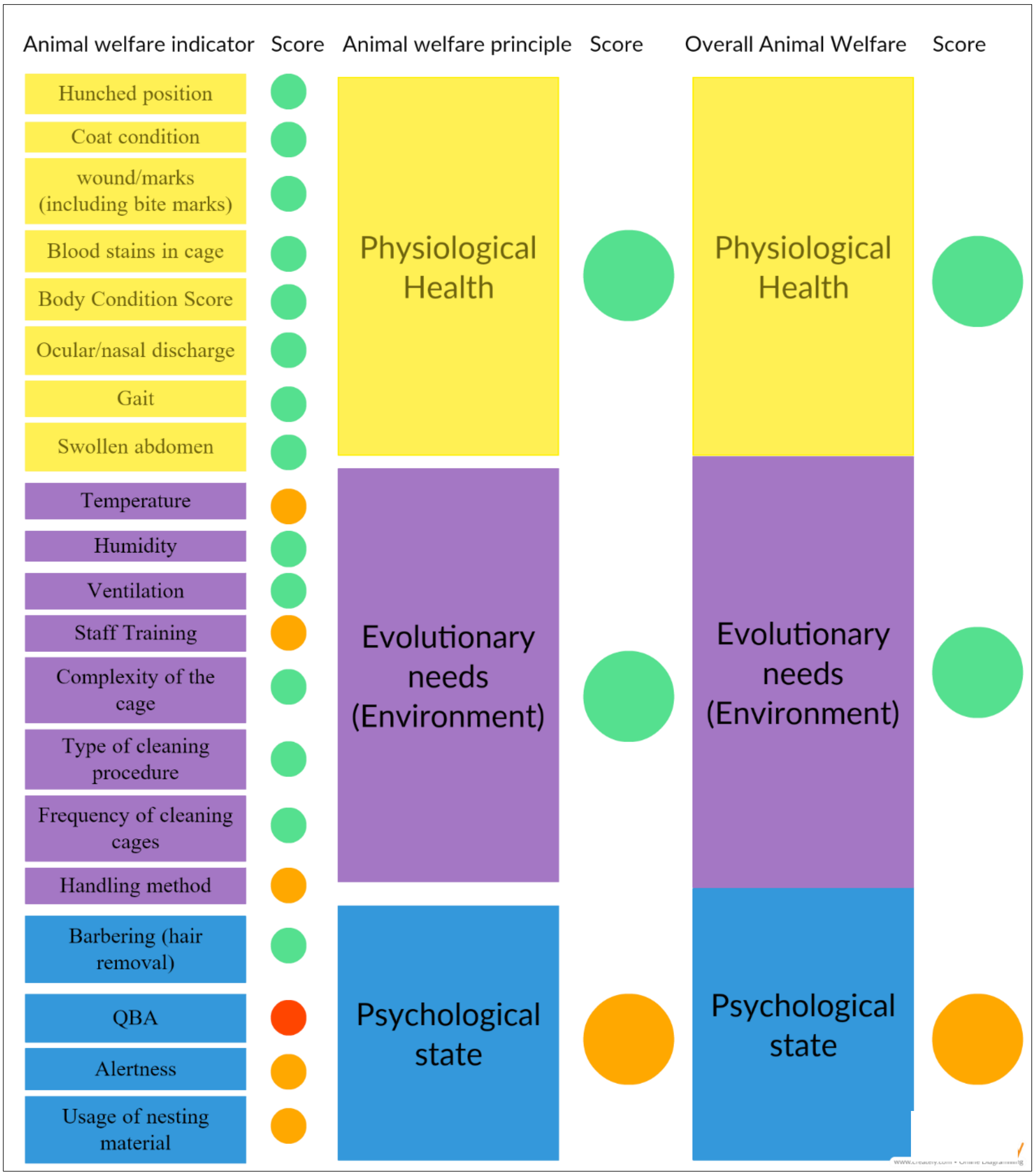


Figure 5.14. Summary of audit welfare assessment protocol results used at Dundee University.

5.5. Reliability assessment of the everyday welfare protocol

5.5.1. Methods

The everyday welfare protocol was carried out at a satellite facility at Newcastle University that houses an average of 3000 mice of which approximately 13% (~ 390 mice) were stock mice. The protocol was carried out by two senior animal technicians who were also NACWOs (Name Animal Care and Welfare Officers) with over ten years of experience each, and a veterinarian with over three years of experience of working with laboratory mice. The sample was determined using the information from previous studies about animal welfare protocols were 5% -10% of the animals are used for the assessment (Welfare Quality®, 2009a; Battini *et al.*, 2015). A total of 9 cages were assessed with 29 female BALB/c mice, who were an average of 6 months of age (Charles River Laboratories, UK). The temperature and humidity were controlled at room level, and ventilation was controlled at cage level with IVCs NexGen Cages (Allentown). The mice were housed in pairs or groups of 3-4 animals per cage. Animal cages included bedding and nesting material (Enviro-Dri® and Sizzle-Pet) as well as a clear Perspex tube (50mm diameter, 150mm length). Animals had *ad-libitum* access to water and food (Special Diet Services, RM3E diet) and sunflower seeds as an additional environmental enrichment. This protocol was used simultaneously by all assessors on four consecutive days, at the same time every day (mid-day). The assessors received information about the protocol one week in advance, to allow time to familiarise themselves with the indicators. On day one of the assessment, the assessors received training on the use of the QBA in other species and how this indicator should be used with mice. Following this short training period, the assessors were instructed not to discuss the scores with each other during the assessment at any time. They were told that the study aimed to assess the reliability of the protocol and not the welfare of the animals. The everyday welfare assessment was carried out following the three levels: in the facility/room indicators, the data was taken from the room, for the cage level assessment each cage was assessed individually after removing the cage from the rack and placed it in a working station. Following the cage level assessment, each individual within the cage was assessed. Both cage level and individual level assessments were performed by observation and involved with no handling.

5.5.2. Data analysis

The data was analysed using the same criteria as the audit welfare assessment protocol (section 5.4.2). The statistical analysis was carried out in Minitab (Minitab 17 Statistical Software, Minitab Inc., State College, PA, USA) and SPPS (Version 16.0. Chicago, SPSS Inc.).

5.5.3. Results

The average time to complete the daily assessment was 60 minutes per day with an average of 6 minutes per cage. Assessors scored humidity and ventilation (facility/room indicators) as 'green' (score of 1) with a 100% agreement. Two assessors scored temperature as 'amber' (score of 2), and one assessor as 'green' (score of 1) thus the level of agreement for this indicator was 66.6% (Table 5.14).

Descriptive Statistics				
	Minimum	Maximum	Mean	Std. Deviation
Temperature	1	2	1.92	.289
Humidity	1	1	1.00	.000
Ventilation	1	1	1.00	.000

Table 5.14. Descriptive statistics for the environment level indicators used in the everyday assessment protocol at Newcastle University. The scores using the traffic light system are represented by a numerical code (1= green, 2= amber and 3= red).

Four indicators were used to assess the welfare of the mice at cage level. The scores can be seen in Table 5.15 and the Attribute Agreement Analysis in Table 5.16. The agreement between the assessors for complexity of the cage and bloodstains on the cage was 100 and scored 'green' (score of 1). The agreement of alertness was lower at 88% with a score of 'green' (score of 1). The usage of nesting material had a low agreement (11%) between the three assessors with a score 'amber' as the mean of the score was 2.

Descriptive Statistics				
	Minimum	Maximum	Mean	Std. Deviation
Complexity of the cage	1	1	1.00	.000
Alertness	1	2	1.01	.096
Usage of nesting material	1	3	1.92	.549
Blood stains in cage	1	1	1.00	.000

Table 5.15. Descriptive statistics for the cage level indicators used in the everyday assessment protocol at Newcastle University. The scores using the traffic light system are represented with a numerical code (1= green, 2= amber and 3= red).

Indicator	Cages inspected	Cages matched	Per cent	95% CI
Complexity of the cage	9	9	100%	71.69, 100
Alertness	9	8	88.89%	51.75, 99.72
Use of nesting material	9	1	11.11%	0.28, 48.25
Blood stains in cage	9	9	100%	71.69, 100

Table 5.16. Attribute agreement analysis between assessors for the cage level indicators used in the everyday protocol.

From the QBA analysis, the two main dimensions explained 70% of the variability between the cages observed. Dimension 1 (52%) was labelled from playful/comfortable to fearful/tense. Dimension 2 (18%) was labelled from agitated/determined to in pain/bored (Figure 5.15).

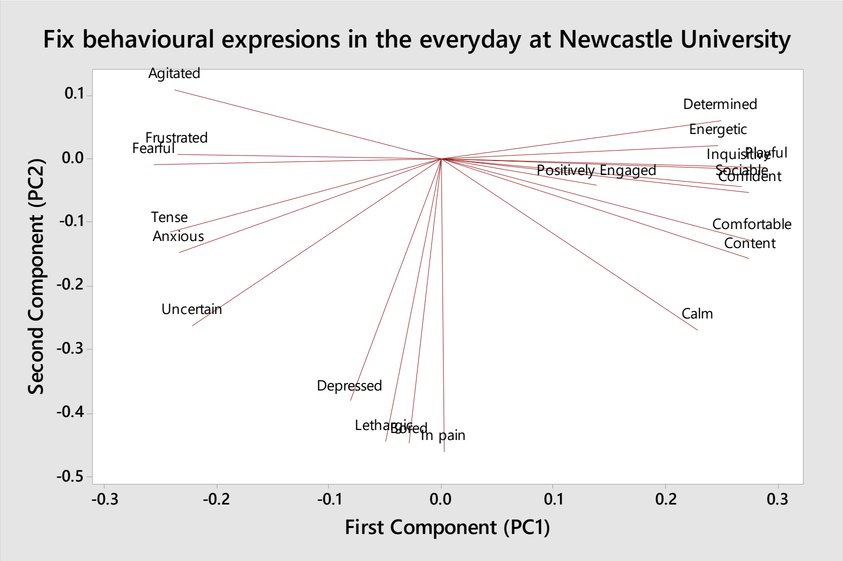


Figure 5.15. Distribution of the 20 fixed behavioural expressions used in Qualitative Behaviour Assessment as part of the everyday welfare assessment at Newcastle University. The x-axis represents dimension 1, and the y-axis represents dimension 2.

Kendall's coefficient of concordance showed a good level of agreement for all the sessions for both dimension 1 (Kendall's *W=0.86*) and dimension 2 (Kendall's *W=0.82*). The analysis of each session (i.e. 4 repeated assessments) separately showed that the agreement between the assessors was high for the two dimensions. For dimension 1: session 1 (*W=0.90*), session 2 (*W*=0.82), session 3 (*W*=0.91) and session 4 (*W*=0.90). For dimension 2: session 1 (*W=1.0*), session 2 (*W*=0.53), session 3 (*W*=1.0) and session 4 (*W*=1.0). The plot of the location of the cages in the dimensions (Figure 5.16) showed that the three assessors allocated the cages into different quadrants. Assessor one placed the cages in the left-upper quadrant described as agitated/determined and fearful tense, and so given a score of 'red'. Assessor two in the right-upper quadrant described as agitated/determined and playful/comfortable and so given a score of 'green' and assessor three in the lower (right and left quadrant) labelled as playful/comfortable for some cages and fearful tense for other cages and so scored 'red' for the cages in the left/low quadrant and 'green' for the cages in the positive/low quadrant.

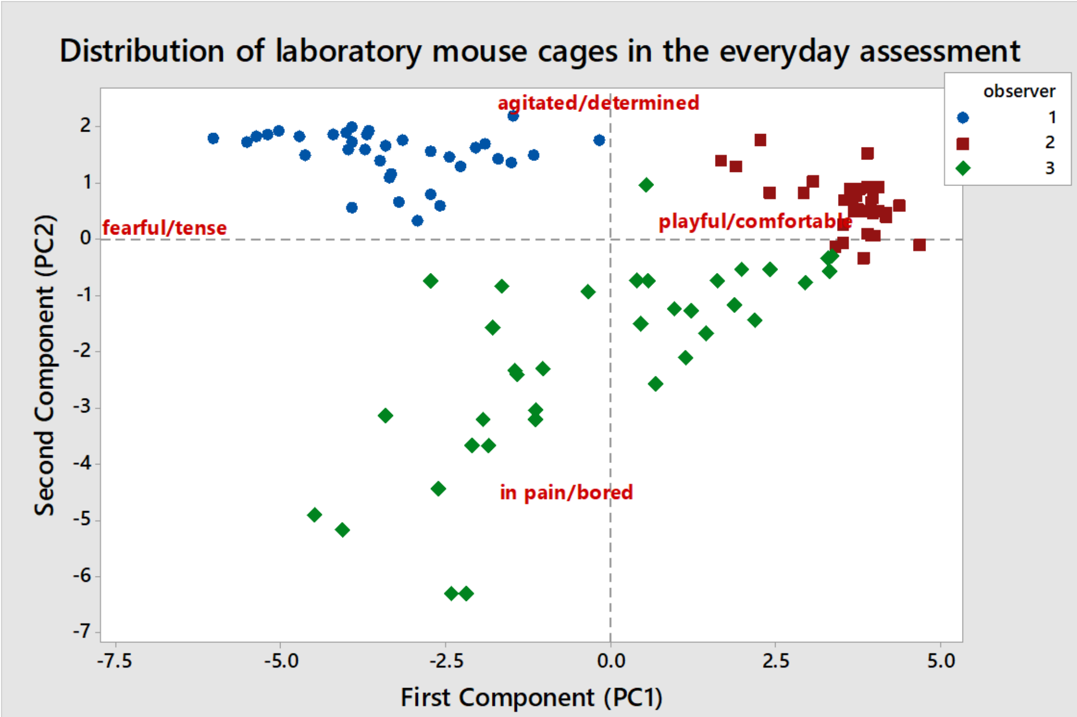


Figure 5.16. Distribution plot of QBA scores for the nine cages scored by three assessors, on four consecutive days as part of the everyday welfare protocol. Each point represents one cage, scored by one observer, on one day. The red legends are the behavioural expressions that represent each dimension.

The agreement for the animal level indicators was high between the observers (Table 5.17), with the mean scores of the indicators being 'green' (score of 1). Agreement between participants was over 95% for coat condition, ocular/nasal discharge, wounds/marks and swollen abdomen. Body Condition Score has an agreement of 93% followed by barbering with 90% agreement. The lowest agreement score was for hunched posture (80%) (Table 5.18).

Descriptive Statistics				
	Minimum	Maximum	Mean	Std. Deviation
Hunched posture	1	3	1.03	.188
Coat condition	1	2	1.00	.053
Gait	1	3	1.01	.158
Ocular /nasal discharge	1	2	1.01	.074
wounds/ marks	1	1	1.00	.000
Body Condition Score	1	2	1.02	.148
Barbering	1	2	1.13	.340
Swollen abdomen	1	1	1.00	.000

Table 5.17. Descriptive statistics for the animal level indicators used in the everyday welfare assessment at Newcastle University. The scores using the traffic light system are represented with a numerical code (1= green, 2= amber and 3= red).

Attribute Agreement Analysis				
Indicator	Mice inspected	Mice matched	Per cent	95% CI
Hunched posture	30	24	80.0%	61.43, 92.29
Coat condition	30	29	96.97%	82.78, 99.92
Gait	30	37	90%	73.47, 97.89
Ocular/nasal discharge	30	29	96.97%	82.78, 99.92
Wounds/marks (including fight wounds)	30	30	100%	90.50, 100.0
Body Condition Score	30	28	93.33%	77.93, 99.18
Barbering	30	27	90%	73.47, 97.89
Swollen Abdomen	30	30	100%	90.50, 100.0

Table 5.18. Attribute agreement analysis between assessors for the animal level indicators used in the everyday welfare assessment at Newcastle University.

The summary of the everyday welfare protocol assessment at Newcastle University is shown in Figure 5.17.

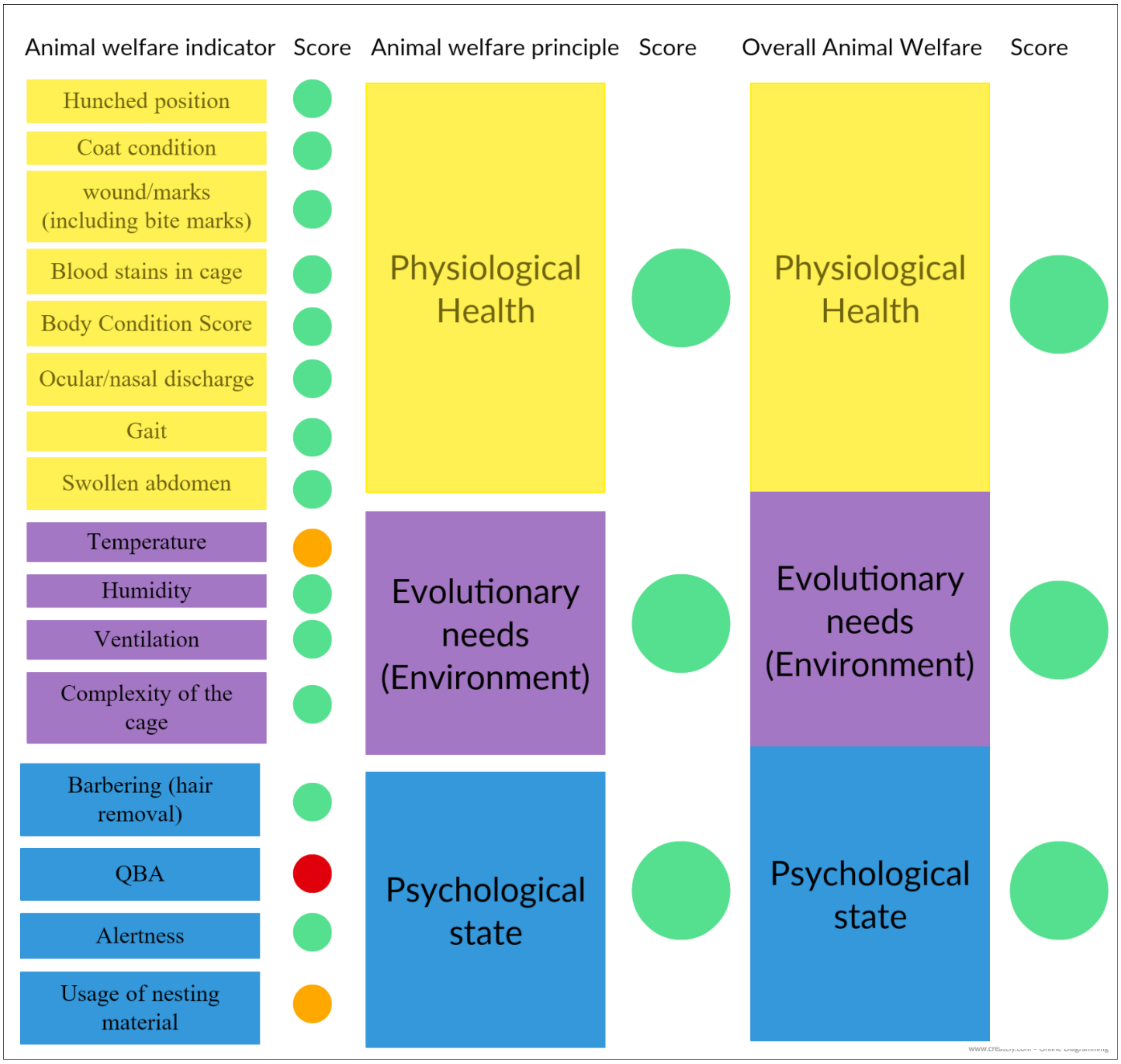


Figure 5.17. Summary of the everyday welfare assessment protocol at Newcastle University.

5.6. Validity assessment of the everyday protocol in experimental mice

5.6.1. Methods

The everyday welfare protocol was used for the assessment of experimental animals at Newcastle University that was enrolled in an unrelated study. Here, the aim was to assess the protocol's ability to identify welfare problems in experimental animals. All the animals enrolled in the experimental study were included in this welfare assessment; 16 male C57BL/6 mice (Charles River Laboratories, UK) at 9 weeks old were assessed using the everyday protocol. They were housed in pairs or groups of 4 in eight IVC NexGen Cages (Allentown) with bedding, nesting material and chew sticks in all the cages (Enviro-Dri® and Sizzle-Pet). The mice were handled using a clear Perspex home cage tube (50mm diameter, 150mm length) during the experiment. The home cages were cleaned once a week and animals were food restricted overnight during the experiment (Special Diet Services, RM3E diet) and *ad-libitum* water. Mice were maintained at 12:12 hour light/dark cycle (lights on at 02:00).

The purpose of this original study was to use the Orofacial Operant Pain Assay (OOPA) (Neubert *et al.*, 2005) to determine the efficacy of analgesic regimens in laboratory mice. Mice were trained to consume a palatable liquid reward when they poked their nose through a small opening, in doing so, their cheeks came into contact with a thermode set at 37ºC. Once the mice reached the pre-selected training threshold (over 800 licks per session), they progressed into the test phase involving the administration of a compound to lower the nociceptive threshold in one cheek that was shaved. Mice then received four treatments (control, saline, high and low analgesia) in a cross-over design. For this study, only the high doses of analgesia were used, and all the treatments were spread out evenly across the 4 days. Therefore, the treatments in this study were control (n=16), saline + compound (n=16) and high doses of analgesia + compound (either 5mg/kg morphine or 20 mg/kg meloxicam, n=16). Following treatment, they were placed in the OOPA again with the option to lick for the palatable reward.

The welfare assessment was made at the same time each day on day 3 of a training phase (Baseline: Pre-treatment) and days 11, 16, 21 and 24 (Test days: Post-treatment) (Figure 5.18). The animals were assessed by a veterinarian with over three years of experience in laboratory mice. The veterinarian was blinded to treatment groups but had general information about the aims of the study and the timeline.

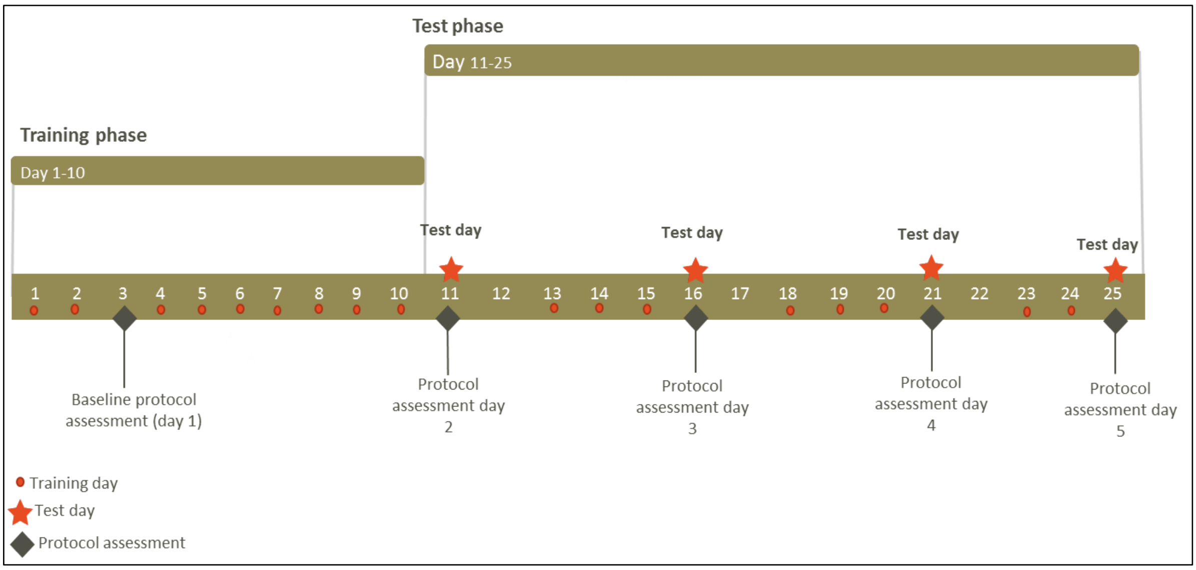


Figure 5.18. Timeline of the analgesic study. The experiment lasted for 25 days in total with 21 days of training and 4 test days. Animals were trained from day 1 to 10, 13 to 16, 18 to 20 and 23 to 24. The everyday protocol was used to make a baseline assessment on day three of the training phase, and post-treatment assessments on days 11, 16, 21 and 25 approximately 2 hours after the compound had been injected and the mice had been placed into the OOPA.

The study used the everyday protocol (16 indicators) with three additional individual indicators specific to the study: presence of superficial burns, cheek swelling and excessive grooming/licking/attention to the face. These 3 indicators were related to the humane endpoints described in the experimental study protocol. The study aimed to assess to what extent the indicators in this protocol were able to detect the changes in mouse welfare that were expected to occur as a result of the study. These changes included: [1] a decrease in welfare from baseline state to when the compound/saline was administered and [2] an improvement in welfare from when compound/saline was administered compared to compound/analgesia.

5.6.2. Data analysis

The scores of the indicators used were calculated through Descriptive statistics. The Attribute Agreement analysis was not included in this study as the assessment was made by only one person. As with the other data analysis, the data was divided to allow facility/room, cage and individual-level data to be analysed separately. Principal Component Analysis (PCA) was used for the analysis of the QBA data (See Chapter 4 for further details). One-way repeated measures ANOVA was performed to assess the difference in the indicator scores with the different treatments (within-subject factor = treatment) with Bonferroni *post hoc,* used for comparing the scores between the treatments and the baseline. The Mauchly's test was performed to test the sphericity assumption in the data for one-way repeated measures ANOVA analysis. The indicators were assessed independently as they measure different components of the animal's welfare. Statistical analysis was carried out in Minitab (Minitab 17 Statistical Software, Minitab Inc., State College, PA, USA) and SPPS (Version 16.0. Chicago, SPSS Inc.).

5.6.3 Results

The average time taken to complete the assessment on a given day was 38 minutes with an average of 4 minutes per cage. For the environmental indicators at the facility level (Table 5.19); Humidity and ventilation was scored as 'green' (score of 1) on all 5 days of assessment. The temperature was scored as 'amber' (score of 2) on all 5 days of assessment. The facility-level indicators were stable during the 5 days assessed, and no variation in the scores was observed.

Descriptive Statistics					
	Range	Minimum	Maximum	Mean	Std. Deviation
Temperature	0	2	2	2.00	.000
Humidity	1	1	2	1.10	.304
Ventilation	0	1	1	1.00	.000

Table 5.19. Descriptive statistics of the environmental indicators used in the everyday protocol for the welfare assessment of experimental animals. The scores using the traffic light system are represented with a numerical code (1= green, 2= amber and 3= red).

For the 4 cage level indicators, a small degree of variation was observed when assessed over the 5 days (Table 5.20). Overall, complexity of the cage, alertness and bloodstains on the cage were scored as 'green' (score of 1) across the 5 days of assessment. Use of nesting material was scored as 'amber' (score of 2).

	Descriptive Statistics				
Assessment day		Minimum	Maximum	Mean	Std. Deviation
day 1(baseline)	Complexity of the cage	1	1	1.0	0.0
	Alertness	1	2	1.1	0.4
	Use of nesting material	1	2	1.1	0.4
	Blood stains in cage	1	1	1.0	0.0
day 2	Complexity of the cage	1	1	1.0	0.0
	Alertness	1	1	1.0	0.0
	Use of nesting material	1	2	1.8	0.5
	Blood stains in cage	1	1	1.0	0.0
day 3	Complexity of the cage	1	1	1.0	0.0
	Alertness	1	1	1.0	0.0
	Use of nesting material	1	2	1.9	0.4
	Blood stains in cage	1	1	1.0	0.0
day 4	Complexity of the cage	1	1	1.0	0.0
	Alertness	1	1	1.0	0.0
	Use of nesting material	1	3	2.0	0.5
	Blood stains in cage	1	1	1.0	0.0
day 5	Complexity of the cage	1	1	1.0	0.0
	Alertness	1	1	1.0	0.0
	Use of nesting material	1	2	1.8	0.5
	Blood stains in cage	1	1	1.0	0.0
overall	Complexity of the cage	1	1	1.0	0.0
	Alertness	1	2	1.0	0.2
	Use of nesting material	1	3	1.7	0.5
	Blood stains in cage	1	1	1.0	0.0

Table 5.20. Descriptive statistics of the cage level indicators from the everyday protocol used for the assessment of experimental mice. The mean was used for the traffic light system scoring 1=green, 2=amber and 3=red.

Some of the animal level indicators used for this assessment showed a high degree of variation. Three indicators: hunched posture, wounds/marks (including bite wounds) and cheek swelling showed a higher variation in scores compared with other animal level indicators (Table 5.21). These indicators were scored as ‘green’ at baseline, however, on test days these indicators showed higher variations and high mean scores compared with the baseline day suggesting that they might be indicating welfare problems (Figure 5.19). For the traffic light system score, hunched posture and wounds/marks (including bite wounds) were score as ‘green’ and cheek swelling was score as ‘amber’. The remaining eight indicators did not show any relevant variation through the study when compared with the baseline levels.

Indicator (N=16)	Mean	Variance	Minimum	Maximum
Hunched posture	**1.50**	**0.254**	**1**	**2**
Coat condition	1.06	0.060	1	2
Gait	1.02	0.016	1	2
Ocular /nasal discharge	1.03	0.031	1	2
wounds/ marks (including bite wounds)	**1.13**	**0.111**	**1**	**2**
Body Condition Score	1.00	0.000	1	1
Barbering	1.00	0.000	1	1
Swollen abdomen	1.00	0.000	1	1
superficial burn	1.06	0.060	1	2
cheek swelling	**1.81**	**0.250**	**1**	**3**
excessive grooming/ licking/ attention to the face	1.08	0.073	1	2

Table 5.21. Descriptive statistics for animal level indicators used for the everyday welfare assessment of experimental mice. The indicators in bold are the three indicators which show a high level of variance on test days compared with the baseline in the assessment.

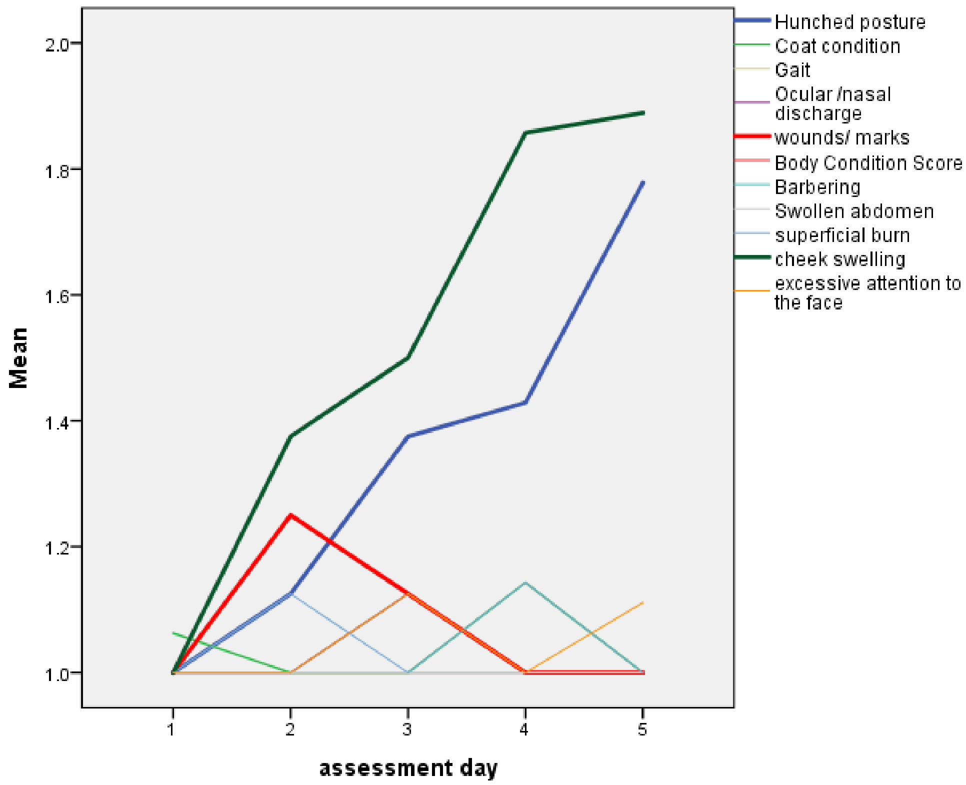


Figure 5.19. Variation of the mean of the animal-based indicators used in the experimental animals at Newcastle University. The x-axis represented the assessment days (from day one [baseline] to day 5). The y-axis represented the mean of the score given for each indicator. A colour represents each indicator in the graph.

For hunched posture, the assumption of sphericity was accepted (X^2 (2) = 0.29, p = 0.98). The results showed that hunched posture scores were significantly affected by the treatment given (F (2, 30) = 13.5, $p < 0.05$). Bonferroni *post hoc* showed that hunched posture scores were significantly higher when the mice received compound/saline compared with compound/analgesia ($p < 0.01$) and baseline ($p < 0.01$). However, there were not significantly different when comparing the scores of hunched posture between compound/analgesia and baseline ($p = 0.12$).

For wounds/marks (including bite wounds) the assumption of sphericity was accepted although Mauchly's test was not computed as there were only two scores levels for this indicator. The results showed that the wounds/mark scores were not significantly affected by the treatments ($F (2, 30) = 3.46$, $p = 0.44$).

For cheek swelling the assumption of sphericity was accepted ($X^2 (2) = 1.43$, $p = 0.46$). The results show that the scores of cheek swelling were significantly affected by the treatment ($F (2, 30) = 17.43$, $p < 0.01$). Bonferroni *post hoc* showed the scores of cheek swelling were significantly higher following compound/saline ($p < 0.01$) and compound/analgesia compared to baseline ($p = 0.019$). The scores of cheek swelling were not significantly different between compound/saline and compound/analgesia groups ($p = 0.05$).

The information of the changes along the days of assessment taking into account the treatment (compound analgesia and compound+saline) of these three indicators (hunched posture, wounds/marks and cheek swelling) can be seen in Figure 5.20. The indicators showed more variation and higher scores for the compound+saline group comparing with the compound+analgesia group.

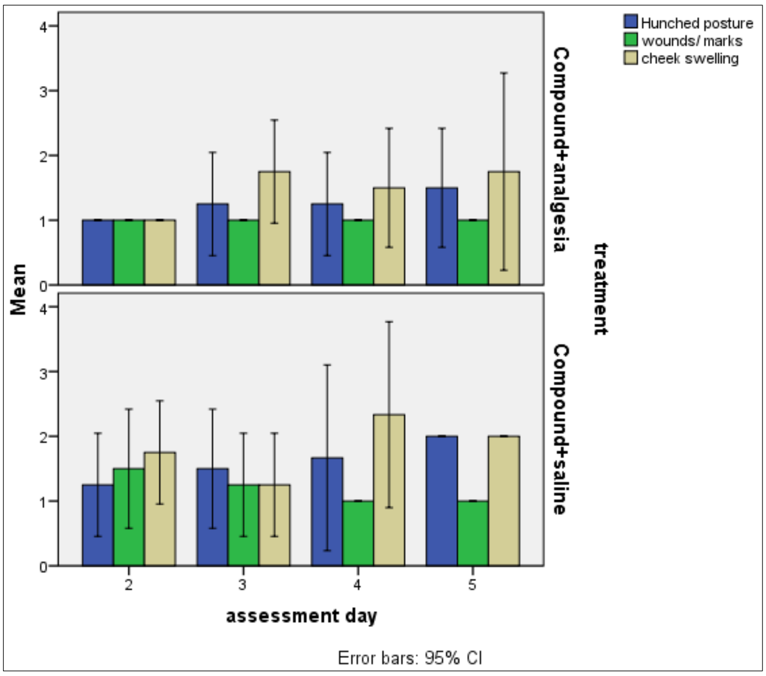


Figure 5.20. Changes of the three indicators in the days of assessment divided by the treatment used for the experimental animals at Newcastle University. The x-axis represented the assessment day. The y-axis represented the scores of the indicators given by the assessors. The error bars represent confidence interval at 95%

From the analysis of the QBA, the two main dimensions explained 47% of the variability between the cages observed. Dimension 1 (33%) was labelled from content/comfortable to lethargic/depressed. Dimension 2 (14%) was labelled from lethargic/depressed to agitated/determined (Figure 5.21).

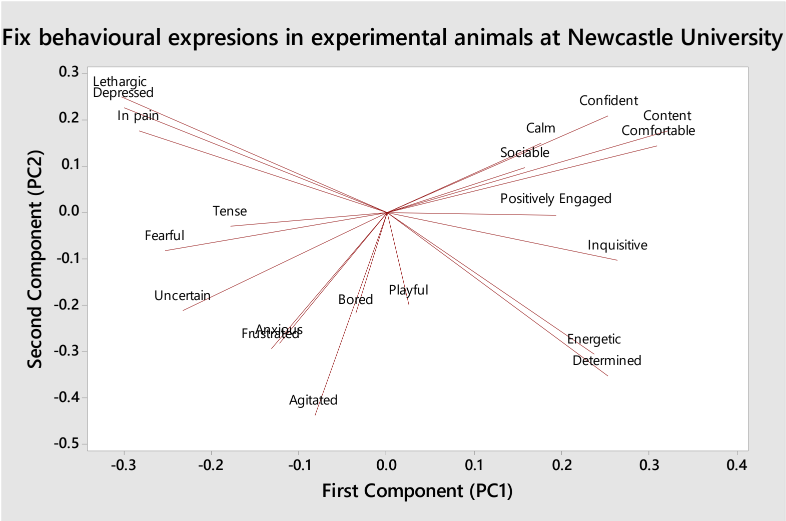


Figure 5.21. Distribution of the 20 fixed behavioural expressions used in the Qualitative Behavioural Assessment as part of the everyday welfare assessment in experimental animals at Newcastle University. The x-axis represented the first dimension or PC1, and the y-axis represented the second dimension of PC2.

The distribution of the cages in the plot is shown in Figure 5.22. The 8 cages were allocated evenly distributed on the plot in the two dimensions. The location of each cage was very similar across the 5 days of assessment.

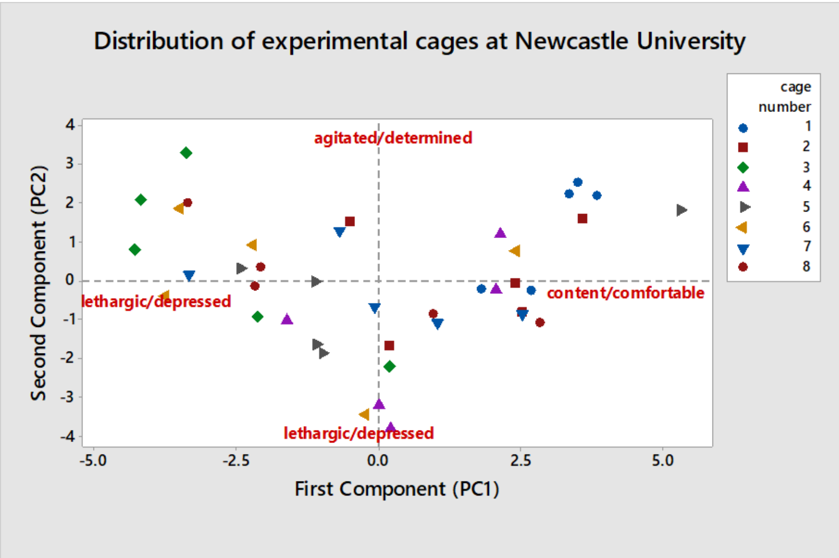
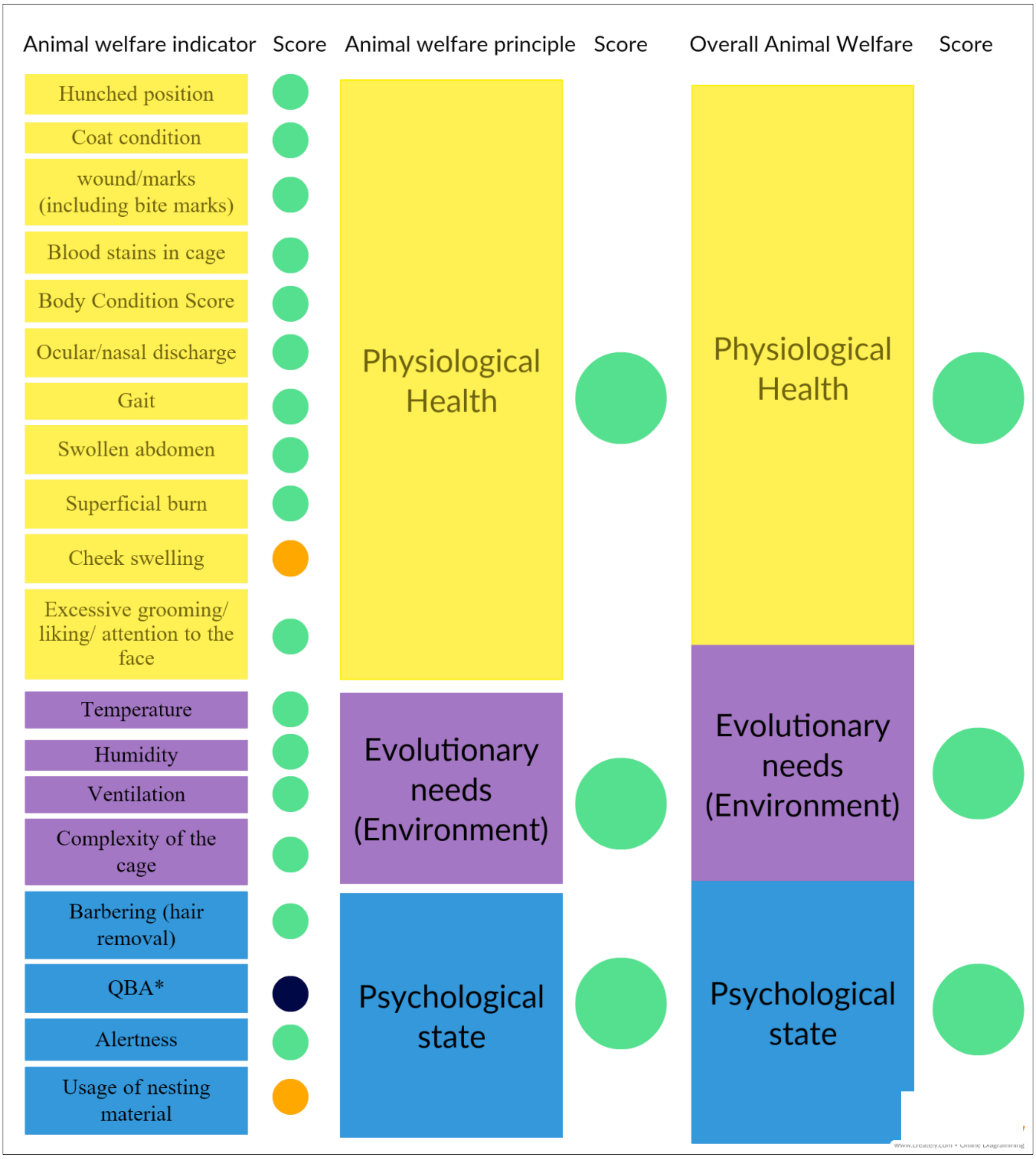


Figure 5.22. Plot distribution of the 8 cages assessed in the everyday welfare assessment in experimental animals at Newcastle University. The x-axis represented dimension 1 or PC1, and the y-axis represents the dimension 2 or PC2. Each colour represented each cage assessed in the 5 days.

The summary of the final scores of the everyday welfare assessment used in the experimental animals at Newcastle University is shown in Figure 5.23.



*QBA was not assessed in this protocol as the study was designed as cross-over with all treatments in different cages at the same time and QBA was used as a cage-level assessment.

Figure 5.23. Summary of the everyday welfare protocol used in experimental animals at Newcastle University

5.7. Discussion

This study aimed to assess the practicability and reliability of the everyday and audit welfare protocols for assessing the welfare of stock mice and the validity of the everyday protocol for assessing the welfare of mice enrolled on a scientific study. The practicability was assessed by measuring the time spent, the resources needed and the ease of scoring. The reliability assessment was conducted by inter-observer agreement analysis using Attribute agreement analysis (Jegstrup *et al.*, 2003), including the assessment of Qualitative Behavioural Assessment (QBA) as a qualitative indicator.

The overall time spending in the assessment for the audit protocol was 68 minutes at Newcastle University and 53 minutes at Dundee University. For the every-day assessment, the total time spent was 60 minutes at Newcastle University and 30 minutes for the experimental animals at Newcastle University. The audit protocol was built based on an audit (4 hours) assessment in a laboratory animal's facility. The time taken for both of the facilities was less than 2 hours, and the assessment included all the 20 indicators proposed with the necessary number of animals considering the proportion of stock animals present in both facilities. However, the number of animals selected for this study (from 5% to 10%) is the minimum number. Therefore, an increase in the number of animals assesses (up to 25%) is it recommended to be aligned with other protocol such as the Welfare Quality Program.

The mean time spent assessing an individual cage was as low as only 5 minutes in both facilities. When using the audit protocol, the time spent was an average of 6 minutes per cage at Newcastle University and 5 minutes per cage at Dundee University. For the everyday protocol, the time spent was 6 minutes per cage for stock mice and 4 minutes per cage for mice enrolled in a study at Newcastle University. The average 5 minutes per cage spent in the everyday protocol for both stock and experimental animals may not be practical in laboratory animal context when assessing all mice in the laboratory facility. The high number of animals which are usually present makes unpractical to spend 5 minutes per cage in a daily assessment, in this case, a sample of the population can be used taking into account the total amount of animals as it is done in other protocols (Leach and Main, 2008). Even if the average time was about the same for both protocols, the everyday protocol used in experimental animals was less time-consuming. One possible explanation might be related to the total number of animals assessed in the experimental study as the assessment was made on 16 mice, which is less than the number assessed with the audit assessment. The time spent in the protocol could be affected by the number of animals because of the individual level indicators which were assessed for each mouse (Hawkins *et al.*, 2011).

In order to ensure the protocol can be executed in a suitable length of time, logical ordering of the indicators on score sheets was essential. Contrasting with other protocols developed in other studies, which have divided the indicators by whether they represent resources (i.e. inputs) or the responses of animals to these resources (i.e. outputs) (Leach *et al.*, 2006). In this study, the indicators were organised by levels of assessment from the facility, down to the single animal. This systematic organisation provides a logical and flowing protocol which assess the welfare from the facilities as a whole, down to individual mice located within the cage. This helps to facilitate the assessment by giving the assessor a pre-determined order in which to conduct the assessment and potentially improving the practicability as both resources and animal outcomes where measured. This offers the advantage of including all the components of animal welfare: the inputs that are given to the animals in terms of environment and housing and the outputs about how animals respond to them as has been done in other protocols (Leach and Main, 2008).

Another critical factor in the practicability of the protocol is the easiness of recording scores during the assessment. In this study, score sheets were used for the audit (Figure 5.2-5.5) and for the everyday protocol (Figure 5.6-5.7) which are considered to be efficient for recording data as they are easy, practical and provided reliable information (Hawkins *et al.*, 2011). The audit and everyday protocols used a scale for each indicator which was determined by recommendations found in the literature and previous studies (see appendix D and F). This scoring system should improve the consistency and objectivity of the protocol by making the assessment practical and reliable.

The results of the reliability analysis showed that participants tended to score indicators in the same manner, with an overall agreement of over 80% with 17 out of 20 indicators in the audit assessment and 13 out of 16 indicators in the everyday assessment, demonstrating a good level of consistency in their scoring. In studies using this form of reliability analysis, agreement over 80% is considered to be good (Landis and Koch, 1977; Van Swieten *et al.*, 1988). Analysing these findings in detail, an inter-observer agreement for facility/room indicators in both protocols was 100%. The agreement between participants for the cage level assessment indicators was over 80% for both protocols; 82.5% at Newcastle University and 87.5% at Dundee University for the audit protocol, and 84.5% for the everyday protocol at Newcastle University. For the animal level indicators, the agreement was 100% for the audit protocol at Newcastle and Dundee Universities and 93.3% for the everyday protocol at Newcastle University. The agreement of cage level indicators at Newcastle University was 85% and 97% for animal level indicators in the audit and everyday assessments respectively. The high level of consistency of the assessors when using the protocols can be related to their experience as well as the easiness for scoring the indicators. All the assessors who participated in the study had experience in working with laboratory mice as technicians, Named Animal Welfare Officers and veterinarians. This experience is reflected in the knowledge about laboratory mouse welfare including the means of assessment. In addition, the indicators used in both protocols are well-known indicators (excluding QBA) which could increase the consistency as the assessors were already familiar with most of the indicators. This previous experience along with the knowledge of the indicators is paramount for improving the reliability of the laboratory mouse welfare protocols as the better the consistency and agreement between the assessor, the better welfare the animals are likely to have, because it will increase the probability of noticing welfare problems with more accuracy (Hawkins *et al.*, 2011).

The everyday protocol was conducted on mice enrolled in a study to assess its validity in a scenario where mice were undergoing scientific procedures. Using this protocol, the assessor was able to recognise animals with compromised welfare, particularly through the assessment of three specific indicators: hunched posture, wounds/marks (including bite wounds) and swollen cheeks. The other indicators from this protocol showed a small variation including coat condition, gait, ocular/nasal discharge/superficial burn and excessive grooming/licking/attention to the face.

The small variation in the other indicators could be related to the expected adverse effects of this study. The protocols involving experimental animals require specific information about possible adverse effects, how they will be measured and what action should be taken if they are observed (Hollands, 1986b). Specific indicators including those that represent the ‘possible worst-case scenario’ are required to be included in the document, and they usually are the most valuable indicators for the project because they are acting as a guide for defining when specific and most likely urgent measures need to be taken. In this study, swollen cheeks, excessive grooming/licking/attention to the face and superficial burn were part of this group of indicators which were specially designed for this study. From those three indicators, superficial burn was described as a possible adverse effect of the protocol. If this indicator were to be present, it would likely represent an equipment malfunction whereby a temperature of >45ºC resulted in tissue damage (Caterina *et al.*, 2000). There were no changes in this indicator during the assessment which is considered not like an unsuccessful detection of the indicator but success in avoiding suffering during the experiment. The other two project-specific indicators (swollen cheeks, excessive grooming/licking/attention) were considered not as severe as superficial burn, and it was expected that they would occur to a certain level following administration of the sensitising compound used in the study.

Furthermore, cheek swelling was observed in some animals and so in this study is useful for identifying animals to keep a close watch over for detecting potential welfare problems. The results of these specific indicators evidence their importance when they are included in the protocols besides the general indicators (Hawkins *et al.*, 2011). Other studies have also included specific indicators into laboratory mice protocols such as genetically modified animal models (Rogers *et al.*, 1997; van der Meer *et al.*, 2001b; Wells *et al.*, 2006b) and mice with abdominal tumours (Paster *et al.*, 2009).

The three indicators which shown high variation in the scores (hunched posture, cheek swelling and marks/wounds) suggested that they might be robust enough for use in this type of OOPA study as these indicators changed when comparing with the baseline levels. Cheek swelling and wounds/marks were indicators related to the physical health of the animal which was able to identify changes in animal's behaviour and physical health which might evidence underlying pain. An inflammation process may have taken place following injection of the compound into the cheek, along with the presentation of wounds in the injection site (Lariviere and Melzack, 1996; Neubert *et al.*, 2005). Hunched posture, on the other hand, is not a specific indicator related to the study, but a general indicator related to mice physical health. Hunched posture is considered a critical measure of laboratory mouse welfare as is a robust indicator of pain and distress (Baumans *et al.*, 1994; Hawkins, 2002; Paster *et al.*, 2009). However, the fact that not all the 19 indicators used as part of the everyday protocol were able to detect possible welfare problem in this study does not mean that the everyday protocol is unimportant for the assessment of welfare in experimental animals. Furthermore, as this protocol was created from a wide range of indicators which measure all the principles of laboratory animal welfare, it can be used for a wide set of circumstances including daily monitoring and experimental conditions, alongside specific procedure linked indicators.

Qualitative Behavioural Assessment at cage level showed variable levels of agreement between assessors for each of the two protocols and in the different locations for the stock mice. For the audit protocol used at Newcastle University, the results show that the assessors agreed in the use of the two dimensions (w>0.8) when assessing the cages. Closer inspection of the distribution of the ten cages showed that the assessors located the cages in different quadrants which described them as playful/determined for assessor one, confident/positive engaged for assessor two and uncertain/tense for assessor three. At Dundee University, however, the agreement between the assessors in the usage of the two dimensions was moderate for dimension one (w=0.5) and low for dimension two (w=0.0). The distribution of the cages in the plot shows that the two assessors allocated the cages in two different sides of the plot, assessor one described the cages as agitated/uncertain whereas assessor two describe the cages as social/positively engaged. This low agreement between the participants might be related to their different backgrounds (veterinarian and senior technician).

For the everyday protocol at Newcastle University, the assessors had a good agreement in the usage of both dimensions when assessing the cages (w>0.8). The distribution of the videos, however, is similar to the audit protocol where the assessors located the cages in different quadrants. Assessor one described the cages as agitated/determined, assessor two as pain/bored and assessor three from fearful/tense to playful/comfortable. The behavioural expressions used in this study provide information about the emotional states of the animals by describing an emotional component that is underline the behaviours observed. Similar to other studies, this information is used for the assessment of the psychological state of the animal based on the observer’s interpretation of the animals’ demeanour (Stevenson-Hinde and Zunz, 1978; Stevenson-Hinde *et al.*, 1980; Wemelsfelder, 1997). Taking the information from these three studies involving QBA as a means of assessing the welfare of mice regarding the psychological component, it can be seen that the reliability of the indicator when used cage-side was variable. The location of the cages in the plot provides information about the welfare of the animal. Each dimension (PC1 and PC2) was labelled with the most representative emotional expressions taken from the PCA analysis thus the location of the cage in those dimensions is correlated with the label of the dimension. In this study, there was no agreement between the participants on the location of the cages in the two dimensions. One possible explanation for this variability could be related to the application of this technique at the cage level, i.e. the assessment was carried out on the cage as a whole, not the individual animals. QBA has routinely been applied at the cage/pen level in other species, such as cattle, poultry and pigs (Welfare Quality®, 2009c) and donkeys (Battini *et al.*, 2015). Compared to the assessment of QBA at the individual level as was done to validate this technique in Chapters 3 and 4. This variability from the individual to the group assessment could be related to the difficulty in the assessment of welfare indicators at group level (Fraser, 2003). QBA can be assessed either at individual or group level. However, the group level assessment involved the integration of the information observed by the assessors, including how the animals interact and how they influence each other. This integration of the information can be challenging, especially when using this indicator for the first time (Wemelsfelder, 2007). These results showed that the practicability of QBA used as part of a cage-side protocol is not straightforward. Although assessors agreed in the usage of the two dimensions (Kendal's coefficient of concordance results), they failed to place the cage in similar quadrants (plot for QBA using PCA). These results could be explained by the lack of calibration in the use of the Visual Analogue Scales (Fleming *et al.*, 2015; Grosso *et al.*, 2016). Another possible explanation for this result is the use of fixed terms. The assessors can interpret the meaning of the terms in different ways which can influence the scoring, even if the meaning of the terms is discussed in advance (Wemelsfelder, 2007; Minero *et al.*, 2016). In order to improve this calibration between the assessors, additional training including practice with the terms and the use of the visual analogue scales would be essential. This training could involve practice with videos or on-site for scoring the fixed behavioural terms using the VAS, comparing and discussing the results obtained and make an agreement within the group in how the VAS should be used for the aim of the assessment (Minero *et al.*, 2016).

QBA was also part of an everyday protocol for the welfare assessment of mice enrolled in a study as a cage-level indicator for the assessment of psychological states. QBA relies on what the assessor perceives in terms of the emotional expressions of the animal as an individual when interacting with others or with the environment as discussed in Chapter 4. The results of QBA in this study are limited as the experiment involved a cross-over design where all the mice received all the treatments on different days. As such, each mouse within a cage had received different treatments when QBA was conducted, therefore the results of QBA are not reliable in this study as the interactions between the different treatments within the cages affect the QBA results.

One of the most challenging steps in building up the protocol was to identify indicators that more directly represent the psychological aspect of welfare. From the list of the final indicators (59), only four were included in this category (use of nesting material, alertness, barbering and QBA). These indicators were included in the final list as they fulfil the requirements from the assessment of validity and reliability from Delphi and practicability assessment (section 5.3.2). The psychological state as a component of welfare in animals has been increasingly included in assessments of welfare (Russell, 1980; Russell, 2003; Scherer, 2005; Thagard and Aubie, 2008). However, this inclusion brings difficulties as its assessment is not straightforward as the methods used are not often fully validated, therefore the proportion of indicators available for this welfare principle are still very limited.

An animal welfare protocol aims to recognise when welfare is compromised and to highlight when welfare is good (Mellor and Reid, 1994; Hawkins *et al.*, 2011). This study shows a good level of practicability and reliability for both audit and everyday protocols and the indicators that comprise them. They have shown to be practical regarding the resources and time needed for assessing stock and experimental animals. The inter-observer agreement showed good levels of reliability as the assessors agreed in the scores given across most indicators when assessing stock animals. It is important to highlight that additional indicators have to be included for experimental animals, taking into account the nature of the study to better assess these animals welfare (Mertens and Rülicke, 1999; Paster *et al.*, 2009). As shown in the validity of the everyday protocol study, specific indicators, e.g. in this case cheek swelling, were useful for detecting animal welfare compromises for mice enrolled in this particular study.

This study is the first step towards the integration of QBA as a cage-side indicator of laboratory mouse welfare. Further studies are needed for the validity of QBA as animal welfare indicator at the cage-side as the results showed that assessors need to be better calibrated in the use of the terms, scales and their interpretation, in the same way that has been shown in other studies (Rutherford *et al.*, 2012; Minero *et al.*, 2016). The study aimed to assess the practicability and reliability of the audit and everyday protocols as well as the validity of the everyday protocol as a welfare assessment tool for mice enrolled in an experimental study. Based on the results obtained, it can be concluded that the audit and everyday protocols used for this study are reliable and practical for the assessment of stock and experimental animals. However, as the results of the everyday protocol in experimental animals have shown, additional indicators should be added, taking into account, the specific welfare problems that different experimental studies can have (Flecknell and Roughan, 2004; Paster *et al.*, 2009).

Chapter 6. General discussion

6.1. Aims

The main objectives of this thesis were [1] to develop a protocol that includes quantitative and qualitative mouse welfare indicators through scientific inquiry and expert opinion that will allow those who care for (i.e. technicians & veterinarians), regulate (i.e. HO inspectors) and use (i.e. scientists) mice in research to assess their welfare practically and reliably. [2] To apply the developed protocol to assess welfare in laboratory mouse facilities, in order to test its practicability and reliability. [3] To investigate whether the novel welfare assessment technique, Qualitative Behavioural Assessment (QBA), may offer a useful and practical means of assessing laboratory mouse welfare.

6.2. Overview of methods

The development of the laboratory mouse welfare protocol, including quantitative and qualitative indicators, involved three phases: [1] Delphi consultation process, [2] validation of Qualitative Behavioural Assessment as a qualitative indicator of laboratory mouse welfare, and [3] the assessment of practicability and reliability of the developed protocol.

Phase one aimed to collect information from experts about the most valid, reliable and practical indicators for the welfare assessment of laboratory mice (Chapter 2). This phase involved a scoping meeting with an expert panel, where preliminary information about indicators that are currently used was sought, and gap analysis in relation to laboratory mouse welfare assessment was carried out. Following this, a pilot survey was carried out to assess the structure and practicability of the survey questionnaires used in the Delphi consultation process. Finally, two rounds of Delphi survey were carried out to reach a consensus on the most important indicators for assessing mouse welfare, based on two hypothetical scenarios; a facility audit and an everyday welfare assessment.

Phase two involved two studies assessing the validity of Qualitative Behavioural Assessment (QBA) as a means of assessing laboratory mouse welfare. The first study assessed the use of QBA through Free Choice Profiling (FCP). The assessment of QBA as a measure of mouse emotional states through behavioural expressions was assessed. The assessment included inter-observer reliability as well as determining if assessors could differentiate between video recordings of mice in varying scenarios (i.e. behavioural tests and during the post-operative phase) associated with different welfare states. The second study for the validation of QBA included the assessment of inter- and intra-reliability of a fixed list of behavioural expressions used in the assessment of laboratory mice from existing video footage of mice in potentially different welfare states.

Phase three of the protocol development process included the practicability assessment of the 59 indicators obtained from the Delphi consultation, by measuring the duration of time taken to use them in an assessment, the resources needed and the easiness for scoring. The final construction of the protocols took into consideration both the results from the Delphi study and the practicability assessment. Following this, a reliability assessment was carried out at Newcastle and Dundee Universities for the audit welfare protocol and at Newcastle University for the everyday protocol. Finally, the validity of the everyday protocol for assessing welfare in experimental animals was assessed at Newcastle University.

6.3. Key findings and discussion.

The process used in this thesis for the construction of an audit welfare and an everyday welfare assessment protocol followed a structured and coordinated strategy for selecting the indicators. Each of the selected indicators went through an assessment that included expert opinion from the Delphi consultation, information found in the literature about the indicators and the practicability assessment. This selection methodology was aligned with recommendations for the construction of laboratory animal welfare protocols (Hawkins *et al.*, 2011).

The Delphi study (Chapter 2) provided a final list of 59 important indicators in ranked order considering validity, practicability and reliability. An agreement about the ranked order of the indicators was reached in the second round of the consultation process, providing a final list of indicators and a separate list of the top ten indicators selected by the experts in both scenarios. The first two indicators on the list (*hunched posture and coat condition*) had 85% agreement between participants and were selected as the first two indicators in both scenarios (audit and every day). The final top ten list for the audit protocol shows that 8 out of 10 indicators were animal-based demonstrating the perceived high credibility of this type of indicator between the experts. These results are consistent with protocols in other species such as the Welfare Quality program in pigs, where a total of 30 out of 33 indicators, were animal-based (Quality, 2009), and in donkey welfare assessment, where 20 out of 24 welfare indicators were animal-based (Battini *et al.*, 2015). From this final list of top ten indicators, three were physiological (*Coat condition, Body condition Score, Weight change*), followed by indicators including behaviour and social interaction and the environment (*hunched posture, Exhibition of abnormal and normal behaviour, Bite/wound marks, and Usage of nesting material*). Only two indicators were resource-based (*Mortality rate and Staff training*). The top ten indicators for everyday welfare assessment were very similar to the indicators for the audit protocol. However, they differ in four indicators (*room temperature, wounds [excluding bite wounds], pups outside the nest,* and *alertness)* which were included in the top ten list for the everyday welfare assessment but not in the audit welfare assessment. These indicators replaced *body condition score, weight change, mortality rate* and *staff training* which were present in the audit top ten list.

The psychological state of an animal is one of the most important elements of animal welfare (Boissy *et al.*, 2007a), but it is considered to be more challenging to assess as there is no means of directly measuring it (Boissy *et al.*, 2007a; Bekoff, 2009; Flecknell *et al.*, 2011). Although behavioural indicators are often used as an indirect means of assessing psychological/emotional states (Dawkins, 2006), as a representation of a possible underlying emotion state (Dawkins *et al.*, 1993; Flecknell *et al.*, 2011). In an attempt to develop a protocol containing measures that may offer a more direct means of assessing the psychological state, the final protocol included four specific indicators: alertness, use of nesting material, barbering and QBA. Alertness, use of nesting material and barbering are related to positive and negative emotional states in laboratory animals and have previously been used in the assessment of laboratory mouse welfare (Lloyd and Wolfensohn, 1999; Garner *et al.*, 2004a; Gaskill *et al.*, 2009). Use of nesting material, for example, is a species-specific welfare indicator which is considered a highly motivated behaviour in mice associated with positive states (Roper, 1973; Hess *et al.*, 2008; Gaskill *et al.*, 2013b). Barbering is a behaviour that is associated with negative psychological states in laboratory mice related either to excessive dominance (Long, 1972; Strozik and Festing, 1981) or to compulsive, repetitive behaviours (Garner *et al.*, 2003; Garner *et al.*, 2004a; Garner *et al.*, 2004b).

QBA is considered a new indicator used recently for the assessment of emotional states in farm animals, and it has been included as a positive welfare indicator in other protocols e.g. pigs (Welfare Quality®, 2009a) and donkeys (Battini *et al.*, 2015). This technique is based on the subjective assessment of an animal's emotional states by integrating behaviour, interactions with other animals and their response to the environment (Wemelsfelder, 2007). This integration makes use of behavioural expressions for animals which describe *how* the animal is performing their activities, rather than *what* the animal is doing (Wemelsfelder, 1997; Wemelsfelder, 2007). For example, a laboratory mouse can be exploring around the home cage but either very anxiously or calmly. The interpretation of the manner that the animal is performing activities factors in both the energy level and valence (the positive and negative character of emotion or some of its aspects (Joffily and Coricelli (2013)) can describe the animal's psychological state (Mendl *et al.*, 2010). QBA is considered a more direct means of assessing psychological state as the behavioural expressions more directly describe the underlying emotions that are behind the behaviour (Wemelsfelder, 2007). In contrast to behavioural indicators (e.g. duration of explorative behaviour), QBA assesses the animals, their environment and the responses to the environment as a whole, involving the assessors in the scientific interpretation of the whole assessment (Wemelsfelder *et al.*, 2012).

The assessment of QBA using the Free Choice Profiling technique, which is a methodology that uses qualitative descriptors for describing animal's emotions (Chapter 3) shows a high level of inter-observer agreement in the qualitative assessment of behavioural expressions of the video sequences shown to the assessors. Two main dimensions which represented 74% of the variation among the videos were found. Assessors with different levels of experience of working with laboratory mice ('experienced' and 'inexperienced') developed a similar vocabulary for describing laboratory mouse emotional states which were correlated with these main two dimensions. The dimension in QBA studies is generally described in terms of energy levels and valence similar to other studies where the assessment of emotion is described in these two components (Mendl *et al.*, 2010). Dimension one is usually related to valence with dimension two more usually describing energy levels. Assessors were also able to differentiate between videos of mice that had undergone surgery, compared with videos of mice that had not undergone surgery. The post-surgery videos were generally assessed as more 'lethargic' and 'in pain' compared with the videos sequences of non-surgery mice which were assessed as more 'calm' and relaxed'. The correlation between the two groups with dimension one shows moderate levels of agreement, labelled "inquisitive/alert" to "in pain/depressed" for experienced observers (animal technicians and veterinarians) and from "agitated/alert" to "in pain/tired" for inexperienced observers (researchers and MSc students). For dimension two, the agreement was high labelled from "calm/relaxed" to "stressed/agitated" for experienced observers and from "relaxed/calm" to "panicked/distressed" for inexperienced observers.

The next validation phase of the QBA involved the assessment of reliability using a fixed-term list, which is a specific 20 terms behavioural expression list (Chapter 4). QBA was able to identify two main dimensions of mouse expressive demeanour from the 20 interaction test mouse videos shown to participants. The results showed that the 19 assessors had a high level of inter-observer agreement on the first dimension (PC1: confident/positively engaged to fearful/anxious). This high level of agreement in the first dimension was also shown when a subset of assessors (9) repeated the scoring process after 1-week (intra-observer reliability). Context blind assessors were also able to differentiate between videos which showed mice previously handled using different methods. Those that had been handled by the tail (considered anxiety-inducing [(Hurst and West, 2010; Gouveia and Hurst, 2013]) were located on the negative valence dimension, whereas the videos showing mice that had been handled by tube (considered to be less or not anxiety-inducing [(Hurst and West, 2010; Gouveia and Hurst, 2013]) were located on the positive valence dimension in both sessions. In addition, the videos of mice handled by the tail displayed more anxious/fearful demeanour than those handled using a tube-based on Figure 4.4. The intra-observer reliability results shown that the 9 assessors reliably scored the videos in the same way for PC1 between the two sessions. This is in line with previous research which suggests that mice that are tail handled are more anxious (Hurst and West, 2010; Gouveia and Hurst, 2013). The current results indicate that assessors could use QBA to differentiate anxious mice from less anxious mice. In contrast to the reliable results of dimension 1, dimension two (PC2: bored/calm to tense/agitated) showed very low inter- or intra-observer reliability. The lack of reliability for this dimension might be related to the difficulty in the assessment of energy levels in laboratory mice. As laboratory mice are animals which are very active and in constant motion, having experience of normal behaviour in mice could be considered to be crucial for the assessment of welfare (Staats, 1966; Heston, 1967; Boursot *et al.*, 1993). Another important result of the validation of QBA was the location of the animals observed into a triangle shape in the FCP study in chapter 3 (figure 3.7). These results showed the distribution of the videos observed by the participants distributed across three dimensions opposite to other studies made in different species. Similar studies in other species have shown the distribution of the videos observed distributed across all quadrants, representing all the four dimensions (good/calm, good/energetic, bad/calm and bad/energetic. However, in this study the quadrant good calm was not represented by any of the videos observed by the participants. This results can also be related to the differences in mouse behaviour compared to the other species. Differences such as high levels of nocturnal activity, increase activity level and reduce live spam might be related to the difficulty in identify this quadrant.

The final stage for the construction of the protocols (audit and every day) involved the practicability assessment of the 59 indicators obtained from the Delphi consultation to define which would be included in these final protocols. These final protocols were constructed based on the definition of laboratory mouse welfare, defined as “the animal’s ability to preserve their biological functions (physiological parameters), their psychological state (mental well-being) and the ability to maintain their evolutionary needs, taking into account the environment where the animal is being kept” (Dawkins, 1990; Broom, 1996; Fraser and Broom, 1997; Dawkins, 1998; Hubrecht, 2014). The audit and everyday protocols comprised of 20 and 17 indicators respectively. These indicators were divided into the three main welfare principles represented in the above definition. Eight indicators represented biological functioning in both protocols, psychological state was represented by 4 indicators in both protocols, and 8 indicators represented the environment in the audit protocol and 5 indicators in the everyday protocol. The indicators were organised based on three levels of assessment (facility/room, cage and individual) to facilitate an efficient scoring process. The reliability of both protocols was shown to be over 80% overall. For the facility/room indicators, there was an agreement of 100% across all the assessors. For the cage level indicators, the agreement between assessors was over 80% and for the animal level indicators was over 90%. The validity of the everyday protocol was determined in experimental animals. This protocol included three additional indicators that were specific to the experimental design (superficial burns, cheek swelling and excessive grooming/licking/attention to the face). From the 19 indicators, three indicators were able to detect potential welfare problems: hunched posture, wounds/marks (including bite wounds) and cheek swelling scores as they were significantly higher for the mice that received the compound (which can cause pain) without analgesia compared with those mice that were administered the compound and analgesic treatment.

The results obtained from the final protocols showed that the differences in the indicators used for the two protocols (audit and every day) were minimal mainly linked with environmental-based indicators. The audit welfare protocol has *Body Condition Score, weight change, mortality rate* and *staff training* which were not present in the everyday welfare protocol. Conversely, *room temperature, wounds (excluding bite wounds), pups outside the nest* and *alertness* were included in the everyday protocol but not in the audit protocol. The most likely explanation for the differences between the protocol indicators could be related to differences in the aim of these two protocols as well as the likely person designated to perform the assessment. The audit welfare assessment is intended to be performed only once and by someone who is unlikely to have any knowledge about the facilities being assessed, such as a Home Office inspector. The everyday assessment, on the other hand, is performed on a daily basis and usually by a person who already has information about the facility such as mortality rate and staff training, such as an animal technician, so those indicators do not need to be included in this protocol. Another possible explanation for the differences in the assessment of these indicators is the difference in the time taken to perform these assessments. In general, the audit welfare protocol usually involves more time due to the additional indicators that need to be assessed related to the environment in the facility or animal rooms.

Both welfare assessment protocols show good levels of practicability when used to assess stock animals that are not undergoing any scientific procedures. The average time of 5 minutes per cage was considered to be acceptable when assessing group-housed mice (Hawkins *et al.*, 2011). Taking into account that the duration and frequency of the assessment can be dependent on the type of animals to be assessed (e.g. experimental or stock), because experimental animals usually required additional indicators specific to the study the duration of time spent on carrying out the assessment may increase (Hawkins *et al.*, 2011). However, when using the everyday protocol in the experimental animals (Chapter 5), this time (4 minutes per cage) did not increase compared with stock animals (6 minutes per cage). One factor that contributes to improving the practicability of the protocols was the sequence in which the indicators were recorded. This sequence was based on three different levels of assessment: facility/room, cage and individual level. This sequence allowed the assessor to focus on the assessment using a “funnel technique” where the assessment starts from the general environment (facility) drilling down to the individual animal. This funnel technique is usually used in social science for the construction of questionnaires and experimental designs and is valuable for improving practicality (Mandel, 1974; Bowling, 2005; Falzarano and Zipp, 2013). Another critical factor for improving practicability was the data scoring system used in the protocols which consisted of a traffic light system (Leach *et al.*, 2006; Sørensen and Fraser, 2010). Protocols for other species have used different schemes of scoring such as mathematical analysis [e.g. Welfare Quality, ARWIN (Botreau *et al.*, 2009; Dai *et al.*, 2016)] or score sheets (Lloyd and Wolfensohn, 1999; Mertens and Rülicke, 1999). However, the traffic light system used in this study provides a simple, visual way of presenting information, based on existing recommendations and best practice for each indicator (Leach *et al.*, 2006; Hawkins *et al.*, 2011). This traffic light system for the final assessment of welfare provided a good general overview of welfare and additionally provides information about specific indicators that need improvement. Similar to other studies that have used this system for providing animal welfare outcomes (Leach *et al.*, 2006; Council, 2009a; Sørensen and Fraser, 2010), the final interpretation of the traffic light allows review of welfare at different levels (e.g. welfare principle, welfare indicator) and highlights positive outcomes (indicators in green).

Both protocols (audit and every day) showed good levels of inter-observer reliability. The assessors demonstrated good levels of agreement when assessing stock animals. However, it is important to underline that this initial assessment was carried out only in stock animals, where high levels of welfare would be expected. As one of the main objectives of an animal welfare protocol is to recognise potential welfare problems, trialling the protocols in experimental animals is essential. Therefore, the everyday protocol was used to assess experimental animals over a 5-day period, with the assessment including additional study-specific indicators (Lloyd and Wolfensohn, 1999; Hawkins *et al.*, 2011). In the future when using the everyday welfare protocol for assessing experimental mice, study plans should be reviewed, and any indicators of potential adverse effects should be included. Indicators such as presence of superficial burns, cheek swelling and excessive grooming/licking/attention to the face in the study used in Chapter 5, stomach filling in breeding genetically modified animals (van der Meer *et al.*, 2001b) or tumour size for cancer models (Paster *et al.*, 2009; Workman *et al.*, 2010) are examples of the inclusion of specific indicators included in protocols.

From the list of the 20 indicators for the everyday protocol, only three indicators identified potential welfare complications in the experimental mice. Hunched posture, wounds/marks and cheek swelling all showed increasing scores when compared with their baseline levels. Hunched posture demonstrated to be an essential non-specific indicator for the assessment of welfare in laboratory mice in this study. As one of the most widely used indicators in experimental mice, especially in relation with pain assessment (Baumans *et al.*, 1994; Hawkins, 2002; Paster *et al.*, 2009), this indicator was also considered the most valid, reliable and practical for experts in the Delphi study. The high reliability of hunched posture is likely relevant as this specific body posture of the laboratory mouse is considered indicative of pain, sickness or distress (Baumans, 2005b; Baumans, 2005a; Paster *et al.*, 2009; Hawkins *et al.*, 2011).

Moreover, this indicator has been widely included in guidelines related to signs of discomfort or pain in experimental animals (Morton and Griffiths, 1985). Wound/marks is a general animal-based indicator related to biological functioning as it assesses superficial and deep wounds in the mouse body (Spangenberg and Keeling, 2015). The presence of wounds/marks in the body can be related to social problems (e.g. fight wounds), management procedures (e.g. trapped tail in the cage) or specific to an experimental design (e.g. surgical wounds). Due to the ease of scoring through observation in laboratory mice, the presence of wounds/marks as one of the main three indicators showed the high reliability of this indicator when assessing laboratory mouse welfare. Cheek swelling, on the other hand, was a specific indicator related to the study that was added on top of the everyday welfare indicators. This indicator was related to possible adverse effects associated with the study and its inclusion, and subsequent results showed the importance of including specific indicators for experimental animals.

The Qualitative Behavioural Assessment was shown to be a valid and reliable indicator for assessing laboratory mouse welfare under research conditions. The high inter-observer agreement showed that assessors were able to agree upon mouse emotional states by using their own descriptors in Free Choice Profiling study. Descriptors such as “in pain” or “lethargic” were used for describing animals that have undergone surgical procedures. This study is the first step towards the validation of QBA as an indicator and has shown that, when given the opportunity to use their own descriptors, assessors were highly consistent in the terms used in the assessment of laboratory mice. These results suggest that QBA could offer a valuable tool for the assessment of laboratory mouse emotional states as a main component of laboratory mouse welfare.

However, QBA requires further development in mice, to produce a fixed terms list that can be practically implemented within an animal unit, as the analysis and interpretation of the data following free choice profiling (FCP) is not straightforward. This FCP technique requires special statistical analysis which is time-consuming and impractical for an on-site welfare assessment (General Procrustes Analysis) (Grosso *et al.*, 2016). Therefore, the next logical step towards the validation of this method was to assess the reliability of the fixed terms. A list of 20 ‘fixed’ terms developed with four experts from the FCP study (Chapter 3). Data collected from the reliability assessment of the fixed term study (Chapter 4) showed a high inter-observer and intra-observer agreement. Assessors agreed that the videos of mice previously handled by the tail were described as anxious and fearful compared with the videos of the mice handled by tube which were described as calm and comfortable which was aligned with studies showing that tail handling results in more anxious mice (Hurst and West, 2010). Using the fixed terms of behavioural expressions for assessing emotional states in laboratory mice, QBA was able to support the studies that have shown tail handling as an aversive method for laboratory mice (Hurst and West, 2010; Gouveia and Hurst, 2013). Similarly to this study, a high correlation between QBA findings and other welfare measurements was observed in other QBA studies. Phythian *et al.* (2016), found a high correlation between QBA and lameness and dull physical demeanour in sheep, which can be valuable when interpreting the broader welfare impact of these health problems. A study in donkeys demonstrated a good correlation of QBA descriptors linked with positive emotional states (e.g. relaxed, happy, friendly) with behavioural measurements such as “positive reaction to the assessor” and “no avoidance distance from the assessor” (Minero *et al.*, 2016).

QBA was also included in the audit and everyday protocols for the assessment of laboratory mouse welfare. The results of QBA as an assessment method in both the audit and everyday welfare protocols were inconsistent. The assessors agreed upon the use of the fixed terms for assessing the cages, but they allocated the cages in different quadrants. The dimensions obtained from PCA analysis along with the scores of each cage represent the location of the animals/cages in the plots of the two dimensions (Brscic *et al.*, 2009). As the dimensions were labelled after the PCA analysis (e.g. dimension 1 labelled from social/positively engaged to agitated/uncertain and dimension 2 from playful to depressed/bored) the location of the cages in the dimensions determined the final behavioural expressions that represented the cage. For example, in the assessment of stock animals using the audit protocol at Dundee University, assessor one located the cages in the dimension labelled as fearful/tense whereas assessor two located the cages in the dimension labelled as playful/comfortable. Also, in the everyday protocol, the assessors located the cages in different dimensions, fearful/tense for assessor 1, playful/comfortable for assessor 2 and fearful/tense for assessor 3. These results showed that even if the assessors agreed in the use of the terms, they have a low agreement when deciding which behavioural expressions best describe each cage. One possible explanation of these results could be a lack of equivalence in how the Visual Analogue Scales were used between observers which were also seen in the previous studies (Chapter 3 and 4). This lack of equivalence of the scale between assessors means training and practice using the scales for the assessment of each behavioural expressions is required. This was carried out to a limited degree in this study and was potentially not clear enough, meaning the assessors may have needed more information about how the scale should be used for giving a numerical value to each term when assessing the animals (Minero *et al.*, 2016).

Taking into account the results of these studies about QBA as an indicator for the assessment of laboratory mouse welfare and in other species, QBA shows promise for welfare assessment of laboratory mice nevertheless more research is needed before it can be implemented as a laboratory welfare indicator routinely.

The everyday protocol was used in the assessment of welfare in experimental animals. The experimental study used for the assessment of the protocol had a moderate severity banding on the Home Office approved project licence. A moderate category implies that the animals are likely to experience short-term moderate pain, suffering or distress or long-lasting mild pain, suffering or distress (Home Office, 2014b). Therefore, following injection of the sensitising compound, the animals were likely to have experienced moderate short-term pain during the experiment, which made the study suitable for the assessment of the reliability of the welfare protocol. However, from the 20 indicators used in this study for the assessment of welfare only three were able to identify potential welfare problems in these experimental animals, demonstrated to be specific for detecting welfare problems, as the scores increased post-administration of a compound for these indicators compared with baseline levels. One of those indicators (cheek swelling) was taken from the license as a specific, study-related indicator. The other two indicators (hunched posture and wound/marks) were indicators which were included as standard in the everyday protocol. These results evidence the importance of including specific indicators related to the experiment as well as the validity of the everyday protocol for detecting potential welfare problems in laboratory mice when they are used in experimental studies.

6.4. Final protocol

Refinements to the two protocols were made based on the results of the reliability and practicability assessment (Chapter 5) to produce two final protocols (shown below). The refinements involved modifying the information from each indicator including the description to make it more concise and practical for the protocol. The traffic light system was also modified in some of the indicators (e.g. frequency of cleaning cages) to include other possible scenarios. In addition, the protocol has been organised into the three levels of assessment (facility/room, cage and individual) for improving the practicality of the protocol. These three levels of assessment include different animal-based and resource-based indicators that were distributed in each level which together represent the three main principles used in the definition of laboratory mouse welfare (biological health, psychological state and environment). The use of these two protocols, which included 20 indicators for the audit protocol and 16 indicators for the everyday protocol, for the assessment of welfare in laboratory mice.

The approach used for the development of these two protocols included the Delphi consultation process, the validation of a qualitative indicator (QBA) for the assessment of psychological states and the reliability and practicability assessment of the protocols. This approach allowed information to be collected from different sources taking into account the expert’s opinion, experimental studies and on-site information about the practicability and created two solid protocols that can be used either in an audit or every day for stock and experimental animals.

6.4.1. Audit welfare assessment protocol

FACILITY ASSESSMENT
DATE: FACILITY: START TIME: FINISH TIME:

Welfare indicator	Scale	Score
Temperature	Current room temperature	
Humidity	Current room humidity	
Ventilation	Current room/cage ventilation	
Handling method	Proportion of animal handling by tail	
	Proportion of animals handling by cup	
	Proportion of animals handling by tunnel	
Frequency of cleaning cages	Frequency of cleaning cages	
	Number of air changes per hour in the cages	
Type of cleaning procedure	Mice are transferred to a clean cage with clean sawdust and the nesting material from the dirty cage	
	Mice are transferred to a clean cage with clean sawdust, dirty sawdust for the dirty cage and clean nesting material	
	Mice are transferred to a clean cage with clean sawdust and clean nesting material.	
Staff training	Mainly all the staff have PDS related to the care of laboratory animals (more than 51%)	
	There is a proportion (30 % to50%) of the staff that have PDS	
	Staff do not have PDS (less than 29%)	

CAGE ASSESSMENT
CAGE ID: DATE: START TIME: FINISH TIME:

Welfare indicator	Scale	Score
Complexity of the cage	Additional elements or structures present	
	Only nesting material as additional element	
	No additional material or elements	
Alertness	Mouse alert when cage lid is open	
	It might respond to the observer but not immediately.	
	Mouse does not show any interest in the environment or the observer.	
Usage of nesting material	Complete or incomplete dome	
	A cup-shaped or flat nest	
	Untouched or scattered nesting material throughout the cage	
Qualitative Behavioural assessment (QBA)	See adjacent scale	
Blood stains in the cage	Blood stains absent in the home cage	
	Blood stains present in the home cage	

INDIVIDUAL ASSESSMENT
CAGE ID: MOUSE ID: START TIME: FINISH TIME:

Indicator	scale	Score
Hunched posture	No hunched posture present	
	Walks slowly in a hunched posture	
	Hunched posture with no movement	
Coat condition	Shiny, clean and well-groomed coat	
	Coat clean but ungroomed	
	Ruffled and untidy coat, it can be greasy and stick together	
Gait	Body weight supported on all limbs	
	Mouse might limp while walking	
	Mouse has difficulty moving forward or is reluctant to walk	
Ocular/nasal discharge	No ocular/nasal discharge	
	Water-like or mucus-like ocular/nasal discharge	
Wounds/marks (including bite wounds)	No wounds	
	Superficial wounds	
	Extensive and deep wounds	
Body Condition Score	The mouse is well-conditioned	
	The mouse in either under-conditioned or overweight	
	The mouse is either emaciated or obese.	
Barbering (hair removal)	No barbering	
	Barbering on different parts of the body	
Swollen abdomen	Abdomen is not swollen	
	Abdomen moderately swollen	
	Abdomen is very swollen	

Qualitative Behavioural Assessment
Cage ID Date:

Inquisitive	Min.	Max.
In pain	Min.	Max.
Positively Engaged	Min.	Max.
Lethargic	Min.	Max.
Energetic	Min.	Max.
Depressed	Min.	Max.
Determined	Min.	Max.
Anxious	Min.	Max.
Confident	Min.	Max.
Agitated	Min.	Max.
Calm	Min.	Max.
Fearful	Min.	Max.
Content	Min.	Max.
Tense	Min.	Max.
Comfortable	Min.	Max.
Uncertain	Min.	Max.
Playful	Min.	Max.
Bored	Min.	Max.
Sociable	Min.	Max.
Frustrated	Min.	Max.

ASSESSMENT AT FACILITY/ROOM LEVEL

GENERAL INFORMATION

These indicators will be assessed at facility/room level. They involve an interview with the technical manager of the laboratory facility and observations of the environmental conditions of the animal holding room. Resources required are:

Scoresheet
Humidity meter
Thermometer

TEMPERATURE

Description
Each housing room in laboratory facilities have a fixed thermometer that can be used to check the temperature on the room
Scoring method
Thermometer to record the current room temperature

Score information

- 🟢 Room temperature above 24°C
- 🟡 Room temperature from 24°C to 20°C
- 🔴 Room temperature below 20°C

HUMIDITY

Description
Each housing room in laboratory facilities have a fixed humidimeter that can be used to check the humidity on the room
Scoring method
Humidity meter to record the current room humidity

Score information

- 🟢 Room humidity 40%-60%
- 🟡 Room humidity outside (+ or -) the recommended (40% - 60%) by 1% - 10%
- 🔴 Room humidity outside (+ or -) the recommended (40% - 60%) by more than 11%

VENTILATION

Description
In most of the facilities, this ventilation is controlled at the cage level when IVCs (Individual Ventilated Cages) are used.
Scoring method
Ventilation information from facility records or IVC (individual Ventilated Cages) in the room

Score information

- 🟢 Room/IVCs ventilation 10 to 20 changes per hour
- 🟡 Room/IVCs ventilation outside (+ or -) the recommended (10 to 20 air changes per hour) by 10%
- 🔴 Room/IVCs ventilation outside (+ or -) the recommended (10 to 20 air changes per hour) by more than 11%

HANDLING METHOD

Description
There are different types of handling methods tail, cupped or tube handling. This indicator assesses the proportion of animals that are handling using the different methods.
Scoring method: Question asked to the personnel of the facility

- The proportion of mice handle by the tail, by cup and by tunnel

Score information

- 🟢 Over 80% of mice handle by tube
- 🟡 Mice handle by tail 50% and handling by tube 50%
- 🔴 Over 80% of mice handle by the tail

FREQUENCY OF CLEANING CAGES

Description
How frequent the animal's cages are cleaned at the facility
Scoring method
Questions asked the personnel of the facility: Frequency of cleaning cages and Ventilation information from facility records or IVC (individual Ventilated Cages) in the room

Score/air changes per hour	30	60	100
🔴	Cleaning cages once every month	Cleaning cages once six weeks	Cleaning cages once every other month
🟡	Cleaning cages once every other week	Cleaning cages once every month	Cleaning cages once six weeks
🟢	Cleaning cages once a week	Cleaning cages once every other week	Cleaning cages once every month
🟡	Cleaning cages twice per week	Cleaning cages once per week	Cleaning cages once every other week
🔴	Cleaning cages three times per week	Cleaning cages twice per week	Cleaning cages once per week

TYPE OF CLEANING PROCEDURE

Description
Information about the cage cleaning procedure
Scoring method
Question asked to the personnel of the facility which of the following options is used in the facility

Score information

- 🟢 Mice are transferred to a clean cage with clean sawdust and the nesting material from the dirty cage.
- 🟡 Mice are transferred to a clean cage with dirty sawdust for the dirty cage and clean nesting material.
- 🔴 Mice are transferred to a clean cage with clean sawdust and clean nesting material.

STAFF TRAINING

Description
The professional development scheme (PDS) (e.g. animal welfare seminars, biosecurity workshops) is valuable for increasing knowledge and animal welfare awareness. This indicator assess the level of training that the staff have in the facility.
Scoring method

Score information

- 🟢 More than 50% of the staff have PDS related to the care of laboratory animals
- 🟡 There is a proportion (30 % to 50%) of the staff that have PDS.
- 🔴 Staff do not have PDS (less than 29%)

ASSESSMENT AT CAGE LEVEL

GENERAL INFORMATION

The following indicators will be assessed at cage level. They involve taking the cage from the rack, placing it the designated area where the assessment will be carried out, opening the cage and removing the food/water hopper.
Resources needed:
Scoresheet

COMPLEXITY OF THE CAGE

Description
Cage complexity refers to the environmental enrichment that is provided within the cage in addition to the primary housing requirements for laboratory mice (Olsson and Dahlborn, 2002).
Scoring method
Observation of the animal's home cage, distribution and structures /elements present in the cage.

Score information

- (green) Cage has additional elements or structures (e.g. nesting material, wood stick)
- (yellow) Cage contains only nesting material as an additional element
- (red) Cage does not contain any additional elements or structures

ALERTNESS

Description
Alertness should be observed taking into account the circadian cycle. The animal should be observed as soon the case is open for 60 seconds (Lloyd and Wolfensohn, 1999).
Scoring method
Observation of the animal in the home cage when the cage is initially opened

Score information

- (green) The mouse is alert when the cage lid is open (focuses its attention towards the observer, may stand up if it is sleeping and sniff towards the observer
- (yellow) The mouse is less alert. It might respond to the observer but not immediately.
- (red) The mouse is immobile. It does not show any interest in the environment or the observer.

USE OF NESTING MATERIAL

Description
This indicator assesses the type of nest built from the available nesting material. The scoring system is modified from Hess *et al.* (2008).

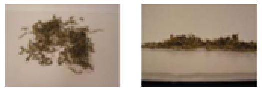
Score 0

(A) Completed dome. The walls completely enclose the nest hollow. A small exit hole may be found on the side or on the top of the nest.
(B) Incomplete dome: bedding material has been gathered to from a nest. The walls reach the widest point of an imaginary sphere that will fit the nest hollow

Score 1

(A) cup: The nest has identifiable walls that form a "cup". These walls would not reach the widest point of an imaginary sphere that would fill the nest hollow.
(B) Flat nest: the nesting material has been gathered for form a nest, identified by a clear nest cavity in the middle of the material or between the material and the cage wall. The nest is flat with no or incomplete walls

Score 2



Untouched or scattered: the nesting material has not been moved or it has been scattered throughout the cage.

Scoring method
Observation of the nest in the home cage when the cage is open

Score information

- (green) Complete or incomplete dome
- (yellow) A cup-shaped or flat nest
- (red) Untouched or scattered nesting material throughout the cage

QUALITATIVE BEHAVIOURAL ASSESSMENT (QBA)

Description
QBA in an indicator that uses the ability that people have to integrate various details in animals behaviour, postures and context into descriptions of animal "body language" using descriptors such as "relaxed", "tense" or "content". These terms have an emotional connotation which is relevant to animal welfare (Wemelsfelder, 2007).
Scoring method
This indicator is assessed by using a visual analogue scale for each descriptive term. The assessment is made at the cage level taking into account the group of animals as a whole. The cage should be left undisturbed with the observer close to it for 5 minutes, after, observation for one minute is made and then, the scales are used for each term in the order they are presented.
This indicator uses a Visual Analogue Scale to score each descriptor where:
0 mm is the minimum (the term is entirely absent) and
125mm is the maximum (the term is entirely present).

TERM	
Inquisitive	The mouse appears curious and interested in others and in exploring the environment. Willing to investigate.
In Pain	The mouse is suffering from physical discomfort leading the mouse to be reluctant to move, or to move with abnormal gait, or showing a tense, hunched or uncomfortable posture. The mouse looks like it is hurting or suffering and is in discomfort.
Positively engaged	The mouse is carrying out activities in a focused, directed and constructive manner. The mouse appears not to be distracted by others or the environment.
Lethargic	The mouse appears fatigued and sluggish. It has a lack of vigour and energy, showing low amounts of movement and any movement is slow and ponderous
Energetic	The mouse is carrying out an activity with a lot of energy and vigour, in a lively and excited manner.
Depressed	The mouse appears unhappy and without hope. It is apathetic, despondent and unresponsive showing little or no response or reaction to anyone or its environment. It appears isolated.
Determined	The mouse is showing an active and rapid reaction to something or someone. It appears to be focused on accomplishing a specific goal or task.
Anxious	The mouse is uneasy, cautious and nervous
Confident	The mouse is displaying assertiveness, behaving assertively with other animals and its environment in a self-assured manner.
Agitated	The mouse appears to be irritable and highly reactive. An excess of physical and/or cognitive activity is present because of anxiety.
Calm	The mouse appears peaceful and without worry. The mouse behaves in a relaxed and untroubled manner.
Fearful	The mouse appears afraid or scared. It seems to be avoiding contact with others and the environment, looks to be hiding, looking for a way out or trying to escape.
Content	The mouse appears happy and satisfied. Expressing happiness, with all its physiological, environmental and psychological needs met.
Tense	The mouse looks worried and emotionally tense. Its posture might evidence physical tension.
Comfortable	The mouse appears physically satisfied with the cage environment and looks relaxed and free from discomfort.
Uncertain	The mouse appears to be insecure; its physical movement is cautious. It moves slow showing alertness and insecurity. Avoidance reactions are showing with all stimuli
Playful	The mouse is engaging in lively movements purely to frolic or for fun, expressing pleasure, happiness and amusement.
Bored	The mouse appears uninterested in its environment and/or cage mates. The way it moves around and orients itself appears to be unfocused and aimless, without much energy, never staying long with a particular activity or aspect of the environment.
Sociable	The mouse actively interacts with others. It is willing to interact with others showing affiliative actions (e.g. grooming, resting in groups, sniffing etc.)
Frustrated	The mouse appears unfulfilled with its environment and/or cage mates. It looks stressed and uneasy showing repetitive and fast movements.

Score information

- 🟢 In the Principal Component Analysis (PCA) plot the laboratory facility is located in either the positive arousal/positive emotional valence quadrant or in the negative arousal/positive emotional valence quadrant
- 🟡 In the PCA plot, the laboratory facility is located in positive arousal/negative emotional valance quadrant
- 🔴 In the PCA plot, the laboratory facility is located in negative arousal/negative emotional valance quadrant

BLOOD STAINS IN THE CAGE

Description
Indicator assesses through observation from the home cage

Scoring method
Observation of the animal's home cage

Score information

- 🟢 Blood stains absent in the home cage
- 🟡 N/A
- 🔴 Blood stains present in the home cage

ASSESSMENT AT INDIVIDUAL LEVEL

GENERAL INFORMATION

These indicators will be assessed at the individual level. They involve close observation and in some cases, physical restraint of the animals.

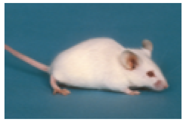
Resources needed:

- Scoresheet

HUNCHED POSTURE

Description

This posture is recognised by a prominent arched back when the mouse is resting/sitting still. See images below for further illustration.


Normal posture


Hunched posture

Scoring method

Observation of the animal in the home cage before any handling.

Score information

- No hunched posture present
- Walks slowly with hunched posture
- Hunched posture with no movement

COAT CONDITION

Description

Animal coat is observed while is in the home cage, the quality of the coat is assesses including the cleanness. See images below for further illustration.


Shiny, clean, and well-groomed coat


Coat clean but ungroomed


Ruffled and untidy coat

Scoring method

Observation of the animal in the home cage before any handling.

Score information

- Shiny, clean, smooth and well-groomed coat
- Coat clean but ungroomed
- Ruffled and untidy coat, it can be greasy and stick together

GAIT

Description

This indicator related to how the animal moves around the cage and is assessed by observing the mouse walking around the home cage.

Scoring method

Observation of the animal walking around the home cage before any handling.

Score information

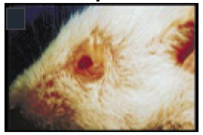
- Body weight supported on all limbs, with its abdomen not touching the ground, and with both hind limbs participating evenly
- The mouse might limp while walking
- The mouse has difficulty moving forward or is reluctant to walk and drags its abdomen along the ground

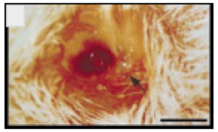
OCULAR/NASAL DISCHARGE

Description

The mouse should have their eyes and nose dry and clean from all types of discharge.


No discharge


water-like discharge


mucus-like discharge

Scoring method

Observation of the animal walking around the home cage before any handling.

Score information

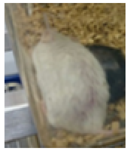
- The mouse does not have ocular/nasal discharge
- N/A
- The mouse has water-like or mucus-like ocular/nasal discharge

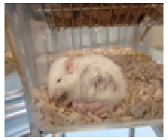
WOUNDS/MARKS (INCLUDING BITE WOUNDS)

Description

This indicator includes bite wounds (usually located in the back, flanks, base of the tail, or genitals) and wounds originated other than for conspecifics, e.g. self-inflicted (excessive grooming) or because of environmental problems, management procedures or hanging problems (e.g. tail trapped in the cage). See images below for further illustration.


No wounds


superficial wounds


Extensive and deep wounds

Scoring method

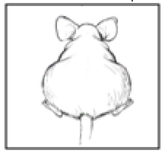
Observation of the animal before handling

Score information

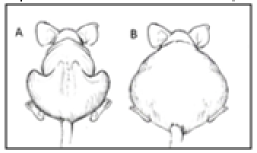
- The mouse does not have wounds/marks
- The mouse has wounds which involve superficial tissue (skin), no bleeding and extension is less than 10% of the body.
- The mouse has wounds which involve deep tissue (e.g. muscle), bleeding and / or the extension of the wounds is more than 10% of body

BODY CONDITION SCORE

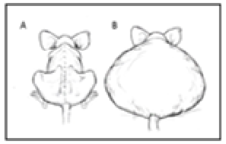
Description
The scoring of the body condition is the process of observing the amount of flesh covering bony protuberances which is mainly in dependent on sex, age, body frame size, and pregnancy status in females (Ullman-Culleré and Foltz, 1999). See diagrams below for further illustration: Mouse body condition score, Modified from (Ullman-Culleré and Foltz, 1999).



Well-conditioned. Vertebral and dorsal pelvis not prominent; palpable with slight pressure



The mouse is either under-conditioned or overweight. Picture A- under conditioned mouse, the segmentation of vertebral column evident. Dorsal pelvic bones are readily palpable. Picture B- overweight mouse, the spine is a continuous column. Vertebrae palpable only with firm pressure



The mouse is either emaciated or obese. Picture A- Emaciated mouse, skeletal structure extremely prominent; little or no flesh covet, vertebrate distinctly segmented. Picture B- Obese mouse, the mouse is smooth and bulky. Bone structure disappears under flesh and subcutaneous fat.

Scoring method
Observation of the animal in the home cage before any handling.

Score information

- 🟢 The mouse is well-conditioned. Vertebral and dorsal pelvis not prominent; palpable with slight pressure
- 🟡 The mouse in either under-conditioned or overweight.
- 🔴 The mouse is either emaciated (A) or obese (B)

BARBERING (HAIR REMOVAL)

Description
The behaviour involves the trimming if whiskers and fur anywhere on the body (Spangenberg and Keeling, 2015). See images below for further illustration.

No barbering.

Barbering on different parts of the body

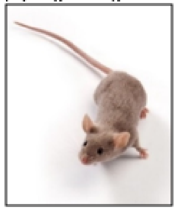
Scoring method
Observation of the animal in the home cage before handling.

Score information

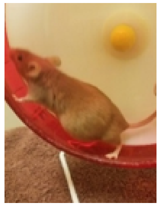
- 🟢 The mouse does not have trimmed whiskers/ fur
- 🟡 The mouse has a small zone of fur-trimmed but no whiskers
- 🔴 The mouse has an extensive area of fur and whiskers trimmed

SWOLLEN ABDOMEN

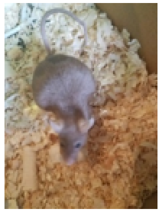
Description
In a normal healthy mouse, the abdomen will be soft with mild resistance to pressure. If the abdomen is swollen or distended it can be an indication of illness such as a tumour or fluid accumulation (Spangenberg and Keeling, 2015). See images below for further illustration.



Abdomen is not swollen



Abdomen moderately swollen



Abdomen is very swollen

Scoring method
Observation of the animal before handling.

Score information

- 🟢 The abdomen is not swollen
- 🟡 The abdomen is moderately swollen
- 🔴 The abdomen is very swollen

6.4.2. Everyday welfare assessment protocol

Assessment sheet for the everyday protocol

Facility/cage level assessment

Facility: cage ID:

Indicator	Scale	Date				
Temperature	Current room temperature					
Humidity	Current room humidity					
Ventilation	Current room/cage ventilation					
Complexity of the cage	Additional elements or structures present					
	Only nesting material as additional element					
	No additional material or elements					
Alertness	Mouse alert when cage lid is open					
	It might respond to the observer but not immediately.					
	Mouse does not show any interest in the environment or the observer.					
Use of nesting material	Complete or incomplete dome					
	A cup-shaped or flat nest					
	Untouched or scattered nesting material throughout the cage					
Qualitative Behavioural assessment (QBA)	See adjacent image					
Blood stains in the cage	Blood stains absent in the home cage					
	Blood stains present in the home cage					

Qualitative Behavioural Assessment

Cage ID Date:

	Min.	Max.
Inquisitive	Min.	Max.
In pain	Min.	Max.
Positively Engaged	Min.	Max.
Lethargic	Min.	Max.
Energetic	Min.	Max.
Depressed	Min.	Max.
Determined	Min.	Max.
Anxious	Min.	Max.
Confident	Min.	Max.
Agitated	Min.	Max.
Calm	Min.	Max.
Fearful	Min.	Max.
Content	Min.	Max.
Tense	Min.	Max.
Comfortable	Min.	Max.
Uncertain	Min.	Max.
Playful	Min.	Max.
Bored	Min.	Max.
Sociable	Min.	Max.
Frustrated	Min.	Max.

Individual assessment

Cage ID: Mouse ID:

Indicator	Scale	Date				
Hunched position	No hunched posture present					
	Walks slowly in a hunched posture					
	Hunched posture with no movement					
Coat condition	Shiny, clean and well-groomed coat					
	Coat clean but ungroomed					
	Ruffled and untidy coat, it can be greasy and stick together					
Gait	Body weight supported on all limbs					
	Mouse might limp while walking					
	Mouse has difficulty moving forward or is reluctant to walk					
Ocular/nasal discharge	No ocular/nasal discharge					
	Water-like or mucus-like ocular/nasal discharge					
Wounds/marks (including bite wounds)	No wounds					
	Superficial wounds					
	Extensive and deep wounds					
Body Condition Score	The mouse is well-conditioned					
	The mouse in either under-conditioned or overweight					
	The mouse is either emaciated or obese.					
Barbering (hair removal)	No barbering					
	Barbering on different parts of the body					
Swollen abdomen	Abdomen is not swollen					
	Abdomen moderately swollen					
	Abdomen is very swollen					

COMMENTS:

Protocol for an everyday laboratory mouse welfare assessment

ASSESSMENT AT FACILITY/ROOM LEVEL

GENERAL INFORMATION

These indicators will be assessed at the facility/room level. They involve observations of the environmental conditions of the animal holding room. Resources required are:

Scoresheet
Humidity meter
Thermometer

TEMPERATURE

Description
Each housing room in laboratory facilities have a fixed thermometer that can be used to check the temperature on the room

Scoring method
Thermometer to record the current room temperature

Score information

- 🟢 Room temperature above 24°C
- 🟡 Room temperature from 24°C to 20°C
- 🔴 Room temperature below 20°C

HUMIDITY

Description
Each housing room in laboratory facilities have a fixed humidimeter that can be used to check the humidity on the room

Scoring method
Humidity meter to record the current room humidity

Score information

- 🟢 Room humidity 40%-60%
- 🟡 Room humidity outside (+ or -) the recommended (40% - 60%) by 1% - 10%
- 🔴 Room humidity outside (+ or -) the recommended (40% - 60%) by more than 11%

VENTILATION

Description
In most of the facilities, this ventilation is controlled at the cage level when IVCs (Individual Ventilated Cages) are used (AAALAC, 2011).

Scoring method
Ventilation information from facility records or IVC (individual Ventilated Cages) in the room

Score information

- 🟢 Room/IVCs ventilation 10 to 20 changes per hour
- 🟡 Room/IVCs ventilation outside (+ or -) the recommended (10 to 20 air changes per hour) by 10%
- 🔴 Room/IVCs ventilation outside (+ or -) the recommended (10 to 20 air changes per hour) by more than 11%

ASSESSMENT AT CAGE LEVEL

GENERAL INFORMATION

The following indicators will be assessed at cage level. They involve taking the cage from the rack, placing it the designated area where the assessment will be carried out, opening the cage and removing the food/water hopper.
Resources needed:
Scoresheet

COMPLEXITY OF THE CAGE

Description
Cage complexity refers to the environmental enrichment that is provided within the cage in addition to the primary housing requirements for laboratory mice (Olsson and Dahlborn, 2002).

Scoring method
Observation of the animal's home cage, distribution and structures /elements present in the cage.

Score information

- 🟢 Cage has additional elements or structures (e.g. nesting material, wood stick)
- 🟡 Cage contains only nesting material as an additional element
- 🔴 Cage does not contain any additional elements or structures

ALERTNESS

Description
Alertness should be observed taking into account the circadian cycle. The animal should be observed as soon the case is open for 60 seconds (Lloyd and Wolfensohn, 1999).

Scoring method
Observation of the animal in the home cage when the cage is initially opened

Score information

- 🟢 The mouse is alert when the cage lid is open (focuses its attention towards the observer, may stand up if it is sleeping and sniff towards the observer
- 🟡 The mouse is less alert. It might respond to the observer but not immediately.
- 🔴 The mouse is immobile. It does not show any interest in the environment or the observer.

USE OF NESTING MATERIAL

Description
This indicator assesses the type of nest built from the available nesting material. The scoring system is modified from Hess *et al.* (2008).

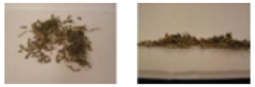
Score 0

(A) Completed dome. The walls completely enclose the nest hollow. A small exit hole may be found on the side or on the top of the nest.
(B) Incomplete dome: bedding material has been gathered to from a nest. The walls reach the widest point of an imaginary sphere that will fit the nest hollow

Score 1

(A) cup: The nest has identifiable walls that form a "cup". These walls would not reach the widest point of an imaginary sphere that would fill the nest hollow.
(B) Flat nest: the nesting material has been gathered for form a nest, identified by a clear nest cavity in the middle of the material or between the material and the cage wall. The nest is flat with no or incomplete walls

Score 2



Untouched or scattered: the nesting material has not been moved or it has been scattered throughout the cage.

Scoring method
Observation of the nest in the home cage when the cage is open

Score information
- 🟢 Complete or incomplete dome
- 🟡 A cup-shaped or flat nest
- 🔴 Untouched or scattered nesting material throughout the cage

QUALITATIVE BEHAVIOURAL ASSESSMENT (QBA)

Description
QBA in an indicator that uses the ability that people have to integrate various details in animals behaviour, postures and context into descriptions of animal "body language" using descriptors such as "relaxed", "tense" or "content". These terms have an emotional connotation which is relevant to animal welfare (Wemelsfelder, 2007).
Scoring method
This indicator is assessed by using a visual analogue scale for each descriptive term. The assessment is made at the cage level taking into account the group of animals as a whole. The cage should be left undisturbed with the observer close to it for 5 minutes, after, observation for one minute is made and then, the scales are used for each term in the order they are presented.
This indicator uses a Visual Analogue Scale to score each descriptor where:
0 mm is the minimum (the term is entirely absent) and
125mm is the maximum (the term is entirely present).

TERM	
Inquisitive	The mouse appears curious and interested in others and in exploring the environment. Willing to investigate.
In Pain	The mouse is suffering from physical discomfort leading the mouse to be reluctant to move, or to move with abnormal gait, or showing a tense, hunched or uncomfortable posture. The mouse looks like it is hurting or suffering and is in discomfort.
Positively engaged	The mouse is carrying out activities in a focused, directed and constructive manner. The mouse appears not to be distracted by others or the environment.
Lethargic	The mouse appears fatigued and sluggish. It has a lack of vigour and energy, showing low amounts of movement and any movement is slow and ponderous
Energetic	The mouse is carrying out an activity with a lot of energy and vigour, in a lively and excited manner.
Depressed	The mouse appears unhappy and without hope. It is apathetic, despondent and unresponsive showing little or no response or reaction to anyone or its environment. It appears isolated.
Determined	The mouse is showing an active and rapid reaction to something or someone. It appears to be focused on accomplishing a specific goal or task.
Anxious	The mouse is uneasy, cautious and nervous
Confident	The mouse is displaying assertiveness, behaving assertively with other animals and its environment in a self-assured manner.
Agitated	The mouse appears to be irritable and highly reactive. An excess of physical and/or cognitive activity is present because of anxiety.
Calm	The mouse appears peaceful and without worry. The mouse behaves in a relaxed and untroubled manner.
Fearful	The mouse appears afraid or scared. It seems to be avoiding contact with others and the environment, looks to be hiding, looking for a way out or trying to escape.
Content	The mouse appears happy and satisfied. Expressing happiness, with all its physiological, environmental and psychological needs met.
Tense	The mouse looks worried and emotionally tense. Its posture might evidence physical tension.
Comfortable	The mouse appears physically satisfied with the cage environment and looks relaxed and free from discomfort.
Uncertain	The mouse appears to be insecure; its physical movement is cautious. It moves slow showing alertness and insecurity. Avoidance reactions are showing with all stimuli
Playful	The mouse is engaging in lively movements purely to frolic or for fun, expressing pleasure, happiness and amusement.
Bored	The mouse appears uninterested in its environment and/or cage mates. The way it moves around and orients itself appears to be unfocused and aimless, without much energy, never staying long with a particular activity or aspect of the environment.
Sociable	The mouse actively interacts with others. It is willing to interact with others showing affiliative actions (e.g. grooming, resting in groups, sniffing etc.)
Frustrated	The mouse appears unfulfilled with its environment and/or cage mates. It looks stressed and uneasy showing repetitive and fast movements.

Score information

- In the Principal Component Analysis (PCA) plot the laboratory facility is located in either the positive arousal/positive emotional valence quadrant or in the negative arousal/positive emotional valence quadrant
- In the PCA plot, the laboratory facility is located in positive arousal/negative emotional valance quadrant
- In the PCA plot, the laboratory facility is located in negative arousal/negative emotional valance quadrant

BLOOD STAINS IN THE CAGE

Description
Indicator assesses through observation from the home cage
Scoring method
Observation of the animal's home cage

Score information

- Blood stains absent in the home cage
- N/A
- Blood stains present in the home cage

ASSESSMENT AT INDIVIDUAL LEVEL

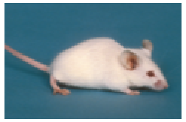
GENERAL INFORMATION

These indicators will be assessed at the individual level. They involve close observation and in some cases, physical restraint of the animals.
Resources needed:

- Scoresheet

HUNCHED POSTURE

Description
This posture is recognised by a prominent arched back when the mouse is resting/sitting still. See images below for further illustration.


Normal posture


Hunched posture

Scoring method
Observation of the animal in the home cage before any handling.

Score information

- No hunched posture present
- Walks slowly with hunched posture
- Hunched posture with no movement

COAT CONDITION

Description
Animal coat is observed while is in the home cage, the quality of the coat is assesses including the cleanness. See images below for further illustration.


Shiny, clean, and well-groomed coat


Coat clean but ungroomed


Ruffled and untidy coat

Scoring method
Observation of the animal in the home cage before any handling.

Score information

- Shiny, clean, smooth and well-groomed coat
- Coat clean but ungroomed
- Ruffled and untidy coat, it can be greasy and stick together

GAIT

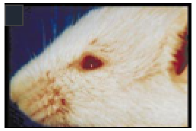
Description
This indicator related to how the animal moves around the cage and is assessed by observing the mouse walking around the home cage.
Scoring method
Observation of the animal walking around the home cage before any handling.

Score information

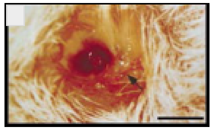
- Body weight supported on all limbs, with its abdomen not touching the ground, and with both hind limbs participating evenly
- The mouse might limp while walking
- The mouse has difficulty moving forward or is reluctant to walk and drags its abdomen along the ground

OCULAR/NASAL DISCHARGE

Description
The mouse should have their eyes and nose dry and clean from all types of discharge.


No discharge


water-like discharge


mucus-like discharge

Scoring method
Observation of the animal walking around the home cage before any handling.

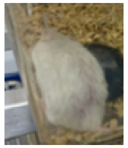
Score information

- 🟢 The mouse does not have ocular/nasal discharge
- 🟡 N/A
- 🔴 The mouse has water-like or mucus-like ocular/nasal discharge

WOUNDS/MARKS (INCLUDING BITE WOUNDS)

Description
This indicator includes bite wounds (usually located in the back, flanks, base of the tail, or genitals) and wounds originated other than for conspecifics, e.g. self-inflicted (excessive grooming) or because of environmental problems, management procedures or hanging problems (e.g. tail trapped in the cage). See images below for further illustration.


No wounds


superficial wounds


Extensive and deep wounds

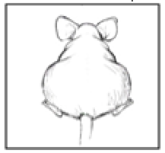
Scoring method
Observation of the animal before handling

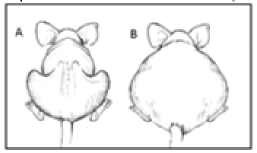
Score information

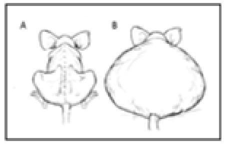
- 🟢 The mouse does not have wounds/marks
- 🟡 The mouse has wounds which involve superficial tissue (skin), no bleeding and extension is less than 10% of the body.
- 🔴 The mouse has wounds which involve deep tissue (e.g. muscle), bleeding and / or the extension of the wounds is more than 10% of body

BODY CONDITION SCORE

Description
The scoring of the body condition is the process of observing the amount of flesh covering bony protuberances which is mainly in dependent on sex, age, body frame size, and pregnancy status in females (Ullman-Culleré and Foltz, 1999). See diagrams below for further illustration: Mouse body condition score, Modified from (Ullman-Culleré and Foltz, 1999).


Well-conditioned. Vertebral and dorsal pelvis not prominent; palpable with slight pressure


The mouse is either under-conditioned or overweight. Picture A- under conditioned mouse, the segmentation of vertebral column evident. Dorsal pelvic bones are readily palpable. Picture B- overweight mouse, the spine is a continuous column. Vertebrae palpable only with firm pressure


The mouse is either emaciated or obese. Picture A- Emaciated mouse, skeletal structure extremely prominent; little or no flesh covet, vertebrate distinctly segmented. Picture B- Obese mouse, the mouse is smooth and bulky. Bone structure disappears under flesh and subcutaneous fat.

Scoring method
Observation of the animal in the home cage before any handling.

Score information

- 🟢 The mouse is well-conditioned. Vertebral and dorsal pelvis not prominent; palpable with slight pressure
- 🟡 The mouse in either under-conditioned or overweight.
- 🔴 The mouse is either emaciated (A) or obese (B)

BARBERING (HAIR REMOVAL)

Description
The behaviour involves the trimming if whiskers and fur anywhere on the body (Spangenberg and Keeling, 2015). See images below for further illustration.


No barbering.



Barbering on different parts of the body

Scoring method
Observation of the animal in the home cage before handling.

Score information

- 🟢 The mouse does not have trimmed whiskers/ fur
- 🟡 The mouse has a small zone of fur-trimmed but no whiskers
- 🔴 The mouse has an extensive area of fur and whiskers trimmed

6.5. Limitations of the study

The process for constructing both protocols had some potential limitations and challenges. The protocols were structured based on the preliminary list of indicators used in the Delphi consultation and the two scenarios created for the study (audit and every day). This Delphi consultation involved the consultation of experts in laboratory mouse welfare about the most valid, reliable and practical indicators for the assessment of laboratory mouse welfare. A potential limitation of this approach can be related to the level of expertise of the participants of the Delphi consultation. Although the level of expertise was assessed taking into account the years of working with laboratory animals, qualifications and job position, the recruitment of the study were voluntary, thus the expertise of the participants could be affected by this voluntary recruitment process.

Laboratory animal welfare assessment is a very complex topic, and initial guidelines need to be provided to ensure the outcomes of the protocols in terms of detecting welfare problems and highlighting instances of positive welfare (Hawkins *et al.*, 2011). The inclusion of an animal welfare definition is determinant for constructing the welfare protocol as this definition provided additional scope against which the indicators were assessed giving additional support to the indicator selection process (Botreau *et al.*, 2009; Hawkins *et al.*, 2011). A potential limitation of this study could be related to the scoring system in the protocols. Generally, in other species, the outcomes of welfare protocols come from complex data analysis and statistics which provide a final numerical value for the welfare scores. For example, in farm animals, the Welfare Quality program involved a series of studies defining the scoring system, including aggregation measures for producing an overall assessment (Bracke *et al.*, 1999; Botreau *et al.*, 2007a). This process is highly time-consuming and requires specific calculations, and data analysis which is not straightforward (Jones and Manteca, 2009) and could not be used cage-side. The traffic light system used in this study had the advantage of being easy and reliable for providing information about the animal’s welfare in a rapid manner. However, the development of this scoring system was difficult as for some of the indicators, the information concerning the means of measurement could be not found in the literature making difficult the inclusion of these indicators in the protocol as the specific recommendation or adequate level of the indicators cannot be found. Consequently, for some indicators, this lack of information meant that these indicators were removed from the final list (e.g. staff training, attitude from staff towards animals), although they may be relevant for laboratory mouse welfare.

The difficulty in assessing laboratory mouse energy levels was demonstrated in both QBA studies, the Free Choice Profiling (Chapter 3) and fixed terms (Chapter 4) studies. This limitation could be related to the selection of the assessors, especially in the fixed terms study. Assessors who participated in this study did not have any previous knowledge of laboratory mouse biology, behaviour, housing or welfare. This lack of knowledge about the species being assessed could have affected the final results as having knowledge about the normal social behaviour, demeanour and the typical housing and environmental conditions of the species being assessed is recommended (Heston, 1967; Boursot *et al.*, 1993). As shown in a QBA study by Minero *et al.* (2016), the assessors with more experience working with buffaloes had a better association between qualitative and quantitative indicators showing that the experience might play an essential role in recognising subtle changes in animals demeanour.

Laboratory mice are very active animals, and spend most of their time engaged in active behaviours (e.g. grooming, walking, exploring) when being observed, i.e. when they are not resting or sleeping in the nest. The differentiation between a calm mouse and a mouse in pain can be difficult to see for someone with a lack of knowledge of the species. The reason for not using experienced participants in the QBA fixed-term study was related to videos used which involved interaction tests, which we did not want the participants to have knowledge of as this may have introduced further bias. The researcher also selected the videos in the fixed terms study with the aim of selecting the most representative videos for each handling method (i.e. tail or tube handling). For future studies, it is recommended that the selection videos be carried out by someone who is blind to the treatments applied to guarantee that they are an accurate representation of the treatments applied, the welfare state induced and animal's emotional states. Another critical point to take into account in the QBA studies is the need for calibration and training in use of the VAS and the fixed terms. Standardisation of the fixed terms to develop a common understating is critical for the success of the assessment when using QBA. Another possible limitation in the QBA studies was related to the use of the indicators at a cage level of assessment. The inclusion of QBA as a cage level assessment indicators was taken align with other welfare protocols that have used these indicators as part of the assessment at pen or farm level (Welfare Quality®, 2009a; Battini *et al.*, 2015). However, when using this same approach to laboratory mouse, the results were not straightforward, especially in the experimental animals. The result showed that QBA need to be used at individual level in experimental animals as the design of these experiments can be very variable, and the assessment of each individual animal is paramount for the study.

Although the indicators used in each protocol were selected considering an organised and strict process, the final use of some of them was not straightforward. Indicators such as frequency of cleaning cages were difficult to assess, and the traffic light system was difficult to develop as there are many factors that affect this indicator (e.g. humidity, type of bedding, presence of nesting material, animal behaviour). General information about the recommendation of these indicators was taken into account for the traffic light system. However, when used in practice, the scale required additional modifications for including all the possible scenarios.

The sample size used in the reliability assessment of both protocols can be considered as a limitation of this study. This sample size was calculated based on other protocols where a large number of animals needed to be assessed. A total of 5% to 10% of stock animals were assessed, however taking into account the time spent on both, the audit and the everyday protocols, this sample size can be increased up to 25% of the animals which might be a better reflection of the welfare status of the facility. In addition, the number of participants used for the assessment of reliability of the protocols could be increased for obtaining more information from different participants, increasing the sample size and improving the reliability of the results.

6.6. Conclusion

The initial process of building up the protocols using the Delphi consultation has been shown to be a reliable tool for obtaining information from experts worldwide. This tool can be used in future studies for gathering more information about laboratory mouse welfare indicators, especially the ones that are currently being used on a daily basis in laboratory facilities. Furthermore, the fact that this study has developed two main protocols which represent the two key ways in which laboratory mouse welfare is assessed (audit and everyday assessment), provides valuable information about the differences between these protocols as well as the potential use in a broad range of situations that can compromise laboratory mouse welfare such as experimental procedures and the impact of everyday housing and husbandry.

The audit and everyday welfare protocols created in this study along with their assessment sheets are shown in section 6.4.1 and 6.4.2. The assessment of welfare is a complex process which must involve the assessment of all principles of animal welfare highlighted by the welfare definition applied. The final protocols created in this study can be used as a one-time-only welfare assessment tool (i.e. an audit) or welfare cheeks conducted daily. The advantage of having the protocol organised into three levels of assessment allowed the assessor to use these protocols for a different purpose (i.e. every day vs audit assessment). Each section (facility, cage and individual level) can be used according to the specific needs, for example, individual indicators can be included according to the specific protocols applied, taking into account the details of the experimental design and what needs to be closely monitored (i.e. humane endpoints). Furthermore, in laboratory mice, these individual indicators can be used to form primary indicators for assessing the welfare of animals enrolled in experiments. However, additional indicators need to be considered that are related to experimental protocols including the assessment of pain, harm and distress that these procedures may cause.

In general, the indicators used in both protocols (audit and every day) showed good levels of reliability between participants. The final outcome of animal welfare was defined by three principles (biological health, psychological state and environment) which also defined in terms of the indicators use for each principle in each protocol. Each individual indicator assesses different principles at different levels (e.g. facility, cage or individual), thus is the aggregation of these indicators which finally produce the outcome of animal welfare. In addition, each indicator was assessed individually in terms of reliability, and they showed to be reliable when used together as part of a protocol. The facility/room indicators showed to be highly reliable and practical. Indicators such as room temperature and ventilation were practical and reliable when measured for the assessors. Indicators used as part of the cage level assessment such as use of nesting material and complexity of the cage also were practical and reliable showing that the aggregation of this specific indicators for the protocol was successful in assessing the welfare of the animals. Furthermore, this study highlights the importance of essential indicators used in the welfare assessment of laboratory mice. Hunched posture, for example, showed to be a very effective indicator for the assessment of laboratory mouse physical health, and it should be included in all protocols when assessing welfare. This was demonstrated by its inclusion in the Delphi consultation process and the practicability and reliability of this indicator in both stock and experimental animals.

The psychological welfare indicators can be challenging to assess as part of a welfare protocol, but they are required for a comprehensive assessment of animal welfare. QBA as an indicator was included in the protocols, showed promising results in the assessment of emotional states. As discussed in other studies, there is a need to find tools that provide a detailed assessment of the animal's entire behavioural repertoire and the environment in relation to emotional states (Flecknell *et al.*, 2011) for which QBA can be a suitable indicator.

6.7. Future research

One of the main objectives of an animal welfare protocol is to recognise potential welfare problems in animals. Therefore, the assessment of the validity, reliability and practicability of the protocols developed in this project need to be assessed in animals undergoing different experimental procedures including models with different experimental designs with different potential welfare impacts and levels of severity.

More research is needed regarding the assessment of practicality and reliability of the protocols at the cage level. Different variables need to be further investigated such as the importance of the different levels of experience of assessors with laboratory animals to assesses if this experience has an impact on the assessment results. Also, further research about the reliability and practicability of QBA, particularly the inclusion of additional assessors when using the fixed terms list in mice. The fixed terms created in chapter 4 need further refinement to assess if all 20 terms are reliable for this assessment of if they need to be modified.

Other indicators that were not included in the protocol due to the lack of literature about the relevance in the assessment of laboratory welfare might need additional investigation. Indicators such as staff attitude towards animals have been researched in other species such as farm animals (Hemsworthlt *et al.*, 1993) and results have shown a positive relationship between a positive attitude and animal's productivity. However, in laboratory mice, there was no information about the impact of people attitudes on the welfare of laboratory animals. Frequency of cleaning cages, for example, showed to be a complex indicator to assess and further research about how other factors affect this indicator is required in order to refine its use in a laboratory mouse welfare protocol.

One of the most challenging topics about welfare is the implementation of measurements for the assessment of positive welfare. Most of the indicators present in these protocols are intended to assess when the animal welfare is compromising assessing evidence of welfare problems (e.g. Weight loss, pain behaviours). However, there has been an increased interest in the assessment of positive welfare as it is considered as the next step towards animal welfare. Recent research about the possible means for positive welfare have shown promising results and have open the possibilities to new and novel indicators (Yeates and Main, 2008; Mellor and Beausoleil, 2015; Webster, 2016). Furthermore one of the indicators, used in these protocols have been included as a positive welfare indicator for farm animals in the European Union Welfare Quality Program (Welfare Quality®, 2009c) . Qualitative Behavioural Assessment, have shown promising results in other species as a positive animal welfare indicator and more research deem to be necessary for laboratory animals.

Appendix

Appendix A. Instructions for the QBA given to the observers prior to part 1 of the Free Choice Profiling

Introduction

Thank you all very much for coming; I really appreciate you taking the time to help me this study.

So what is Qualitative Behaviour Assessment?

Qualitative Behaviour Assessment or QBA is a method of welfare assessment that focuses on the whole animal - how the whole animal behaves. Normally in science, we are used to measuring separate elements of behaviour, for example, walking, sitting, biting, sniffing, etc. However, when looking at an animal, we see more than just whether it walks or lies or sniffs, all these behaviours are also done in a certain **style**, with a certain **expressive quality**. *Any* behaviour can be done in different ways; an animal can walk around in a way that is calm and relaxed, or tense and agitated; it can look at something in a curious way, or in a fearful way. So we can see the same behaviour – e.g. walking - but done in different styles. We see not only *what* the animal does, but also *how* it does what it does. QBA is about an animal's experience and how they feel in their environment.

This method is based on the fact that the way animals behave has a certain **expressive quality, a body language** so to speak. Moreover, describing this body language provides us with interesting information about the animal's welfare state - whether it is relaxed or tense is relevant to understanding welfare.

To assess body language, you look at posture, gaze, speed + direction of movement, but what counts is the whole picture, how the whole animal interacts with its environment and how they feel.

QBA was originally developed in livestock species and has not yet been applied to laboratory animals, so the aim of the sessions over the next few days is to determine if this method is suitable for mice because they are obviously very different to cows, sheep and pigs. My aim is not to test your ability to assess the welfare of these mice–

I just want to see if the method works, how you perceive mouse body language and if you agree with each other about what do you perceive.

This study takes place in 2 phases:

Phase 1: Today. In this phase, all you do is create your own terms for describing the mouse expressions. We will watch 20 video clips, and you should write down these terms on the forms. More details in a minute.

Phase 2: in the next meeting, we will use your own list of terms to score the animals' expressions on a quantitative rating scale. So you stick with your own terms, they are not mixed with other people's terms. This will be explained next time.

A. The videos: what you will see:

- 20 video clips which show a number of laboratory mouse. You will be focusing on just one animal per clip as they were filmed one mouse at the time.
- The clips are 45 seconds long – break after 10 clips or so.
- When the clip stops, you see a white screen. You will have 1 and a half minutes to write down terms on the form; I will check you are ready and then the next clip will start.
- The mice in the clips I will show you are laboratory animals housed at an academic facility in the UK. They were in a filming cage at the time of filming, so this is not its home cage. The mice could have experienced various interventions; however, I do not know what.

B. The videos: what we ask you to do:

- Basically, you sit back, relax, and focus on **how** the animal behaves. You try and assess the expressive quality of how they are engaging with the environment. Is it calm, or nervous, or tense? You will not see this clearly right away. It emerges over time; you gradually gain an impression of the overall expression of the interaction.
- When the clip ends, let everything you have seen sink in. What was this animal like? Search for the words that best capture that image, and write them down.
- Use as many terms as you like, until you feel you have captured the different aspects of the animal's expression. The expression has different layers: an animal can be confident but tense, or confident and relaxed, and it can be relaxed and lively, or relaxed and calm. Try and think about these different layers.

- Don't think: "this term is not scientific enough"; what counts is how *you* perceive the animal's style of behaving.
- Do not write down sentences, or descriptions of *what* the animal is doing (e.g. chewing, running around, vocalising). Only write down terms that describe the qualities of *how* the animal is doing what it is doing. These should mostly be single terms (e.g. calm), and can sometimes be composite terms (e.g. easy-going, attention-seeking, etc.). Do not add quantifiers to your terms (e.g. very worried, a little tense).
- Quality is more important than quantity! – It is not better to have lots and lots of terms. Individual preferences: some people find it easy to come up with words; others say everything in a few words. All OK.
- It is OK to scribble down terms while watching, but not as a rule as this is distracting
- You have 1 minute to do all this, and then the next clip starts.

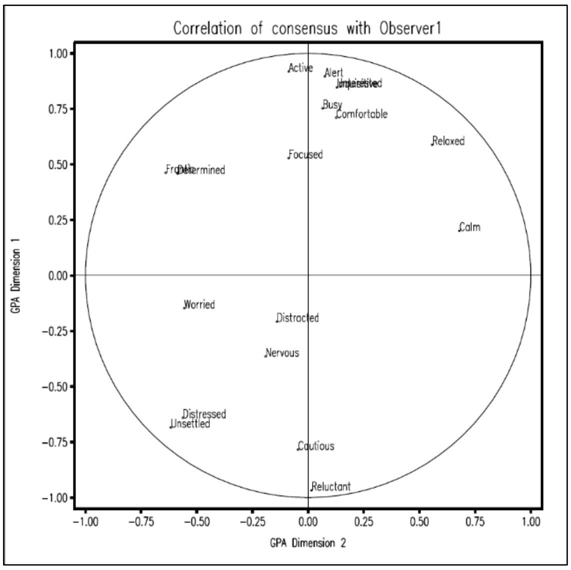
The form:

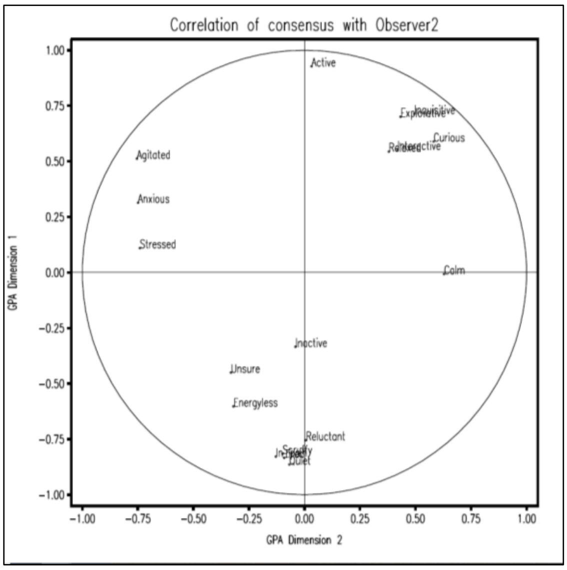
- Please fill in your name on every page of the form – I need to know whom to give the VAS to.
- Each clip has its own section for writing down words.
- The number of each clip is shown above the video clip.
- With every new clip, you are again completely free to choose your terms: you can use new terms, or ones you have used before; you can choose 1-2 terms or 10. What matters is to choose those terms that are best for the animal you have just watched.
- You need to warm up, that is OK. Do not go back to change what you have done before.

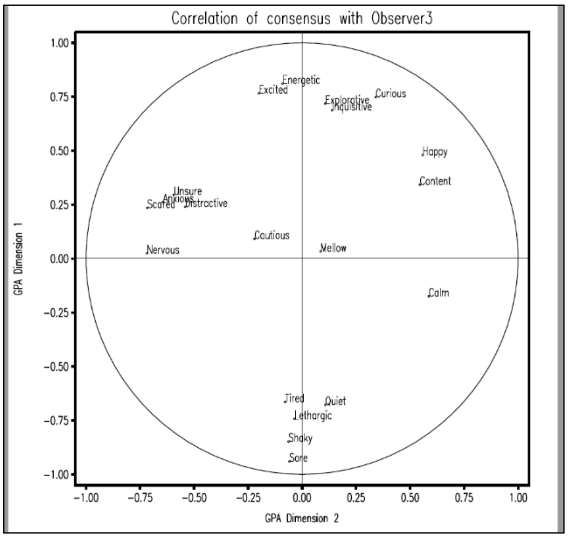
A few final points:

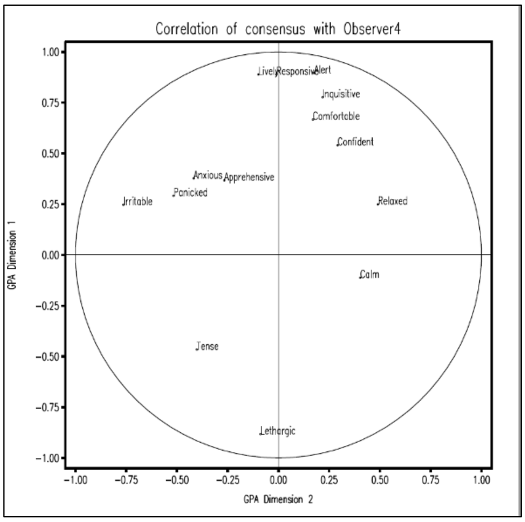
- It is not possible to make a mistake. What matters is how **YOU** interpret the animal's expressions.
- Please write so that we can read it.
- Do not talk about expressions to each other (break!). Your choice of terms has to be <u>independent of other observers!</u>

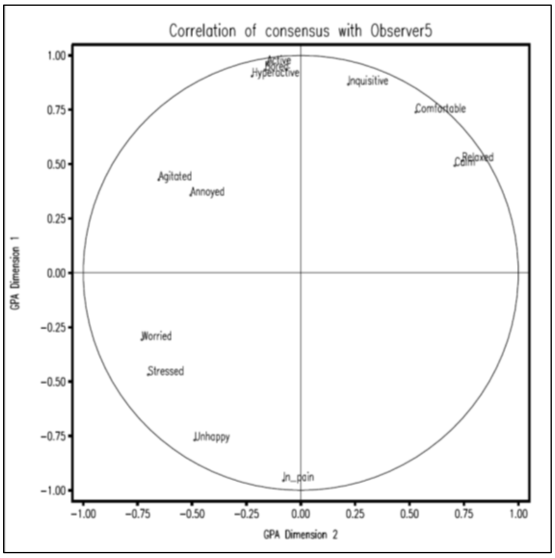
Appendix B. Word charts created from the QBA-FCP data analysis for each of the 20 Observers

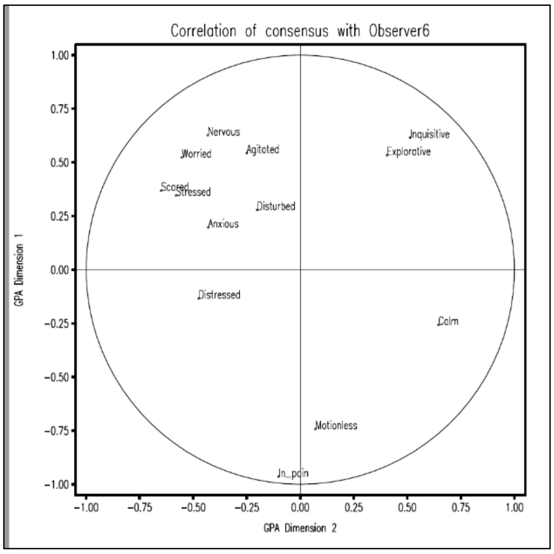


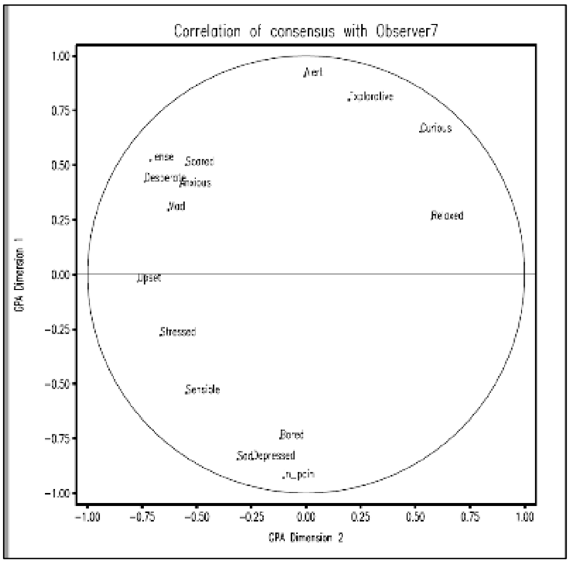


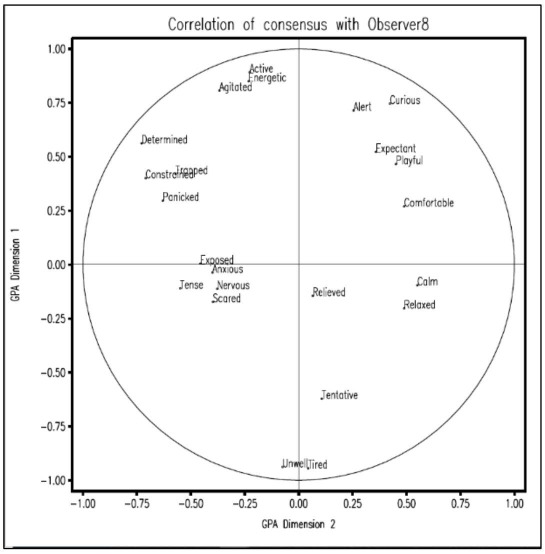


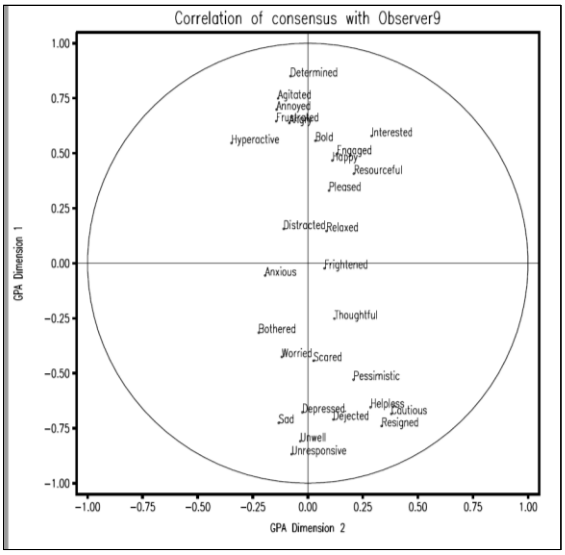


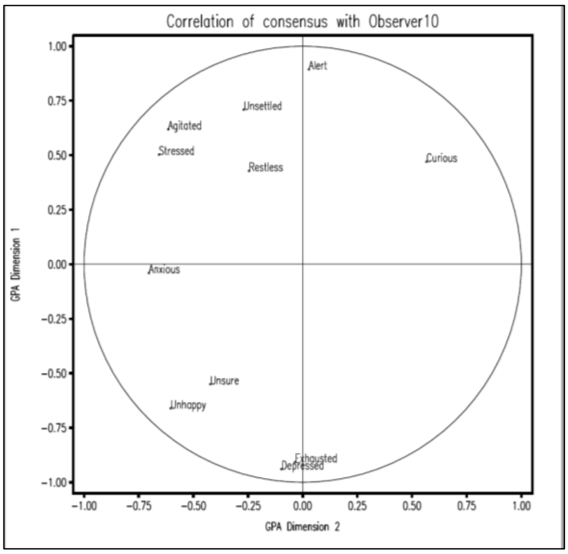


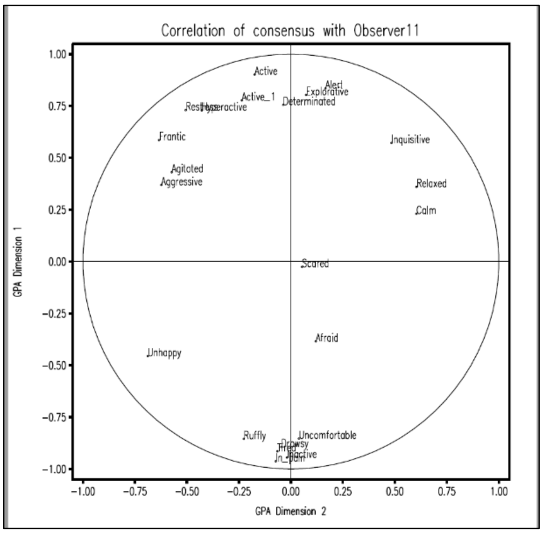


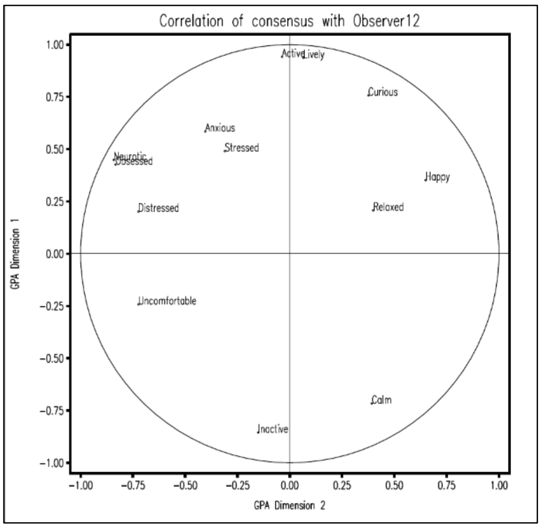

Correlation of consensus with Observer7
GPA Dimension 1
GPA Dimension 2
Alert
Explorative
Curious
Tense
Scared
Desperate
Anxious
Mad
Relaxed
Upset
Stressed
Sensible
Bored
Sad
Depressed
In_pain

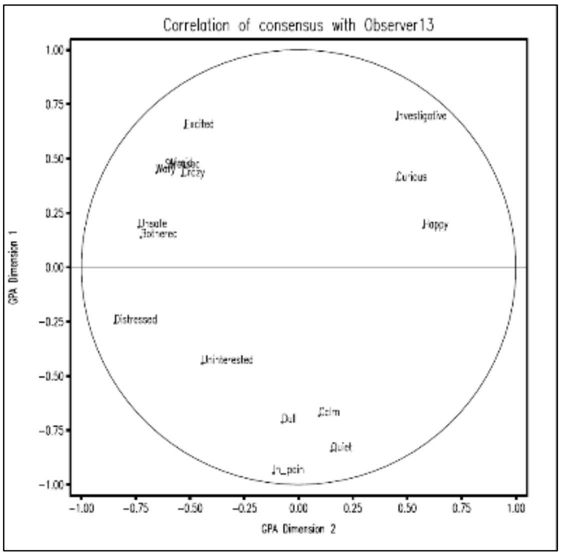

Correlation of consensus with Observer8
GPA Dimension 1
GPA Dimension 2
Active
Energetic
Agitated
Alert
Curious
Determined
Expectant
Playful
Trapped
Panicked
Comfortable
Exposed
Anxious
Tense
Nervous
Scared
Relieved
Calm
Relaxed
Tentative
Unwell
Tired

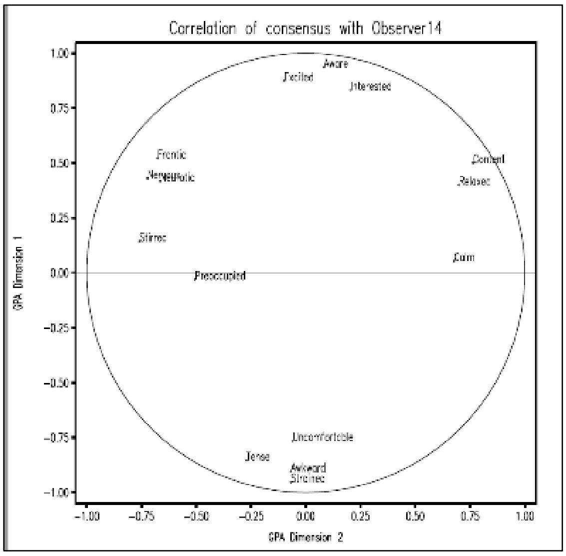

Correlation of consensus with Observer9
GPA Dimension 1
GPA Dimension 2
Determined
Agitated
Annoyed
Hyperactive
Bold
Interested
Engaged
Happy
Resourceful
Pleased
Distracted
Relaxed
Anxious
Frightened
Thoughtful
Bothered
Worried
Scared
Pessimistic
Depressed
Dejected
Helpless
Cautious
Resigned
Sad
Unwell
Unresponsive

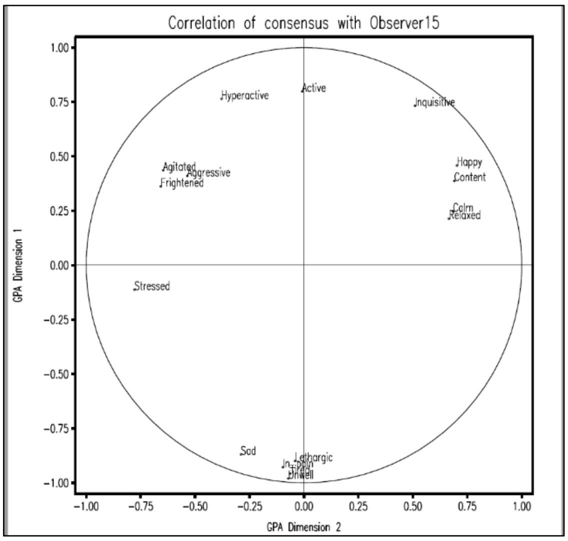

Correlation of consensus with Observer10
GPA Dimension 1
GPA Dimension 2
Alert
Unsettled
Agitated
Stressed
Restless
Curious
Anxious
Unsure
Unhappy
Exhausted
Depressed

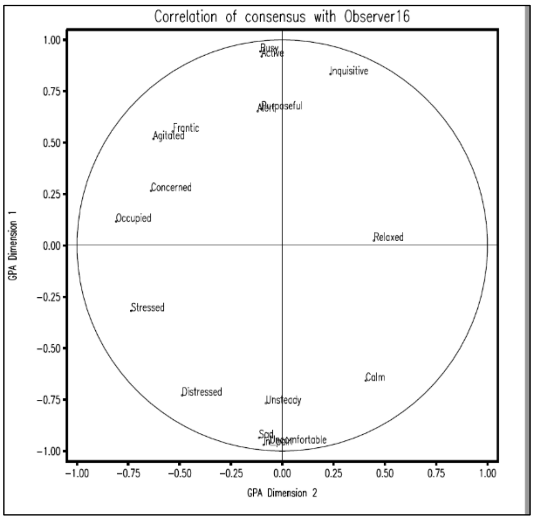

Correlation of consensus with Observer11
GPA Dimension 1
GPA Dimension 2
Active
Alert
Explorative
Active_1
Determinated
Hyperactive
Frantic
Inquisitive
Agitated
Aggressive
Relaxed
Calm
Scared
Afraid
Unhappy
Ruffly
Uncomfortable
Drowsy
Inactive

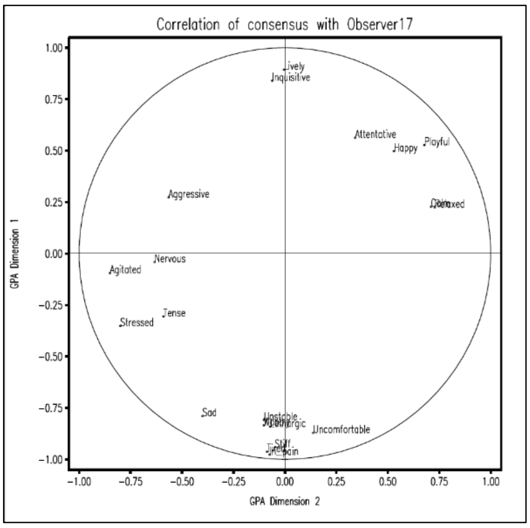

Correlation of consensus with Observer12
GPA Dimension 1
GPA Dimension 2
Active
Lively
Curious
Anxious
Stressed
Neurotic
Diseased
Happy
Distressed
Relaxed
Uncomfortable
Calm
Inactive

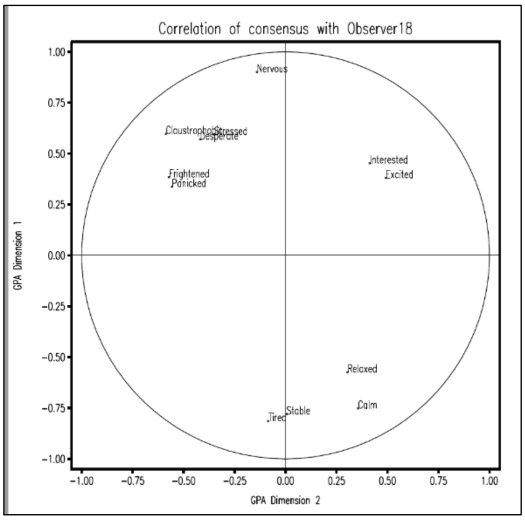

Correlation of consensus with Observer13
GPA Dimension 1
GPA Dimension 2
Excited
Investigative
Curious
Happy
Unsafe
Distressed
Uninterested
Calm
Dull
Quiet
In_pain

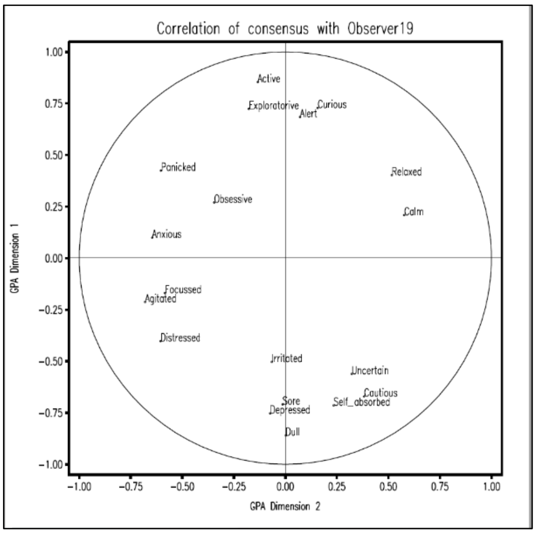

Correlation of consensus with Observer14
GPA Dimension 1
GPA Dimension 2
Aware
Excited
Interested
Frantic
Content
Relaxed
Stirred
Calm
Preoccupied
Uncomfortable
Tense
Awkward

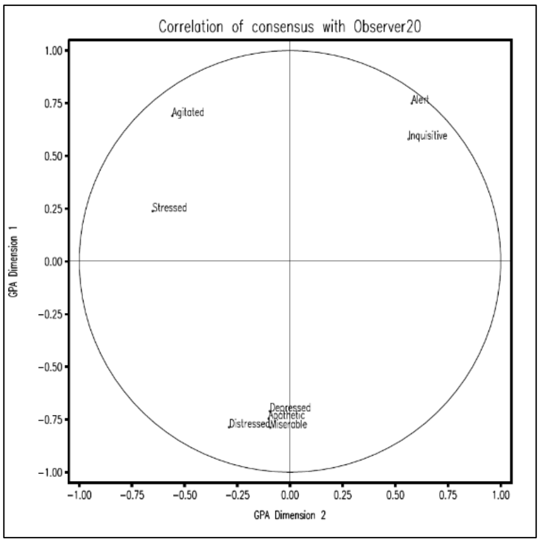

Correlation of consensus with Observer15
GPA Dimension 1
GPA Dimension 2
Active
Hyperactive
Inquisitive
Agitated
Aggressive
Frightened
Happy
Content
Calm
Relaxed
Stressed
Sad
Lethargic


Correlation of consensus with Observer16
GPA Dimension 1
GPA Dimension 2
Inquisitive
Frantic
Concerned
Occupied
Relaxed
Stressed
Calm
Distressed
Unsteady


Correlation of consensus with Observer17
GPA Dimension 1
GPA Dimension 2
Lively
Inquisitive
Attentative
Playful
Happy
Aggressive
Nervous
Agitated
Tense
Stressed
Sad
Uncomfortable


Correlation of consensus with Observer18
GPA Dimension 1
GPA Dimension 2
Nervous
Interested
Excited
Frightened
Panicked
Relaxed
Calm
Stable
Tired


Correlation of consensus with Observer19
GPA Dimension 1
GPA Dimension 2
Active
Exploratorive
Curious
Alert
Panicked
Relaxed
Obsessive
Calm
Anxious
Focussed
Agitated
Distressed
Irritated
Uncertain
Cautious
Sore
Self_absorbed
Depressed
Dull


Correlation of consensus with Observer20
GPA Dimension 1
GPA Dimension 2
Agitated
Alert
Inquisitive
Stressed
Depressed
Apathetic
Distressed
Miserable

Appendix C. Assessment sheets for the audit protocol

FACILITY ASSESSMENT

DATE:

FACILITY:

START TIMES: FINISH TIME:

Welfare indicator	Scale	SCORE
Temperature	Current room temperature	
Humidity	Current room humidity	
Ventilation	Current room/cage ventilation	
Handling method	Proportion of animal handled by tail	
	Proportion of animals handled by cupping	
	Proportion of animals handled by tunnel	
Frequency of cleaning cages	Frequency of cleaning cages	
	Number of air changes per hour in the cages	
Type of cleaning procedure	Mice are transferred to a clean cage with clean sawdust and the nesting material from the dirty cage	
	Mice are transferred to a clean cage with clean sawdust, dirty sawdust for the dirty cage and clean nesting material	
	Mice are transferred to a clean cage with clean sawdust and clean nesting material.	
Staff training	Mainly all the staff have Professional Development Scheme (PDS) related to the care of laboratory animals (more than 51%)	
	There is a proportion (30 % to 50%) of the staff that have PDS	
	Staff do not have PDS (less than 29%)	

CAGE ASSESSMENT

START TIME: FINISH TIME:

CAGE ID:

Welfare indicator	Scale	score
Complexity of the cage	Additional elements or structures present	
	Only nesting material as additional element	
	No additional material or elements	
Alertness	Mouse alert when cage lid is open	
	Mouse might respond to the observer but not immediately.	
	Mouse does not show any interest in the environment or the observer.	
Use of nesting material	Complete or incomplete dome	
	A cup-shaped or flat nest	
	Untouched or scattered nesting material throughout the cage	
Qualitative Behavioural assessment (QBA)	See next page	
Bloodstains in the cage	Bloodstains absent in the home cage	
	Bloodstains present in the home cage	

QUALITATIVE BEHAVIOURAL ASSESSMENT

CAGE ID:

DATE:

	Min.	Max.
Inquisitive		
In pain		
Positively Engaged		
Lethargic		
Energetic		
Depressed		
Determined		
Anxious		
Confident		
Agitated		
Calm		
Fearful		
Content		
Tense		
Comfortable		
Uncertain		
Playful		
Bored		
Sociable		
Frustrated		

INDIVIDUAL ASSESSMENT

START TIME: FINISH TIME:

CAGE ID

MOUSE ID

Indicator	scale	Score
Hunched posture	No hunched posture present	
	Walks slowly in a hunched posture	
	Hunched posture with no movement	
Coat condition	Shiny, clean and well-groomed coat	
	Coat clean but ungroomed	
	Ruffled and untidy coat, it may be greasy and stick together	
Gait	Bodyweight supported on all limbs	
	Mouse might limp while walking	
	Mouse has difficulty moving forward or is reluctant to walk	
Ocular/nasal discharge	No ocular/nasal discharge	
	Water-like or mucus-like ocular/nasal discharge	
Wounds/marks (including bite wounds)	No wounds	
	Superficial Wounds	
	Extensive and deep wounds	
Body Condition Score	The mouse is well-conditioned	
	The mouse in either under-conditioned or overweight	
	The mouse is either emaciated or obese.	
Barbering (hair removal)	No barbering	
	Barbering on different parts of the body	
Swollen abdomen	Abdomen is not swollen	
	Abdomen moderately swollen	
	Abdomen is very swollen	

Appendix D. Protocol for audit laboratory mouse welfare assessment

ASSESSMENT AT FACILITY/ROOM LEVEL

GENERAL INFORMATION

These indicators will be assessed at facility/room level. They involve an interview with the technical manager of the laboratory facility and observations of the environmental conditions of the animal holding room. Resources required are:

- Scoresheet
- Humidity meter
- Thermometer

TEMPERATURE

Description

Room temperature for laboratory mice is usually maintained at 20°C to 24°C as suggested by the Code of Practice for Housing and Care of Laboratory Animals (Home Office, 2014a). However, recent scientific recommendations have suggested that the thermoneutral mouse zone is 26°C to 29°C (Gaskill *et al.*, 2009). To reduce the need for an increase in laboratory ambient temperature, the provision of adequate nesting material in the home cage can alleviate thermal distress (Gaskill *et al.*, 2012).

Scoring method

Thermometer to record the current room temperature

Score information

- (green) Room temperature from 24°C to 29°C
- (amber) Room temperature from 24°C to 20°C
- (red) Room temperature below 20°C

HUMIDITY

Description

As a general rule, the humidity in a mouse room should be kept at 40 to 60% (Home Office, 2014a). This indicator is measured at room level, and records should be kept to avoid any sudden or prolonged periods below 40% or above 60% (Home Office, 2014a). High levels of humidity are usually associated with health problems (Chesler *et al.*, 2002) (e.g. reproduction) and behavioural changes, e.g. (Nelson *et al.*, 1990).

Scoring method
Humidity meter to record the current room humidity

Score information

🟢 Room humidity 40%-60%

🟠 Room humidity outside (+ or -) the recommended (40% - 60%) by 1% - 10%

🔴 Room humidity outside (+ or -) the recommended (40% - 60%) by more than 11%

VENTILATION

Description
The purpose of ventilation in an animal room is to provide good quality air and to minimise the levels of gases, odours and infectious agents. It is advisable to have 10 to 20 air changes per hour in a fully stocked room (Reeb-Whitaker *et al.*, 2001). In most of the facilities, this ventilation is controlled at the cage level when IVCs (Individual Ventilated Cages) are used (AAALAC, 2011).

Scoring method

Ventilation information from facility records or IVC (individual Ventilated Cages) in the room

Score information

🟢 Room/IVCs ventilation 10 to 20 changes per hour

🟠 Room/IVCs ventilation outside (+ or -) the recommended (10 to 20 air changes per hour) by 1 – 10%

🔴 Room/IVCs ventilation outside (+ or -) the recommended (10 to 20 air changes per hour) by more than 11%

HANDLING METHOD

Description

The type of handling used it is also essential for animals welfare. Animals should be handled with care by trained personnel (Hurst and West, 2010). There are different types of handling methods tail, cupped or tube handling. This indicator assesses the proportion of animals that are handling using different methods. Tube handling is considered the best method for handling laboratory mouse as different studies have shown a decreasing in anxiety levels when this method is used as common practice (Hurst and West, 2010)

Scoring method

Question asked to the personnel of the facility

- The proportion of mice handle by the tail
- The proportion of mice handle by cup
- The proportion of mice handle by tunnel

	Score information
	Over 80% of mice handle by tube
	Mice handle by tail 50% and handling by tube 50%
	Over 80% of mice handle by the tail

FREQUENCY OF CLEANING CAGES

Description

The frequency of cage cleaning in laboratory mice depends on the type of cage, the stocking density and the capacity of the ventilation system for keeping the air quality (Home Office, 2014a). In general, cages are changed once a week when the ventilation rate is at 30 air changes per hour (Home Office, 2014a). In some cases, cages are also changed every 14 and 30 days if the ventilation rate is at 60 and 100 air changes per hour respectively (Reeb-Whitaker *et al.*, 2001). The frequency of cleaning cages has a direct impact in animal's welfare as affecting the cleanliness of the bedding and the relative humidity which have an impact on the proliferation of bacteria and ammonia production (Reeb-Whitaker *et al.*, 2001). The scores are being attributed taking into account the table below:

Score/air changes per hour	30	60	100
	Cleaning cages once a week	Cleaning cages once every other week	Cleaning cages once every month
	Cleaning cages twice per week	Cleaning cages once per week	Cleaning cages once every other week
	Cleaning cages three times per week	Cleaning cages twice per week	Cleaning cages once per week

Scoring method

Questions asked the personnel of the facility

- Frequency of cleaning cages
- Ventilation information from facility records or IVC (individual Ventilated Cages) in the room

TYPE OF CLEANING PROCEDURE

Description

Type of cleaning procedure affects the social interaction between laboratory mice. Cages can be cleaned in different ways; the most common and appropriate method is to transfer mice to a clean cage with clean sawdust and transfer their nesting material from the dirty cage (Van Loo *et al.*, 2000). Animals can also be transferred to clean cages with clean sawdust plus additional sawdust containing both urine and faeces from the dirty cage (Van Loo *et al.*, 2000). Another cleaning procedure involves transfer mice to a clean cage with clean sawdust and clean nesting material. The last two procedures have been reported as cause increasing levels of aggression in laboratory mice (Van Loo *et al.*, 2000). Scores were attributed taking into account current best practices.

Scoring method

Question asked to the personnel of the facility which of the following options is used in the facility

Score information

- 🟢 Mice are transferred to a clean cage with clean sawdust and the nesting material from the dirty cage.
- 🟡 Mice are transferred to a clean cage with dirty sawdust for the dirty cage and clean nesting material.
- 🔴 Mice are transferred to a clean cage with clean sawdust and clean nesting material.

STAFF TRAINING

Description

Appropriately trained staff, in matters related to animal's care and welfare, are paramount for identifying welfare concerns, in order to identify health and welfare compromises quickly and accurately. For general staff, there are no special requirements by law, but formal education about animal care is encouraged. The professional development scheme (PDS) (e.g. animal welfare seminars, biosecurity workshops) is also valuable for increasing knowledge and animal welfare awareness. Courses covering laboratory mouse behaviour, peri-operative care, and recognition of pain and its management are included in this category. This indicator assesses the level of training that the staff have in the facility.

Scoring method

Question asked to the personnel of the facility which of the following options is used in the facility

Score information

- 🟢 More than 50% of the staff have PDS related to the care of laboratory animals
- 🟡 There is a proportion (30 % to 50%) of the staff that have PDS.
- 🔴 Staff do not have PDS (less than 29%)

ASSESSMENT AT CAGE LEVEL

GENERAL INFORMATION

The following indicators will be assessed at cage level. They involve taking the cage from the rack, placing it the designated area where the assessment will be carried out, opening the cage and removing the food/water hopper. Resources needed:

- Scoresheet

COMPLEXITY OF THE CAGE

Description

Cage complexity refers to the environmental enrichment that is provided within the cage in addition to the primary housing requirements for laboratory mice (Olsson and Dahlborn, 2002). Mice require additional enrichment in the cage to be able to cope with stressors and perform natural behaviours (Olsson and Dahlborn, 2002). Structures that increase the complexity of the cage, such as material for building a shelter or a nest, additional pellets or seeds as well as chewing material among others are recommended (Olsson and Dahlborn, 2002).

Scoring method

Observation of the animal's home cage, distribution and structures /elements present in the cage.

Score information

- 🟢 Cage has additional elements or structures (e.g. nesting material, wood stick)
- 🟡 Cage contains only nesting material as an additional element
- 🔴 Cage does not contain any additional elements or structures

ALERTNESS

Description

Mice should show interest in the environment, wanting to explore and be attentive. Alertness should be observed taking into account the circadian cycle. As mice are crepuscular animals, it would be usual to find the mouse sleep and resting during daylight hours. The animal should be observed as soon the case is open for 60 seconds (Lloyd and Wolfensohn, 1999).

Scoring method

Observation of the animal in the home cage when the cage is initially opened

Score information

- 🟢 The mouse is alert when the cage lid is open (focuses its attention towards the observer, may stand up if it is sleeping and sniff towards the observer)
- 🟡 The mouse is less alert. It might respond to the observer but not immediately.
- 🔴 The mouse is immobile. It does not show any interest in the environment or the observer.

USE OF NESTING MATERIAL

Description

The usage of nesting material to build a nest in the home cage is related to the mouse-specific behaviour used for controlling the environment and thermoregulation (Hess *et al.*, 2008). This indicator assesses the type of nest built from the available nesting material. The scoring system is modified from Hess *et al.* (2008).

Score 0

(A) Completed dome. The walls completely enclose the nest hollow. A small exit hole may be found on the side or on the top of the nest.

(B) Incomplete dome: bedding material has been gathered to from a nest. The walls reach the widest point of an imaginary sphere that will fit the nest hollow

Score 1

(A) cup: The nest has identifiable walls that form a "cup". These walls would not reach the widest point of an imaginary sphere that would fill the nest hollow.

(B) Flat nest: the nesting material has been gathered for form a nest, identified by a clear nest cavity in the middle of the material or between the material and the cage wall. The nest is flat with no or incomplete walls

Score 2

Untouched or scattered: the nesting material has not been moved or it has been scattered throughout the cage.

Scoring method

Observation of the nest in the home cage when the cage is open

Score information

- (green) Complete or incomplete dome
- (amber) A cup-shaped or flat nest
- (red) Untouched or scattered nesting material throughout the cage

QUALITATIVE BEHAVIOURAL ASSESSMENT (QBA)

Description

QBA in an indicator that uses the ability that people have to integrate various details in animals behaviour, postures and context into descriptions of animal "body language" using descriptors such as "relaxed", "tense" or "content". These terms have an emotional connotation which is relevant to animal welfare (Wemelsfelder, 2007).

Scoring method

This indicator is assessed by using a visual analogue scale for each descriptive term. The assessment is made at the cage level taking into account the group of animals as a whole. The cage should be left undisturbed with the observer close to it for 5 minutes, after, observation for one minute is made and then, the scales are used for each term in the order they are presented.

This indicator uses a Visual Analogue Scale to score each descriptor where:

0 mm is the minimum (the term is entirely absent) and

125mm is the maximum (the term is entirely present).

TERM	
Inquisitive	The mouse appears curious and interested in others and in exploring the environment. Willing to investigate.
In Pain	The mouse is suffering from physical discomfort leading the mouse to be reluctant to move, or to move with abnormal gait, or showing a tense, hunched or uncomfortable posture. The mouse looks like it is hurting or suffering and is in discomfort.
Positively engaged	The mouse is carrying out activities in a focused, directed and constructive manner. The mouse appears not to be distracted by others or the environment.
Lethargic	The mouse appears fatigued and sluggish. It has a lack of vigour and energy, showing low amounts of movement and any movement is slow and ponderous
Energetic	The mouse is carrying out an activity with a lot of energy and vigour, in a lively and excited manner.
Depressed	The mouse appears unhappy and without hope. It is apathetic, despondent and unresponsive showing little or no response or reaction to anyone or its environment. It appears isolated.
Determined	The mouse is showing an active and rapid reaction to something or someone. It appears to be focused on accomplishing a specific goal or task.
Anxious	The mouse is uneasy, cautious and nervous

Confident	The mouse is displaying assertiveness, behaving assertively with other animals and its environment in a self-assured manner.
Agitated	The mouse appears to be irritable and highly reactive. An excess of physical and/or cognitive activity is present because of anxiety.
Calm	The mouse appears peaceful and without worry. The mouse behaves in a relaxed and untroubled manner.
Fearful	The mouse appears afraid or scared. It seems to be avoiding contact with others and the environment, looks to be hiding, looking for a way out or trying to escape.
Content	The mouse appears happy and satisfied. Expressing happiness, with all its physiological, environmental and psychological needs met.
Tense	The mouse looks worried and emotionally tense. Its posture might evidence physical tension.
Comfortable	The mouse appears physically satisfied with the cage environment and looks relaxed and free from discomfort.
Uncertain	The mouse appears to be insecure; its physical movement is cautious. It moves slow showing alertness and insecurity. Avoidance reactions are showing with all stimuli
Playful	The mouse is engaging in lively movements purely to frolic or for fun, expressing pleasure, happiness and amusement.
Bored	The mouse appears uninterested in its environment and/or cage mates. The way it moves around and orients itself appears to be unfocused and aimless, without much energy, never staying long with a particular activity or aspect of the environment.
Sociable	The mouse actively interacts with others. It is willing to interact with others showing affiliative actions (e.g. grooming, resting in groups, sniffing etc.)
Frustrated	The mouse appears unfulfilled with its environment and/or cage mates. It looks stressed and uneasy showing repetitive and fast movements.

Score information

In the Principal Component Analysis (PCA) plot the laboratory facility is located in either the positive arousal/positive emotional valence quadrant or in the negative arousal/positive emotional valence quadrant

In the PCA plot, the laboratory facility is located in positive arousal/negative emotional valance quadrant

In the PCA plot, the laboratory facility is located in negative arousal/negative emotional valance quadrant

BLOODSTAINS IN THE CAGE

Description

Bloodstains might be an indication of fighting behaviour. These stains can be because of fighting and other lesions produced by the environment (e.g. tail trapped) of experimental procedures (e.g. surgery) (Spangenberg and Keeling, 2015).

Scoring method

Observation of the animal's home cage

Score information

- 🟢 Bloodstains absent in the home cage
- 🟠 N/A
- 🔴 Bloodstains present in the home cage

GENERAL INFORMATION

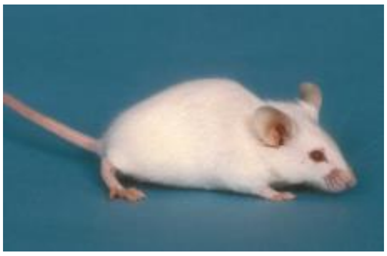
These indicators will be assessed at the individual level. They involve close observation and in some cases, physical restraint of the animals. Resources needed:

- Scoresheet

HUNCHED POSTURE

Description

This posture is recognised by a prominent arched back when the mouse is resting/sitting still. It is considered a reliable indicator of pain, sickness or distress (Baumans *et al.*, 1994; Hawkins, 2002; Paster *et al.*, 2009). See images below for further illustration.



Normal posture

Hunched posture

Scoring method

Observation of the animal in the home cage before any handling.

Score information

- (green) No hunched posture present
- (amber) Walks slowly with hunched posture
- (red) Hunched posture with no movement

COAT CONDITION

Description

The state of the coat in a healthy mouse is usually shiny and well-groomed. When the animal is ill the coat is ruffled and untidy (Paster *et al.*, 2009). See images below for further illustration.

Shiny, clean, and well-groomed coat

Coat clean but ungroomed

Ruffled and untidy coat, it can be greasy and stick together

Scoring method
Observation of the animal in the home cage before any handling.

Score information

- 🟢 Shiny, clean, smooth and well-groomed coat
- 🟠 Coat clean but ungroomed
- 🔴 Ruffled and untidy coat, it can be greasy and stick together

GAIT

Description

This indicator related to how the animal moves around the cage and any change is usually related to injury or pain (Arras *et al.*, 2007). Gait can be used as a measurement of coordination and muscle function (Guyenet *et al.*, 2010). This indicator is assessed by observing the mouse walking around the home cage.

Scoring method
Observation of the animal walking around the home cage before any handling.

Score information

- 🟢 Bodyweight supported on all limbs, with its abdomen not touching the ground, and with both hind limbs participating evenly
- 🟠 The mouse might limp while walking
- 🔴 The mouse has difficulty moving forward or is reluctant to walk and drags its abdomen along the ground

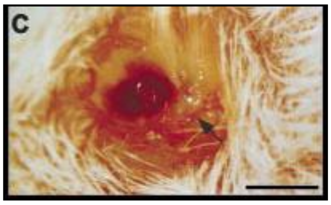
OCULAR/NASAL DISCHARGE

Description

Discharges such as mucus-like or water-like from eyes or nose in a mouse are considered indicative of disease (Hawkins, 2002). The mouse should have their eyes and nose dry and clean from all types of discharge.

No discharge

water-like discharge



mucus-like discharge

Scoring method

Observation of the animal walking around the home cage before any handling.

Score information

- The mouse does not have ocular/nasal discharge
- N/A
- The mouse has water-like or mucus-like ocular/nasal discharge

WOUNDS/MARKS (INCLUDING BITE WOUNDS)

Description

A wound is a lesion to the skin which could be superficial as an area without hair or more profound as a laceration into other tissues (Spangenberg and Keeling, 2015). This indicator includes bite wounds (usually located in the back, flanks, base of the tail, or genitals) and wounds originated other than for conspecifics, e.g. self-inflicted (excessive grooming) or because of environmental problems, management procedures or hanging problems (e.g. tail trapped in the cage). See images below for further illustration.

No wounds

superficial wounds

Extensive and deep wounds

Scoring method
Observation of the animal before handling

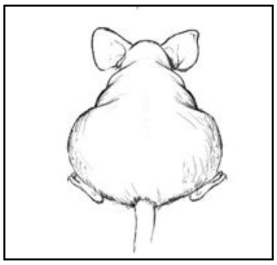
Score information

- 🟢 The mouse does not have wounds/marks
- 🟡 The mouse has wounds which involve superficial tissue (skin), no bleeding and extension is less than 10% of the body.
- 🔴 The mouse has wounds which involve deep tissue (e.g. muscle), bleeding and / or the extension of the wounds is more than 10% of body

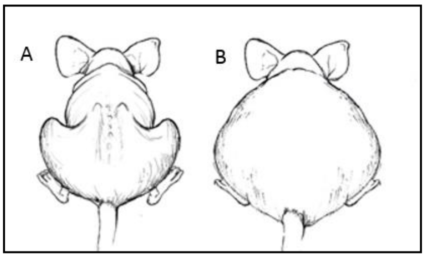
BODY CONDITION SCORE

Description

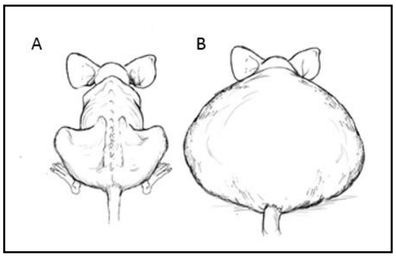
The scoring of the body condition is the process of observing the amount of flesh covering bony protuberances which is mainly in dependant on sex, age, body frame size, and pregnancy status in females (Ullman-Culleré and Foltz, 1999). See diagrams below for further illustration: Mouse body condition score, Modified from (Ullman-Culleré and Foltz, 1999).



Well-conditioned. Vertebral and dorsal pelvis not prominent; palpable with slight pressure



The mouse is either under-conditioned or overweight.
Picture A- under conditioned mouse, the segmentation of vertebral column evident. Dorsal pelvic bones are readily palpable.
Picture B- overweight mouse, the spine is a continuous column. Vertebrae palpable only with firm pressure



The mouse is either emaciated or obese.
Picture A- Emaciated mouse, skeletal structure extremely prominent; little or no flesh covet, vertebrate distinctly segmented.
Picture B- Obese mouse, the mouse is smooth and bulky. Bone structure disappears under flesh and subcutaneous fat.

Scoring method

Observation of the animal in the home cage before any handling.

Score information

- The mouse is well-conditioned. Vertebral and dorsal pelvis not prominent; palpable with slight pressure
- The mouse in either under-conditioned or overweight.
- The mouse is either emaciated or obese

BARBERING (HAIR REMOVAL)

Description

Barbering is considered an abnormal, compulsive behaviour with the risk of pain and welfare compromises if the mouse is barbered (Spangenberg and Keeling, 2015). The behaviour involves the trimming if whiskers and fur anywhere on the body (Spangenberg and Keeling, 2015). See images below for further illustration.

No barbering.

Barbering on different parts of the body

Scoring method
Observation of the animal in the home cage before handling.

Score information

- Mouse does not have trimmed whiskers/ fur
- Mouse has a small zone of fur-trimmed but no whiskers
- The mouse has an extensive area of fur and whiskers trimmed

SWOLLEN ABDOMEN

Description

In a normal healthy mouse, the abdomen will be soft with mild resistance to pressure. If the abdomen is swollen or distended it can be an indication of illness such as a tumour or fluid accumulation (Spangenberg and Keeling, 2015). See images below for further illustration.

Abdomen is not swollen

Abdomen moderately swollen

Abdomen is very swollen

Scoring method

Observation of the animal before handling.

Score information

- The abdomen is not swollen
- The abdomen is moderately swollen
- The abdomen is very swollen

Appendix E. Assessment sheet for the everyday protocol

FACILITY/CAGE ASSESSMENT:

FACILITY:

CAGE ID:

Indicator	Scale	Date				
Temperature	Current room temperature					
Humidity	Current room humidity					
Ventilation	Current room/cage ventilation					
Complexity of the cage	Additional elements or structures present					
	Only nesting material as additional element					
	No additional material or elements					
Alertness	Mouse alert when cage lid is open					
	It might respond to the observer but not immediately.					
	Mouse does not show any interest in the environment or the observer.					
Use of nesting material	Complete or incomplete dome					
	A cup-shaped or flat nest					
	Untouched or scattered nesting material throughout the cage					
Qualitative Behavioural assessment (QBA)	See next page					
Bloodstains in the cage	Bloodstains absent in the home cage					
	Bloodstains present in the home cage					

QUALITATIVE BEHAVIOURAL ASSESSMENT

CAGE ID:

DATE:

Inquisitive	Min.	Max.
In pain	Min.	Max.
Positively Engaged	Min.	Max.
Lethargic	Min.	Max.
Energetic	Min.	Max.
Depressed	Min.	Max.
Determined	Min.	Max.
Anxious	Min.	Max.
Confident	Min.	Max.
Agitated	Min.	Max.
Calm	Min.	Max.
Fearful	Min.	Max.
Content	Min.	Max.
Tense	Min.	Max.
Comfortable	Min.	Max.
Uncertain	Min.	Max.
Playful	Min.	Max.
Bored	Min.	Max.
Sociable	Min.	Max.
Frustrated	Min.	Max.

INDIVIDUAL ASSESSMENT

CAGE ID

MOUSE ID

Indicator	Scale	Date				
Hunched position	No hunched posture present					
	Walks slowly in a hunched posture					
	Hunched posture with no movement					
Coat condition	Shiny, clean and well-groomed coat					
	Coat clean but ungroomed					
	Ruffled and untidy coat, it can be greasy and stick together					
Gait	Bodyweight supported on all limbs					
	Mouse might limp while walking					
	Mouse has difficulty moving forward or is reluctant to walk					
Ocular/nasal discharge	No ocular/nasal discharge					
	Water-like or mucus-like ocular/nasal discharge					
Wounds/marks (including bite wounds)	No wounds					
	Superficial wounds					
	Extensive and deep wounds					
Body Condition Score	The mouse is well-conditioned					
	The mouse in either under-conditioned or overweight					
	The mouse is either emaciated or obese.					
Barbering (hair removal)	No barbering					
	Barbering on different parts of the body					
Swollen abdomen	Abdomen is not swollen					
	Abdomen moderately swollen					
	Abdomen is very swollen					

COMMENTS:

Appendix F. Protocol for an everyday laboratory mouse welfare assessment

ASSESSMENT AT FACILITY/ROOM LEVEL

GENERAL INFORMATION

These indicators will be assessed at the facility/room level. They involve an interview with the technical manager of the laboratory facility and observations of the environmental conditions of the animal holding room. Resources required are:

- Scoresheet
- Humidity meter
- Thermometer

ROOM TEMPERATURE

Description

Room temperature for laboratory mice is usually maintained at 20°C to 24°C as suggested by the Code of Practice for Housing and Care of Laboratory Animals (Home Office, 2014a). However, recent scientific recommendations have suggested that the thermoneutral mouse zone is 26°C to 29°C (Gaskill *et al.*, 2009). To reduce the need for an increase in laboratory ambient temperature, the provision of adequate nesting material in the home cage can alleviate thermal distress (Gaskill *et al.*, 2012).

Scoring method

Thermometer to record the current room temperature

Score information

Green	Room temperature from 24°C to 29°C
Amber	Room temperature from 24°C to 20°C
Red	Room temperature below 24°C

HUMIDITY

Description

As a general rule, the humidity in a mouse room should be kept at 40 to 60% (Home Office, 2014a). This indicator is measured at room level and records should be kept to avoid any sudden or prolonged periods below 40% or above 60% (Home Office, 2014a). High levels of humidity are usually associated with health problems (Chesler *et al.*, 2002) (e.g. reproduction) and behavioural changes, e.g. (Nelson *et al.*, 1990).

Scoring method

Humidity meter to record the current room humidity

Score information

Green	Room humidity 40%-60%
Amber	Room humidity outside (+ or -) the recommended (40% - 60%) by 1% - 10%
Red	Room humidity outside (+ or -) the recommended (40% - 60%) by more than 11%

VENTILATION

Description

The purpose of ventilation in an animal room is to provide good quality air and to minimise the levels of gases, odours and infectious agents. It is advisable to have 10 to 20 air changes per hour in a fully stocked room (Reeb-Whitaker *et al.*, 2001). In most of the facilities, this ventilation is controlled at cage level when IVCs (Individual Ventilated Cages) are used (AAALAC, 2011).

Scoring method

Ventilation information from facility records or IVC (individual Ventilated Cages) in the room

Score information

- (Green) Room/IVCs ventilation 10 to 20 changes per hour
- (Amber) Room/IVCs ventilation outside (+ or -) the recommended (10 to 20 air changes per hour) by 1 – 10%
- (Red) Room/IVCs ventilation outside (+ or -) the recommended (10 to 20 air changes per hour) by more than 11%

GENERAL INFORMATION

These indicators will be assessed at cage level. They involve taking the cage from the rack, placing it the designated area where the assessment will be carried out, opening the cage and removing the food/water hopper. Resources needed:

- Scoresheet

COMPLEXITY OF THE CAGE

Description

Cage complexity refers to the environmental enrichment that is provided within the cage on top of the basic housing requirements for the laboratory mouse (Olsson and Dahlborn, 2002). Mice require additional enrichment in the cage to be able to cope with stressors and perform natural behaviours (Olsson and Dahlborn, 2002). Structures that increase the complexity of the cage, such as material for building a shelter or a nest, additional pellets or seeds as well as chewing material among others are recommended (Olsson and Dahlborn, 2002).

Scoring method

Observation of the animal's home cage, distribution and structures /elements present in the cage.

Score information

- 🟢 Cage is provided with additional elements or structures (e.g. nesting material, wood stick)
- 🟡 Cage contains only nesting material as an additional element
- 🔴 Cage does not contain any additional elements or structures

ALERTNESS

Description

Mice are active and alert animals by nature (Lloyd and Wolfensohn, 1999. They should show interest in the environment, wanting to explore and be attentive. Alertness should be observed taking into account the circadian cycle as is normal to find the mouse sleep and resting during daylight hours. The animal should be observed as soon the case is open for 60 seconds (Lloyd and Wolfensohn, 1999).

Scoring method

Observation of the animal in the home cage when the cage is initially opened

Score information

- 🟢 The mouse is alert when the cage lid is open (focus its attention towards the observer, may stand up if it is sleeping and sniff towards the observer)
- 🟡 The mouse is less alert. It might respond to the observer but not immediately.
- 🔴 The mouse is immobile. It does not show any interest in the environment or the observer.

USAGE OF NESTING MATERIAL

Description

The usage of nesting material to build a nest in the home cage is related to the mouse-specific behaviour used for controlling the environment and thermoregulation (Hess *et al.*, 2008). This indicator assesses the type of nest built from the available nesting material. The scoring system is modified from Hess *et al.* (2008).

Score 0

(A) Completed dome. The walls completely enclose the nest hollow. A small exit hole may be found on the side or on the top of the nest.

(B) Incomplete dome: bedding material has been gathered to from a nest. The walls reach the widest point of an imaginary sphere that will fit the nest hollow

Score 1

(A) cup: The nest has identifiable walls that form a "cup". These walls would not reach the widest point of an imaginary sphere that would fill the nest hollow.

(B) Flat nest: the nesting material has been gathered for form a nest, identified by a clear nest cavity in the middle of the material or between the material and the cage wall. The nest is flat with no or incomplete walls

Score 2

Untouched or scattered: the nesting material has not been moved or it has been scattered throughout the cage.

Scoring method

Observation of the nest in the home cage when the cage is open

Score information

- Complete or incomplete dome
- A cup-shaped or flat nest
- Untouched or scattered nesting material throughout the cage

QUALITATIVE BEHAVIOURAL ASSESSMENT (QBA)

Description

QBA in an indicator that uses the ability that people have to integrate various details in animals behaviour, postures and context into descriptions of animal "body language" using descriptors such as "relaxed", "tense" or "content". These terms have an emotional connotation which is relevant to animal welfare (Wemelsfelder, 2007).

Scoring method

This indicator is assessed by using a visual analogue scale for each term. The assessment is made at the cage level taking into account the group of animals as a whole. The cage should be left undisturbed with the observer close to it for 5 minutes, after, observation for one minute is made and then, the scales are used for each term in the order they are presented.

This indicator uses a Visual Analogue Scale to score each descriptor where:

0 mm is the minimum (the term is absent) and

125mm is the maximum (the term is entirely present).

TERM	
Inquisitive	The mouse appears curious and interested in others and in exploring the environment. Willing to investigate.
In Pain	The mouse is suffering from physical discomfort leading the mouse to be reluctant to move, or to move with abnormal gait, or showing a tense, hunched or uncomfortable posture. The mouse looks like it is hurting or suffering and is in discomfort.
Positively engaged	The mouse is carrying out activities in a focused, directed and constructive manner. The mouse appears not to be distracted by others or the environment.
Lethargic	The mouse appears fatigued and sluggish. It has a lack of vigour and energy, showing low amounts of movement and any movement is slow and ponderous
Energetic	The mouse is carrying out an activity with a lot of energy and vigour, in a lively and excited manner.
Depressed	The mouse appears unhappy and without hope. It is apathetic, despondent and unresponsive showing little or no response or reaction to anyone or its environment. It appears isolated.
Determined	The mouse is showing an active and rapid reaction to something or someone. It appears to be focused on accomplishing a specific goal or task.
Anxious	The mouse is uneasy, cautious and nervous
Confident	The mouse is displaying assertiveness, behaving assertively with other animals and its environment in a self-assured manner.

Agitated	The mouse appears to be irritable and highly reactive. An excess of physical and/or cognitive activity is present because of anxiety.
Calm	The mouse appears peaceful and without worry. The mouse behaves in a relaxed and untroubled manner.
Fearful	The mouse appears afraid or scared. It seems to be avoiding contact with others and the environment, looks to be hiding, looking for a way out or trying to escape.
Content	The mouse appears happy and satisfied. Expressing happiness, with all its physiological, environmental and psychological needs met.
Tense	The mouse looks worried and emotionally tense. Its posture might evidence physical tension.
Comfortable	The mouse appears physically satisfied with the cage environment and looks relaxed and free from discomfort.
Uncertain	The mouse appears to be insecure; its physical movement is cautious. The mouse moves slowly showing alertness and insecurity. Avoidance reactions are showing with all stimuli
Playful	The mouse is engaging in lively movements purely to frolic or for fun, expressing pleasure, happiness and amusement.
Bored	The mouse appears uninterested in its environment and cage mates. The way it moves around and orients itself appears to be unfocused and aimless, without much energy, never staying long with a particular activity or aspect of the environment.
Sociable	The mouse actively interacts with others. It is willing to interact with others showing affiliative actions (e.g. grooming, resting in groups, sniffing etc.)
Frustrated	The mouse appears unfulfilled with its environment and/or cage mates. It looks stressed and uneasy showing repetitive and fast movements.

Score information

In the Principal Component Analysis (PCA) plot the laboratory facility is located in either the positive arousal/positive emotional valence quadrant or in the negative arousal/positive emotional valence quadrant

In the PCA plot, the laboratory facility is located in positive arousal/negative emotional valance quadrant

In the PCA plot, the laboratory facility is located in negative arousal/negative emotional valance quadrant

BLOOD STAINS IN THE CAGE

Description

Bloodstains might be an indication of fighting behaviour. These stains can be because of fighting and other lesions produced by the environment (e.g. tail trapped) of experimental procedures (e.g. surgery) (Spangenberg and Keeling, 2015).

Scoring method

Observation of the animal's home cage

Score information

- 🟢 Bloodstains absent in the home cage
- 🟡 N/A
- 🔴 Bloodstains present in the home cage

GENERAL INFORMATION

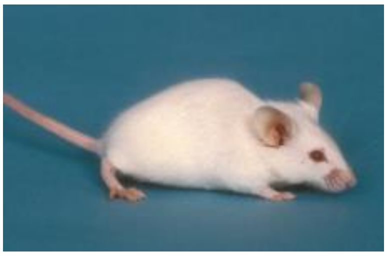
These indicators will be assessed at individual level. They involve close observation and physical restraint of the animals. Resources needed:

- Scoresheet

HUNCHED POSITION

Description

This posture is recognised by a prominent arched back when the mouse is resting/sitting still. It is considered a strong indicator of pain, sickness or distress (Baumans *et al.*, 1994; Hawkins, 2002; Paster *et al.*, 2009). See images below for further illustration.



Normal posture

Hunched posture

Scoring method

Observation of the animal in the home cage before any handling.

Score information

- (green) No hunched posture present
- (amber) Walks slowly with no hunched posture
- (red) Hunched posture with no movement

COAT CONDITION

Description

The state of the coat in a healthy mouse is usually shiny and well-groomed. When the animal is ill the coat is ruffled and untidy (Paster *et al.*, 2009). See images below for further illustration.

Shiny, clean, and well-groomed coat

Coat clean but ungroomed

Ruffled and untidy coat, it can be greasy and stick together

Scoring method
Observation of the animal in the home cage before any handling.

Score information

- 🟢 Shiny, clean, smooth and well-groomed coat
- 🟡 Coat clean but ungroomed
- 🔴 Ruffled and untidy coat, it can be greasy and stick together

GAIT

Description

This indicator related to how the animal moves around the cage and any change is usually related to injury or pain (Arras *et al.*, 2007). Gait can be used as a measurement of coordination and muscle function (Guyenet *et al.*, 2010). This indicator is assessed by observing the mouse walking around the home cage.

Scoring method
Observation of the animal walking around home cage before any handling.

Score information

- 🟢 Bodyweight supported on all limbs, with its abdomen not touching the ground, and with both hind limbs participating evenly
- 🟡 The mouse might limp while walking
- 🔴 The mouse has difficulty moving forward or is reluctant to walk and drags its abdomen along the ground

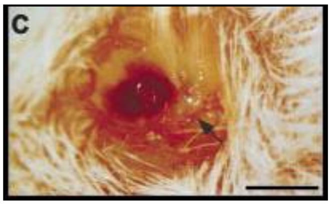
OCULAR/NASAL DISCHARGE

Description

Discharges such as mucus-like or water-like from eyes or nose in a mouse are considered indicative of disease (Hawkins, 2002). The mouse should have their eyes and nose dry and clean from all types of discharge.

No discharge

water-like discharge



mucus-like discharge

Scoring method

Observation of the animal walking around the home cage before any handling.

Score information

- (green) The mouse does not have ocular/nasal discharge
- (amber) N/A
- (red) The mouse has water-like or mucus-like ocular/nasal discharge

WOUNDS/MARKS (INCLUDING BITE WOUNDS)

Description

A wound is a lesion to the skin which could be superficial as an area without hair or more profound as a laceration into other tissues (Spangenberg and Keeling, 2015). This indicator includes bite wounds (usually located in the back, flanks, base of the tail, or genitals) and wounds originated other than for conspecifics, e.g. self-inflicted (excessive grooming) or because of environmental problems, management procedures or hanging problems (e.g. tail trapped in the cage). See images below for further illustration.


No wounds


superficial wounds


Extensive and deep wounds

Scoring method
Observation of the animal before handling

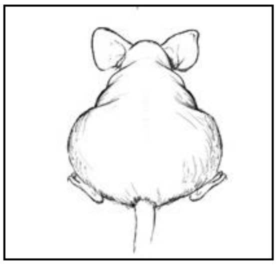
Score information

- (green) The mouse does not have wounds/marks
- (amber) The mouse has wounds which involve superficial tissue (skin), no bleeding and extension is less than 10% of the body.
- (red) The mouse has wounds which involve deep tissue (e.g. muscle), bleeding and the extension of the wounds is more than 10% of body

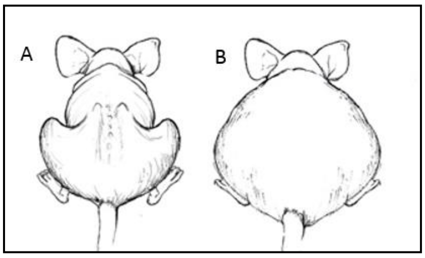
BODY CONDITION SCORE

Description

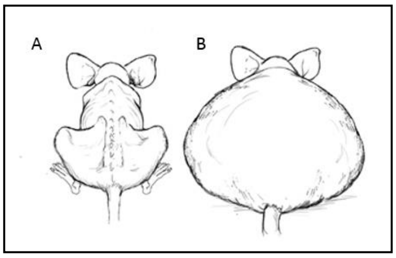
The scoring of the body condition is the process of observing the amount of flesh covering bony protuberances which is mainly in dependant on sex, age, body frame size, and pregnancy status in females (Ullman-Culleré and Foltz, 1999). See diagrams below for further illustration: Mouse body condition score, Modified from (Ullman-Culleré and Foltz, 1999).



Well-conditioned. Vertebral and dorsal pelvis not prominent; palpable with slight pressure



The mouse is either under-conditioned or overweight.
Picture A- under conditioned mouse, the segmentation of vertebral column evident. Dorsal pelvic bones are readily palpable.
Picture B- overweight mouse, the spine is a continuous column. Vertebrae palpable only with firm pressure



The mouse is either emaciated or obese.
Picture A- Emaciated mouse, skeletal structure extremely prominent; little or no flesh covet, vertebrate distinctly segmented.
Picture B- Obese mouse, the mouse is smooth and bulky. Bone structure disappears under flesh and subcutaneous fat.

Scoring method
Observation of the animal in the home cage before any handling.

Score information

- 🟢 The mouse is well-conditioned. Vertebral and dorsal pelvis not prominent; palpable with slight pressure
- 🟡 The mouse in either under-conditioned or overweight.
- 🔴 The mouse is either emaciated or obese

BARBERING

Description

Barbering is considered an abnormal, compulsive behaviour with the risk of pain and welfare compromises if the mouse is barbered (Spangenberg and Keeling, 2015). The behaviour involves the trimming if whiskers and fur anywhere on the body (Spangenberg and Keeling, 2015). See images below for further illustration.

No barbering.

Barbering on different parts of the body

Scoring method
Observation of the animal in the home cage before handling.

Score information

- The mouse does not have trimmed whiskers/ fur
- The mouse has a small zone of fur-trimmed but no whiskers
- The mouse has an extensive area of fur and whiskers trimmed

SWOLLEN ABDOMEN

Description

In a normal healthy mouse, the abdomen will be soft with mild resistance to pressure. If the abdomen is swollen or distended it can be an indication of illness such as a tumour or fluid accumulation (Spangenberg and Keeling, 2015). See images below for further illustration.

Abdomen is not swollen

Abdomen moderately swollen

Abdomen is very swollen

Scoring method
Observation of the animal before handling.

Score information

- The abdomen is not swollen
- The abdomen is moderately swollen
- The abdomen is very swollen

References

AAALAC (2011) *Guide for the care and use of laboratory animals*. 8 Edition edn. Association for Assessment and Accreditation of Laboratory Animal Care Internacional.

Adler, M. and Ziglio, E. (1996) *Gazing into the oracle: The Delphi method and its application to social policy and public health*. Jessica Kingsley Publishers.

Akintola, T., Raver, C., Studlack, P., Uddin, O., Masri, R. and Keller, A. (2017) 'The grimace scale reliably assesses chronic pain in a rodent model of trigeminal neuropathic pain', *Neurobiology of Pain*, 2, pp. 13-17.

Aksoy, B. and Orbak, Â.Y. (2009) 'Reducing the quantity of reworked parts in a robotic arc welding process', *Quality and Reliability Engineering International*, 25(4), pp. 495-512.

Algers, B. (2012) 'Animal welfare–recent developments in the field', *Animal Science Reviews 2011*, p. 19.

Anders, S., Lotze, M., Erb, M., Grodd, W. and Birbaumer, N. (2004) 'Brain activity underlying emotional valence and arousal: A response-related fMRI study', *Human brain mapping*, 23(4), pp. 200-209.

Andreasen, S.N., Wemelsfelder, F., Sandøe, P. and Forkman, B. (2013) 'The correlation of Qualitative Behavior Assessments with Welfare Quality® protocol outcomes in on-farm welfare assessment of dairy cattle', *Applied Animal Behaviour Science*, 143(1), pp. 9-17.

Anil, S.S., Anil, L. and Deen, J. (2002) 'Challenges of pain assessment in domestic animals', *Journal of the American Veterinary Medical Association*, 220(3), pp. 313-319.

Animal and Plant Health Inspection Service (2015) *Annual Report Animal Usage by Fiscal Year 2016*. Available at: https://speakingofresearch.files.wordpress.com/2008/03/usda-annual-report-animal-usage-in-research-2016.pdf.

Antonacopoulos, N.M.D. and Pychyl, T.A. (2010) 'An examination of the potential role of pet ownership, human social support and pet attachment in the psychological health of individuals living alone', *Anthrozoös*, 23(1), pp. 37-54.

Appleby, M.C. (2011) *Animal welfare*. 2nd edn. Wallingford, Oxfordshire, UK ; Cambridge, MA: CABI.

Appleby, M.C. and Hughes, B.O. (1997) *Animal welfare*. New York, NY: CAB International.

Arras, M., Autenried, P., Rettich, A., Spaeni, D. and Rülicke, T. (2001) 'Optimization of intraperitoneal injection anesthesia in mice: drugs, dosages, adverse effects, and anesthesia depth', *Comparative medicine*, 51(5), pp. 443-456.

Arras, M., Rettich, A., Cinelli, P., Kasermann, H.P. and Burki, K. (2007) 'Assessment of post-laparotomy pain in laboratory mice by telemetric recording of heart rate and heart rate variability', *BMC veterinary research*, 3(1), p. 16.

Asif T. Chinwalla et al. , L.L.C., Kimberly D. Delehaunty1, Ginger A. Fewell1, Lucinda A. Fulton1,47, et al. (2002) 'Initial sequencing and comparative analysis of the mouse genome', *Nature*, 420(6915), pp. 520-562.

Association of Veterinary, T. and Research, W. (1989) *Guidelines for the recognition and assessment of pain in animals*. Univ. Federation for Animal Welfare.

Augustsson, H. and Meyerson, B.J. (2004) 'Exploration and risk assessment: a comparative study of male house mice (Mus musculus musculus) and two laboratory strains', *Physiology & Behavior*, 81(4), pp. 685-698.

Balcombe, J. (2009) 'Animal pleasure and its moral significance', *Applied Animal Behaviour Science*, 118(3–4), pp. 208-216.

Balcombe, J.P. (2006) 'Laboratory environments and rodents' behavioural needs: a review', *Laboratory animals*, 40(3), pp. 217-235.

Balcombe, J.P., Barnard, N.D. and Sandusky, C. (2004) 'Laboratory routines cause animal stress', *Journal of the American Association for Laboratory Animal Science*, 43(6), pp. 42-51.

Banks, E.M. (1982) 'Behavioral research to answer questions about animal welfare', *Journal of Animal Science*, 54(2), pp. 434-446.

Barnett, J.L. and Hemsworth, P.H. (1990) 'The validity of physiological and behavioural measures of animal welfare', *Applied Animal Behaviour Science*, 25(1), pp. 177-187.

Barrett, L.F. (1998) 'Discrete emotions or dimensions? The role of valence focus and arousal focus', *Cognition & Emotion*, 12(4), pp. 579-599.

Bateson, M., Desire, S., Gartside, Sarah E. and Wright, Geraldine A. (2011) 'Agitated Honeybees Exhibit Pessimistic Cognitive Biases', *Current Biology*, 21(12), pp. 1070-1073.

Bateson, M. and Nettle, D. (2015) *Development of a cognitive bias methodology for measuring depression-like mood in chimpanzees* (2167-9843). PeerJ PrePrints.

Bateson, P. (1991) 'Assessment of pain in animals', *Animal Behaviour*, 42(5), pp. 827-839.

Bateson, P. (2004) 'Do animals suffer like us? – the assessment of animal welfare', *The Veterinary Journal*, 168(2), pp. 110-111.

Battini, M., Stilwell, G., Vieira, A., Barbieri, S., Canali, E. and Mattiello, S. (2015) 'On-Farm Welfare Assessment Protocol for Adult Dairy Goats in Intensive Production Systems', *Animals : an Open Access Journal from MDPI*, 5(4), pp. 934-950.

Baumans, V. (2004) *The welfare of laboratory mice in The Welfare of Laboratory Animals*. Springer,. Available at: http://www.NCL.eblib.com/patron/FullRecord.aspx?p=198029.

Baumans, V. (2005a) 'Environmental enrichment for laboratory rodents and rabbits: requirements of rodents, rabbits, and research', *ILAR journal*, 46(2), pp. 162-170.

Baumans, V. (2005b) 'Science-based assessment of animal welfare: laboratory animals', *Revue scientifique et technique (International Office of Epizootics)*, 24(2), pp. 503-513.

Baumans, V., Brain, P.E., Brugère, H., Clausing, P., Jeneskog, T. and Perretta, G. (1994) 'Report of the FELASA Working Group on pain and distress', *Lab Anim*, 28, pp. 97-112.

Bekoff, M. (2009) 'Animal emotions, wild justice and why they matter: Grieving magpies, a pissy baboon, and empathic elephants', *Emotion, Space and Society*, 2(2), pp. 82-85.

Bennett, R.M., Broom, D.M., Henson, S.J., Blaney, S.J.P. and Harper, G. (2004) 'Assessment of the impact of government animal welfare policy on farm animal welfare in the UK', *Animal Welfare*, 13(1), pp. 1-11.

Berry, A., Bellisario, V., Capoccia, S., Tirassa, P., Calza, A., Alleva, E. and Cirulli, F. (2012) 'Social deprivation stress is a triggering factor for the emergence of anxiety- and depression-like behaviours and leads to reduced brain BDNF levels in C57BL/6J mice', *Psychoneuroendocrinology*, 37(6), pp. 762-772.

Black, T.R. (1999) *Doing quantitative research in the social sciences: An integrated approach to research design, measurement and statistics*. Sage.

Blank, R. (2004) *The Basics of Reliability*. SteinerBooks.

Blokhuis, H.J. (2013) *Improving farm animal welfare: science and society working together: the welfare quality approach*. Springer.

Blokhuis, H.J., Veissier, I., Miele, M. and Jones, B. (2010) 'The Welfare Quality® project and beyond: Safeguarding farm animal well-being', *Acta Agriculturae Scand Section A*, 60(3), pp. 129-140.

Blom, H., Van Tintelen, G., Van Vorstenbosch, C., Baumans, V. and Beynen, A. (1996a) 'Preferences of mice and rats for types of bedding material', *Laboratory Animals*, 30(3), pp. 234-244.

Blom, H.J.M., Van Tintelen, G., Van Vorstenbosch, C., Baumans, V. and Beynen, A.C. (1996b) 'Preferences of mice and rats for types of bedding material', *Laboratory Animals*, 30(3), pp. 234-244.

Blom, H.J.M., Van Vorstenbosch, C., Baumans, V., Hoogervorst, M.J.C., Beynen, A.C. and Van Zutphen, L.F.M. (1992) 'Description and validation of a preference test system to evaluate housing conditions for laboratory mice', *Applied Animal Behaviour Science*, 35(1), pp. 67-82.

Boissy, A., Arnould, C., Chaillou, E., Désiré, L., Duvaux-Ponter, C., Greiveldinger, L., Leterrier, C., Richard, S., Roussel, S. and Saint-Dizier, H. (2007a) 'Emotions and cognition: a new approach to animal welfare', *Animal welfare*, 16(Supplement 1), pp. 37-43.

Boissy, A., Manteuffel, G., Jensen, M.B., Moe, R.O., Spruijt, B., Keeling, L.J., Winckler, C., Forkman, B., Dimitrov, I., Langbein, J., Bakken, M., Veissier, I. and Aubert, A. (2007b) 'Assessment of positive emotions in animals to improve their welfare', *Physiol Behav*, 92(3), pp. 375-97.

Bolboaca, S.-D. and Jäntschi, L. (2006) 'Pearson versus Spearman, Kendall’s tau correlation analysis on structure-activity relationships of biologic active compounds', *Leonardo Journal of Sciences*, 5(9), pp. 179-200.

Botreau, R., Bonde, M., Butterworth, A., Perny, P., Bracke, M.B.M., Capdeville, J. and Veissier, I. (2007a) 'Aggregation of measures to produce an overall assessment of animal welfare. Part 1: a review of existing methods'.

Botreau, R., Veissier, I., Butterworth, A., Bracke, M.B.M. and Keeling, L.J. (2007b) 'Definition of criteria for overall assessment of animal welfare', *ANIMAL WELFARE-POTTERS BAR THEN WHEATHAMPSTEAD-*, 16(2), p. 225.

Botreau, R., Veissier, I. and Perny, P. (2009) 'Overall assessment of animal welfare: strategy adopted in Welfare Quality®', *Animal welfare (South Mimms, England)*, (18), pp. 363-370.

Boursot, P., Auffray, J.C., Britton-Davidian, J. and Bonhomme, F. (1993) 'The evolution of house mice', *Annual Review of Ecology and Systematics*, pp. 119-152.

Bowling, A. (2005) *Techniques of questionnaire design*. Open University Press, Maidenhead, UK.

Braastad, B.O. (1998) 'Effects of prenatal stress on behaviour of offspring of laboratory and farmed mammals', *Applied Animal Behaviour Science*, 61(2), pp. 159-180.

Bracke, M.B.M. (2006) 'Expert opinion regarding environmental enrichment materials for pigs', *Animal Welfare*, 15(1), pp. 67-70.

Bracke, M.B.M., Metz, J.H.M. and Spruijt, B.M. (1999) 'Overall animal welfare reviewed. Part 2: Assessment tables and schemes', *NJAS wageningen journal of life sciences*, 47(3), pp. 293-305.

Bradley, M.M. and Lang, P.J. (1994) 'Measuring emotion: the self-assessment manikin and the semantic differential', *Journal of behavior therapy and experimental psychiatry*, 25(1), pp. 49-59.

Brambell, F.W.R. (1965) *Report of the Technical Committee to Enquire Into the Welfare of Animals Kept Under Intensive Livestock Husbandry Systems: Presented to Parliament by the Secretary of State for Scotland and the Minister of Agriculture, Fisheries and Food by Command of Her Majesty December, 1965*. HM Stationery Office.

Branchi, I., Santucci, D. and Alleva, E. (2001) 'Ultrasonic vocalisation emitted by infant rodents: a tool for assessment of neurobehavioural development', *Behavioural Brain Research*, 125(1–2), pp. 49-56.

Brilot, B.O., Asher, L. and Bateson, M. (2010) 'Stereotyping starlings are more ‘pessimistic'', *Animal Cognition*, 13(5), pp. 721-731.

Broom, D.M. (1986) 'Indicators of poor welfare', *British veterinary journal*, 142(6), pp. 524-526.

Broom, D.M. (1988) 'The scientific assessment of animal welfare', *Applied Animal Behaviour Science*, 20(1–2), pp. 5-19.

Broom, D.M. (1991a) 'Assessing welfare and suffering', *Behavioural Processes*, 25(2–3), pp. 117-123.

Broom, D.M. (1993) *Stress and animal welfare*. Springer Science & Business Media.

Broom, D.M. (1996) 'Animal welfare defined in terms of attempts to cope with the environment', *Acta Agriculturae Scandinavica. Section A. Animal Science. Supplementum (Denmark)*.

Broom, D.M. (2007) 'Quality of life means welfare: how is it related to other concepts and assessed?', *Animal Welfare* 16, p. 45.

Broom, D.M. (2010) 'Cognitive ability and awareness in domestic animals and decisions about obligations to animals', *Applied Animal Behaviour Science*, 126(1–2), pp. 1-11.

Broom, D.M. (2011) 'A history of animal welfare science', *Acta Biotheoretica*, 59(2), pp. 121-137.

Brown, J.D. (2008) 'Effect size and eta squared', *JALT Testing & Evaluation SIG News*.

Brscic, M., Wemelsfelder, F., Tessitore, E., Gottardo, F., Cozzi, G. and Van Reenen, C. (2009) 'Welfare assessment: correlations and integration between a Qualitative Behavioural Assessment and a clinical/health protocol applied in veal calves farms', *Italian Journal of Animal Science*, 8(sup2), pp. 601-603.

Buchanan-Smith, H.M., Rennie, A., Vitale, A., Pollo, S., Prescott, M.J. and Morton, D.B. (2005) 'Harmonising the definition of refinement'.

Bushman, L. and Schnitker, B. (1995) 'Teacher Attitudes on Portfolio Assessment, Implementation, and Practicability'.

Cardinal, R.N., Parkinson, J.A., Hall, J. and Everitt, B.J. (2002) 'Emotion and motivation: the role of the amygdala, ventral striatum, and prefrontal cortex', *Neuroscience & Biobehavioral Reviews*, 26(3), pp. 321-352.

Carter, C.S. (2001) 'Is there a neurobiology of good welfare', *Coping with Challenge: Welfare of Animals including Humans. Broom DM (ed.), vols*, pp. 11-30.

Caterina, M.J., Leffler, A., Malmberg, A., Martin, W., Trafton, J., Petersen-Zeitz, K., Koltzenburg, M., Basbaum, A. and Julius, D. (2000) 'Impaired nociception and pain sensation in mice lacking the capsaicin receptor', *science*, 288(5464), pp. 306-313.

Champy, M.F., Selloum, M., Piard, L., Zeitler, V., Caradec, C., Chambon, P. and Auwerx, J. (2004) 'Mouse functional genomics requires standardization of mouse handling and housing conditions', *Mammalian genome*, 15(10), pp. 768-783.

Chesler, E.J., Wilson, S.G., Lariviere, W.R., Rodriguez-Zas, S.L. and Mogil, J.S. (2002) 'Identification and ranking of genetic and laboratory environment factors influencing a behavioral trait, thermal nociception, via computational analysis of a large data archive', *Neuroscience & Biobehavioral Reviews*, 26(8), pp. 907-923.

Clarkson, J.M., Dwyer, D.M., Flecknell, P.A., Leach, M.C. and Rowe, C. (2018) 'Handling method alters the hedonic value of reward in laboratory mice', *Scientific reports*, 8(1), p. 2448.

Clough, G. (1982) 'Environmental effects on animals used in biomedical research', *Biological Reviews*, 57(3), pp. 487-523.

Cohen, C. (1997) 'Do animals have rights?', *Ethics & behavior*, 7(2), pp. 91-102.

Collins, A.J. (1991) 'The use of generalised procrustes techniques in sensory analysis', *Journées Agro-Industries et Méthodes Statistiques, Nantes, France*.

Collins, J., Hanlon, A., More, S.J., Wall, P.G. and Duggan, V. (2009) 'Policy Delphi with vignette methodology as a tool to evaluate the perception of equine welfare', *The Veterinary Journal*, 181(1), pp. 63-69.

Council, F.A.W. (2009a) *Farm animal welfare in Great Britain: Past, present and future*. Farm Animal Welfare Council.

Council, N.F.A.C. (2009b) *Code of Practice for the Care and Handling of Dairy Cattle*. Dairy Farmers of Canada.

Croney, C.C. and Millman, S.T. (2007) 'Board-invited review: The ethical and behavioral bases for farm animal welfare legislation', *Journal of animal science*, 85(2), pp. 556-565.

Dai, F., Dalla Costa, E., Murray, L.M.A., Canali, E. and Minero, M. (2016) 'Welfare conditions of donkeys in Europe: initial outcomes from On-farm assessment', *Animals*, 6(1), p. 5.

David, J.M., Knowles, S., Lamkin, D.M. and Stout, D.B. (2013) 'Individually ventilated cages impose cold stress on laboratory mice: a source of systemic experimental variability', *Journal of the American Association for Laboratory Animal Science*, 52(6), pp. 738-744.

Dawkins, M.S. (1983) 'Cage size and flooring preferences in litter-reared and cage-reared hens', *Br Poult Sci*, 24(2), pp. 177-82.

Dawkins, M.S. (1985) 'Cage height preference and use in battery-kept hens', *Vet Rec*, 116(13), pp. 345-7.

Dawkins, M.S. (1988) 'Behavioural deprivation: A central problem in animal welfare', *Applied Animal Behaviour Science*, 20(3–4), pp. 209-225.

Dawkins, M.S. (1990) 'From an animal's point of view: motivation, fitness, and animal welfare', *Behavioral and brain sciences*, 13(01), pp. 1-9.

Dawkins, M.S. (1998) 'Evolution and animal welfare', *Quarterly Review of Biology*, pp. 305-328.

Dawkins, M.S. (2003) 'Behaviour as a tool in the assessment of animal welfare1', *Zoology*, 106(4), pp. 383-387.

Dawkins, M.S. (2006) 'Through animal eyes: What behaviour tells us', *Applied Animal Behaviour Science*, 100(1–2), pp. 4-10.

Dawkins, M.S., Dawkins, M.S., Dawkins, M.S., Zoologiste, G.-B., Dawkins, M.S. and Zoologist, G.B. (1993) *Through our eyes only?: the search for animal consciousness*. WH Freeman Oxford.

Dawkins, M.S., Donnelly, C.A. and Jones, T.A. (2004) 'Chicken welfare is influenced more by housing conditions than by stocking density', *Nature*, 427(6972), pp. 342-4.

Deacon, R.M. (2006) 'Housing, husbandry and handling of rodents for behavioral experiments', *Nature protocols*, 1(2), p. 936.

DeGrazia, D. and Rowan, A. (1991) 'Pain, suffering, and anxiety in animals and humans', *Theoretical Medicine*, 12(3), pp. 193-211.

Der Fels-Klerx, V., Ine, H.J., Goossens, L.H.J., Saatkamp, H.W. and Horst, S.H.S. (2002) 'Elicitation of quantitative data from a heterogeneous expert panel: formal process and application in animal health', *Risk Analysis*, 22(1), pp. 67-81.

Désiré, L., Boissy, A. and Veissier, I. (2002) 'Emotions in farm animals:: a new approach to animal welfare in applied ethology', *Behavioural processes*, 60(2), pp. 165-180.

Donnelley, S. (1989) 'Animals, science, and ethics--Section II. Animals in science: the justification issue', *The Hastings Center Report*, 20(3), pp. S8-13.

Douglas-Hamilton, I., Bhalla, S., Wittemyer, G. and Vollrath, F. (2006) 'Behavioural reactions of elephants towards a dying and deceased matriarch', *Applied Animal Behaviour Science*, 100(1–2), pp. 87-102.

Duncan, I. (2005a) 'Science-based assessment of animal welfare: farm animals', *Revue scientifique et technique-Office international des epizooties*, 24(2), p. 483.

Duncan, I.J. and Petherick, J.C. (1991) 'The implications of cognitive processes for animal welfare', *Journal of Animal Science*, 69(12), pp. 5017-22.

Duncan, I.J.H. (2005b) 'Science-based assessment of animal welfare: farm animals', *Revue scientifique et technique-Office international des epizooties*, 24(2), p. 483.

E Marian, D. and P Shimamura, A. (2013) *Contextual Influences on Dynamic Facial Expressions*.

Eaton, L.G. and Funder, D.C. (2001) 'Emotional experience in daily life: Valence, variability, and rate of change', *Emotion*, 1(4), p. 413.

European Commission (2013) 'REPORT FROM THE COMMISSION TO THE COUNCIL AND THE EUROPEAN PARLIAMENT Seventh Report on the Statistics on the Number of Animals used for Experimental and other Scientific Purposes in the Member States of the European Union'.

European Union (1997) *Treaty of Amsterdam*. Luxembourg: Office for Official Publications of the European Communities.

Falzarano, M. and Zipp, G.P. (2013) 'Seeking consensus through the use of the Delphi technique in health sciences research', *Journal of allied health*, 42(2), pp. 99-105.

Farm Animal Welfare Council (1992) *The five freedoms*. Farm Animal Welfare Council London.

Feaver, J., Mendl, M. and Bateson, P. (1986) 'A method for rating the individual distinctiveness of domestic cats', *Animal Behaviour*, 34(4), pp. 1016-1025.

Feldman, L.A. (1995) 'Valence focus and arousal focus: Individual differences in the structure of affective experience', *Journal of personality and social psychology*, 69(1), p. 153.

Flecknell, P. (2002) 'Replacement, reduction and refinement', *Altex*, 19(2), pp. 73-78.

Flecknell, P., Leach, M. and Bateson, M. (2011) 'Affective state and quality of life in mice', *Pain*, 152(5), pp. 963-4.

Flecknell, P.A. (1986) 'Recognition and alleviation of pain in animals', in *Advances in Animal Welfare Science 1985*. Springer, pp. 61-77.

Flecknell, P.A. and Liles, J.H. (1991) 'The effects of surgical procedures, halothane anaesthesia and nalbuphine on locomotor activity and food and water consumption in rats', *Laboratory Animals*, 25(1), pp. 50-60.

Flecknell, P.A. and Roughan, J.V. (2004) 'Assessing pain in animals-putting research into practice', *Animal Welfare*, 13, pp. S71-S76.

Fleming, P.A., Paisley, C.L., Barnes, A.L. and Wemelsfelder, F. (2013) 'Application of Qualitative Behavioural Assessment to horses during an endurance ride', *Applied Animal Behaviour Science*, 144(1-2), pp. 80-88.

Fleming, P.A., Wickham, S.L., Stockman, C.A., Verbeek, E., Matthews, L. and Wemelsfelder, F. (2015) 'The sensitivity of QBA assessments of sheep behavioural expression to variations in visual or verbal information provided to observers', *animal*, 9(05), pp. 878-887.

Fox, J.G., Barthold, S., Davisson, M., Newcomer, C.E., Quimby, F.W. and Smith, A. (2006) *The mouse in biomedical research: normative biology, husbandry, and models*. Elsevier.

Francione, G.L. and Garner, R. (2010) *The animal rights debate: abolition or regulation?* Columbia University Press.

Fraser, A.F. and Broom, D.M. (1997) *Farm animal behaviour and welfare*. CAB international.

Fraser, D. (1989) 'Role of ethology in determining farm animal well-being', in *Science and Animals: Addressing Contemporary Issues*. Scientists Center for Animal Welfare Bethesda, MD, pp. 95-102.

Fraser, D. (1999) 'Animal ethics and animal welfare science: bridging the two cultures', *Applied Animal Behaviour Science*, 65(3), pp. 171-189.

Fraser, D. (2003) 'Assessing animal welfare at the farm and group level: the interplay of science and values'.

Fraser, D. (2009) 'Animal behaviour, animal welfare and the scientific study of affect', *Applied Animal Behaviour Science*, 118(3-4), pp. 108-117.

Fraser, D. and Nicol, C.J. (2011) 'Preference and motivation research', *Animal welfare*, pp. 183-199.

Fraser, D., Weary, D.M., Pajor, E.A. and Milligan, B.N. (1997) 'A Scientific Conception of Animal Welfare that Reflects Ethical Concerns', *Animal Welfare*, 6(3), pp. 187-205.

Frewer, L.J., Fischer, A.R.H., Wentholt, M.T.A., Marvin, H.J.P., Ooms, B.W., Coles, D. and Rowe, G. (2011) 'The use of Delphi methodology in agrifood policy development: Some lessons learned', *Technological Forecasting and Social Change*, 78(9), pp. 1514-1525.

Garner, J., Wayne, C., Würbel, H. and Mench, J. (2003) *Proceedings of the 37th International Congress of The International Society for Applied Ethology, Abano Terme, Italy: Fondazione Iniziative Zooprofilattiche e Zootecniche. p*.

Garner, J.P., Dufour, B., Gregg, L.E., Weisker, S.M. and Mench, J.A. (2004a) 'Social and husbandry factors affecting the prevalence and severity of barbering (‘whisker trimming’) by laboratory mice', *Applied Animal Behaviour Science*, 89(3–4), pp. 263-282.

Garner, J.P., Weisker, S.M., Dufour, B. and Mench, J.A. (2004b) 'Barbering (fur and whisker trimming) by laboratory mice as a model of human trichotillomania and obsessive-compulsive spectrum disorders', *Comparative medicine*, 54(2), pp. 216-224.

Gaskill, B.N., Gordon, C.J., Pajor, E.A., Lucas, J.R., Davis, J.K. and Garner, J.P. (2012) 'Heat or insulation: behavioral titration of mouse preference for warmth or access to a nest', *PloS one*, 7(3), p. e32799.

Gaskill, B.N., Gordon, C.J., Pajor, E.A., Lucas, J.R., Davis, J.K. and Garner, J.P. (2013a) 'Impact of nesting material on mouse body temperature and physiology', *Physiology & Behavior*, 110–111, pp. 87-95.

Gaskill, B.N., Karas, A.Z., Garner, J.P. and Pritchett-Corning, K.R. (2013b) 'Nest Building as an Indicator of Health and Welfare in Laboratory Mice', *Journal of Visualized Experiments : JoVE*, (82), p. 51012.

Gaskill, B.N., Rohr, S.A., Pajor, E.A., Lucas, J.R. and Garner, J.P. (2009) 'Some like it hot: Mouse temperature preferences in laboratory housing', *Applied Animal Behaviour Science*, 116(2–4), pp. 279-285.

Geist, M.R. (2010) 'Using the Delphi method to engage stakeholders: A comparison of two studies', *Evaluation and Program Planning*, 33(2), pp. 147-154.

Gillingham, M.B., Clark, M.D., Dahly, E.M., Krugner-Higby, L.A. and Ney, D.M. (2001) 'A Comparison of Two Opioid Analgesics for Relief of Visceral Pain Induced by Intestinal Resection in Rats', *Journal of the American Association for Laboratory Animal Science*, 40(1), pp. 21-26.

Gonyou, H.W., Hemsworth, P.H. and Barnett, J.L. (1986) 'Effects of frequent interactions with humans on growing pigs', *Applied Animal Behaviour Science*, 16(3), pp. 269-278.

Gouveia, K. and Hurst, J.L. (2013) 'Reducing Mouse Anxiety during Handling: Effect of Experience with Handling Tunnels', *PLOS ONE*, 8(6), p. e66401.

Gouveia, K. and Hurst, J.L. (2017) 'Optimising reliability of mouse performance in behavioural testing: the major role of non-aversive handling', *Scientific reports*, 7, p. 44999.

Griffin, G., Clark, J.M., Zurlo, J. and Ritskes-Hoitinga, M. (2014) 'Scientific uses of animals: harm-benefit analysis and complementary approaches to implementing the three Rs', *Revue scientifique et technique (International Office of Epizootics)*, 33(1), pp. 265-272.

Grosso, L., Battini, M., Wemelsfelder, F., Barbieri, S., Minero, M., Dalla Costa, E. and Mattiello, S. (2016) 'On-farm Qualitative Behaviour Assessment of dairy goats in different housing conditions', *Applied Animal Behaviour Science*, 180, pp. 51-57.

Guillen, J. (2012) 'FELASA Guidelines and Recommendations', *Journal of the American Association for Laboratory Animal Science*, 51(3), pp. 311-321.

Guyenet, S.J., Furrer, S.A., Damian, V.M., Baughan, T.D., La Spada, A.R. and Garden, G.A. (2010) 'A simple composite phenotype scoring system for evaluating mouse models of cerebellar ataxia', *Journal of visualized experiments: JoVE*, (39).

Haemisch, A. and Gärtner, K. (1994) 'The cage design affects intermale aggression in small groups of male laboratory mice: strain specific consequences on social organization, and endocrine activations in two inbred strains (DBA/2J and CBA/J)', *Journal of Experimental Animal Science*, 36(4-5), pp. 101-116.

Hagelin, J., Carlsson, H.-E. and Hau, J. (2003) 'An overview of surveys on how people view animal experimentation: some factors that may influence the outcome', *Public Understanding of Science*, 12(1), pp. 67-81.

Harvey, P.W. and Chevins, P.F.D. (1985) 'Crowding pregnant mice affects attack and threat behavior of male offspring', *Hormones and behavior*, 19(1), pp. 86-97.

Hasson, F., Keeney, S. and McKenna, H. (2000) 'Research guidelines for the Delphi survey technique', *Journal of advanced nursing*, 32(4), pp. 1008-1015.

Hau, J. and Schapiro, S.J. (2002) *Handbook of laboratory animal science: essential principles and practices*. CRC press.

Hawkins, P. (2002) 'Recognizing and assessing pain, suffering and distress in laboratory animals: a survey of current practice in the UK with recommendations', *Laboratory Animals*, 36(4), pp. 378-395.

Hawkins, P., Morton, D.B., Burman, O., Dennison, N., Honess, P., Jennings, M., Lane, S., Middleton, V., Roughan, J.V. and Wells, S. (2011) 'A guide to defining and implementing protocols for the welfare assessment of laboratory animals: eleventh report of the BVAAWF/FRAME/RSPCA/UFAW Joint Working Group on Refinement', *Laboratory animals*, 45(1), pp. 1-13.

Heizmann, V., Jonas, I., Hirschenauer, K. and Havelec, L. (1998) 'Choice tests with groups of mice: nestbox, nesting material and tubes as enrichment items for laboratory mice', *Journal of experimental animal science*, 39(1-2), pp. 43-60.

Hem, A., Smith, A.J. and Solberg, P. (1998) 'Saphenous vein puncture for blood sampling of the mouse, rat, hamster, gerbil, guineapig, ferret and mink', *Laboratory animals*, 32(4), pp. 364-368.

Hemsworthlt, P., Barnett, J. and Coleman, G. (1993) 'The human-animal relationship in agriculture and its consequences for the animal', *Animal Welfare*, 2(1), pp. 33-51.

Hess, G.R. and King, T.J. (2002) 'Planning open spaces for wildlife: I. Selecting focal species using a Delphi survey approach', *Landscape and Urban Planning*, 58(1), pp. 25-40.

Hess, S.E., Rohr, S., Dufour, B.D., Gaskill, B.N., Pajor, E.A. and Garner, J.P. (2008) 'Home improvement: C57BL/6J mice given more naturalistic nesting materials build better nests', *Journal of the American Association for Laboratory Animal Science*, 47(6), pp. 25-31.

Heston, W.E. (1967) 'Biology of the laboratory mouse', *Journal of Heredity*, 58(1), pp. 16-16.

Hintze, S., Murphy, E., Bachmann, I., Wemelsfelder, F. and Würbel, H. (2017) 'Qualitative Behaviour Assessment of horses exposed to short-term emotional treatments', *Applied animal behaviour science*, 196, pp. 44-51.

Hockly, E., Cordery, P.M., Woodman, B., Mahal, A., Van Dellen, A., Blakemore, C., Lewis, C.M., Hannan, A.J. and Bates, G.P. (2002) 'Environmental enrichment slows disease progression in R6/2 Huntington's disease mice', *Annals of neurology*, 51(2), pp. 235-242.

Hollands, C. (1986a) 'The Animals (scientific procedures) Act 1986', *Lancet (London, England)*, 2(8497), p. 32.

Hollands, C. (1986b) 'The animals (scientific procedures) act 1986', *The Lancet*, 328(8497), pp. 32-33.

Holmes, A., Yang, R.J., Murphy, D.L. and Crawley, J.N. (2002) 'Evaluation of antidepressant-related behavioral responses in mice lacking the serotonin transporter', *Neuropsychopharmacology*, 27(6), pp. 914-923.

Home Office (2014a) 'Code of Practice for the Housing and Care of Animals Bred, Supplied or Used for Scientific Purposes: Rodents and Rabbits', p. 65.

Home Office (2018) *Annual Statistics of Scientific Procedures on Living Animals, Great Britain 2017*.

Home Office, U.K. (2013) *Annual Statistics of Scientific.*
Procedures on Living Animals Great Britain 2013.

Home Office, U.K. (2014b) *Working to reduce the use of animals in scientific research*. United Kingdom.

Hsu, C.-C. and Sandford, B.A. (2007) 'The Delphi technique: making sense of consensus', *Practical assessment, research & evaluation*, 12(10), pp. 1-8.

Hubrecht, R. (2014) *The Welfare of Animals Used in Research: Practice and Ethics*. UFAW Animal Welfare Series.

Hubrecht, R., Hurst, J., Morton, D., Peters, A., Raymond, R., Sales, G.D., Sherwin, C.M. and West, C. (1993) 'Refining rodent husbandry: the mouse', *Lab Anim*, 27, pp. 301-329.

Hubrecht, R.C. and Kirkwood, J. (2010) *The UFAW handbook on the care and management of laboratory and other research animals*. John Wiley & Sons.

Hurst, J.L., Barnard, C.J., Tolladay, U., Nevision, C.M. and West, C.D. (1999) 'Housing and welfare in laboratory rats: effects of cage stocking density and behavioural predictors of welfare', *Animal Behaviour*, 58(3), pp. 563-586.

Hurst, J.L. and West, R.S. (2010) 'Taming anxiety in laboratory mice', *Nature Methods*, 7(10), pp. 825-826.

IASP (2008) 'The Kyoto protocol of IASP Basic Pain Terminology☆', *Pain*, 137(3), pp. 473-477.

Irwin, S. (1968) 'Comprehensive observational assessment: Ia. A systematic, quantitative procedure for assessing the behavioral and physiologic state of the mouse', *Psychopharmacologia*, 13(3), pp. 222-257.

Ishii, T., Yoshida, K., Hasegawa, M., Mizuno, S., Okamoto, M., Tajima, M. and Kurosawa, T. (1998) 'Invention of a forced-air-ventilated micro-isolation cage and rack system Environment within cages: temperature and ammonia concentration', *Applied Animal Behaviour Science*, 59(1), pp. 115-123.

Jacobsen, K.R., Kalliokoski, O., Teilmann, A.C., Hau, J. and Abelson, K.S. (2012) 'Postsurgical food and water consumption, fecal corticosterone metabolites, and behavior assessment as noninvasive measures of pain in vasectomized BALB/c mice', *Journal of the American Association for Laboratory Animal Science*, 51(1), pp. 69-75.

Jegstrup, I., Thon, R., Hansen, A.K. and Hoitinga, M.R. (2003) 'Characterization of transgenic mice—a comparison of protocols for welfare evaluation and phenotype characterization of mice with a suggestion on a future certificate of instruction', *Laboratory Animals*, 37(1), pp. 1-9.

Jensen, K.K. and Sandøe, P. (1997) 'Animal welfare: relative or absolute?', *Applied Animal Behaviour Science*, 54(1), pp. 33-37.

Joffily, M. and Coricelli, G. (2013) 'Emotional valence and the free-energy principle', *PLoS computational biology*, 9(6), p. e1003094.

Jolliffe, I. (2011) 'Principal component analysis', in *International encyclopedia of statistical science*. Springer, pp. 1094-1096.

Jones, B. and Manteca, X. (2009) *Practical strategies for improving farm animal welfare: an information resource: The Welfare Quality Project*.Welfare Quality®.

Kaliste, E. (2004) *The welfare of laboratory animals*. Springer.

Keeney, S., Hasson, F. and McKenna, H. (2006) 'Consulting the oracle: ten lessons from using the Delphi technique in nursing research', *Journal of advanced nursing*, 53(2), pp. 205-212.

Keeney, S., McKenna, H. and Hasson, F. (2010) *The Delphi technique in nursing and health research*. John Wiley & Sons.

Kirkden, R.D. and Pajor, E.A. (2006) 'Using preference, motivation and aversion tests to ask scientific questions about animals' feelings', *Applied Animal Behaviour Science*, 100(1-2), pp. 29-47.

Koo, T.K. and Li, M.Y. (2016) 'A guideline of selecting and reporting intraclass correlation coefficients for reliability research', *Journal of chiropractic medicine*, 15(2), pp. 155-163.

Landers, R. (2015) 'Computing Intraclass Correlations (ICC) as Estimates of Interrater Reliability in SPSS, The Winnower 2: e143518. 81744, 2015, DOI: 10.15200/winn. 143518.81744 Landers '.

Landis, J.R. and Koch, G.G. (1977) 'The measurement of observer agreement for categorical data', *biometrics*, pp. 159-174.

Lane-Petter, W. (1966) 'UFAW handbook on the care and management of laboratory animals'.

Lariviere, W.R. and Melzack, R. (1996) 'The bee venom test: a new tonic-pain test', *PAIN*®, 66(2), pp. 271-277.

Latham, N. and Mason, G. (2004) 'From house mouse to mouse house: the behavioural biology of free-living Mus musculus and its implications in the laboratory', *Applied Animal Behaviour Science*, 86(3–4), pp. 261-289.

Lawlor, M. (1983) 'Behavioural approaches to rodent management', *Standards in laboratory animal management UFAW (ed.), UFAW, Potters Bar*, pp. 40-49.

Leach, M.C., Ambrose, N., Bowell, V.J. and Morton, D.B. (2000) 'The development of a novel form of mouse cage enrichment', *Journal of applied animal welfare science*, 3(2), pp. 81-91.

Leach, M.C., Bowell, V.A., Allan, T.F. and Morton, D.B. (2004) 'Measurement of aversion to determine humane methods of anaesthesia and euthanasia', *Animal Welfare*, 13(1), pp. 77-86.

Leach, M.C., Klaus, K., Miller, A.L., Di Perrotolo, M.S., Sotocinal, S.G. and Flecknell, P.A. (2012) 'The assessment of post-vasectomy pain in mice using behaviour and the Mouse Grimace Scale', *PloS one*, 7(4), p. e35656.

Leach, M.C. and Main, D.C.J. (2008) 'An assessment of laboratory mouse welfare in UK animal units', *Animal Welfare*, 17(2), pp. 171-187.

Leach, M.C., Thornton, P.D. and Main, D.C.J. (2006) 'Development of a welfare-benchmarking scheme for laboratory mice', *In Proceedings of the ninth FELASA Symposium on Internationalisation and Harmonisation of Laboratory Animal Care and Use Issues*, p. 68.

Leach, M.C., Thornton, P.D. and Main, D.C.J. (2008a) 'Identification of appropriate measures for the assessment of laboratory mouse welfare', *Animal Welfare*, 17(2), pp. 161-170.

Leach, M.C., Thornton, P.D. and Main, D.C.J. (2008b) *Identification of appropriate measures for the assessment of laboratory mouse welfare/ M. C. Leach, P. D. Thornton and D. C. J. Main*. South Mimms, Potters Bar, Herts. UK: Universities Federation for Animal Welfare.

Liles, J.H. and Flecknell, P.A. (1992) 'The use of non-steroidal anti-inflammatory drugs for the relief of pain in laboratory rodents and rabbits', *Laboratory Animals*, 26(4), pp. 241-255.

Linstone, H.A. and Turoff, M. (2002) 'The Delphi Method', *Techniques and applications*, 53.

Litwin, M.S. and Fink, A. (1995) *How to measure survey reliability and validity*. Sage.

Lloyd, M.H. and Wolfensohn, S.E. (1999) 'Practical use of distress scoring systems in the application of humane endpoints'. London: Royal Society of Medicine Press, pp. 48-53.

Long, S.Y. (1972) 'Hair-nibbling and whisker-trimming as indicators of social hierarchy in mice', *Animal behaviour*, 20(1), pp. 10-12.

Lund, T.B., Mørkbak, M.R., Lassen, J. and Sandøe, P. (2014) 'Painful dilemmas: A study of the way the public’s assessment of animal research balances costs to animals against human benefits', *Public Understanding of Science*, 23(4), pp. 428-444.

Mandel, J.E. (1974) 'A strategy for selecting and phrasing questions in an interview', *The Journal of Business Communication (1973)*, 12(1), pp. 17-23.

Marashi, V., Barnekow, A., Ossendorf, E. and Sachser, N. (2003) 'Effects of different forms of environmental enrichment on behavioral, endocrinological, and immunological parameters in male mice', *Hormones and Behavior*, 43(2), pp. 281-292.

Mason, G., Clubb, R., Latham, N. and Vickery, S. (2007) 'Why and how should we use environmental enrichment to tackle stereotypic behaviour?', *Applied Animal Behaviour Science*, 102(3–4), pp. 163-188.

Mason, G.J. (1991) 'Stereotypies: a critical review', *Animal behaviour*, 41(6), pp. 1015-1037.

Matfield, M.J. (1991) 'Animal liberation or animal research?', *Trends in pharmacological sciences*, 12, pp. 411-415.

McClelland, G.H. and Judd, C.M. (1993) 'Statistical difficulties of detecting interactions and moderator effects', *Psychological bulletin*, 114(2), p. 376.

McKenna, H. and Hasson, F. (2002) 'A study of skill mix issues in midwifery: a multimethod approach', *Journal of Advanced Nursing*, 37(1), pp. 52-61.

McKenna, H.P. (1994) 'The essential elements of a practitioners' nursing model: a survey of psychiatric nurse managers', *Journal of Advanced Nursing*, 19(5), pp. 870-877.

McKenna, H.P., Bradley, M. and Keeney, S. (2001) 'Primary care nursing: a study exploring key issues for future developments', *Coleraine: University of Ulster*.

Meaney, M.J., Diorio, J., Francis, D., Widdowson, J., LaPlante, P., Caldji, C., Sharma, S., Seckl, J.R. and Plotsky, P.M. (1996) 'Early Environmental Regulation of Forebrain Glucocorticoid Receptor Gene Expression: Implications for Adrenocortical Responses to Stress; pp. 61–72', *Developmental neuroscience*, 18(1-2), pp. 61-72.

Meek, L.R., Dittel, P.L., Sheehan, M.C., Chan, J.Y. and Kjolhaug, S.R. (2001) 'Effects of stress during pregnancy on maternal behavior in mice', *Physiology & behavior*, 72(4), pp. 473-479.

Mellor, D. and Reid, C. (1994) 'Concepts of animal well-being and predicting the impact of procedures on experimental animals', *Improving the well-being of animals in the research environment*, pp. 3-18.

Mellor, D.J. and Beausoleil, N. (2015) 'Extending the ‘Five Domains’ model for animal welfare assessment to incorporate positive welfare states', *Animal Welfare*, 24(3), pp. 241-253.

Mendl, M., Burman, O.H.P., Parker, R.M.A. and Paul, E.S. (2009) 'Cognitive bias as an indicator of animal emotion and welfare: emerging evidence and underlying mechanisms', *Applied Animal Behaviour Science*, 118(3), pp. 161-181.

Mendl, M., Burman, O.H.P. and Paul, E.S. (2010) 'An integrative and functional framework for the study of animal emotion and mood', *Proceedings of the Royal Society of London B: Biological Sciences*, 277(1696), pp. 2895-2904.

Mendl, M. and Paul, E.S. (2004) 'Consciousness, emotion and animal welfare: insights from cognitive science', *Animal Welfare*, 13, pp. S17-S26.

Mertens, C. and Rülicke, T. (1999) 'Score sheets for the monitoring of transgenic mice', *Animal Welfare*, 8(4), pp. 433-438.

Miller, A.L., Flecknell, P.A., Leach, M.C. and Roughan, J.V. (2011) 'A comparison of a manual and an automated behavioural analysis method for assessing post-operative pain in mice', *Applied Animal Behaviour Science*, 131(3-4), pp. 138-144.

Miller, A.L., Kitson, G.L., Skalkoyannis, B., Flecknell, P.A. and Leach, M.C. (2016) 'Using the mouse grimace scale and behaviour to assess pain in CBA mice following vasectomy', *Applied Animal Behaviour Science*, 181(Supplement C), pp. 160-165.

Minero, M., Dalla Costa, E., Dai, F., Murray, L.A.M., Canali, E. and Wemelsfelder, F. (2016) 'Use of Qualitative Behaviour Assessment as an indicator of welfare in donkeys', *Applied Animal Behaviour Science*, 174, pp. 147-153.

Minero, M., Tosi, M.V., Canali, E. and Wemelsfelder, F. (2009) 'Quantitative and qualitative assessment of the response of foals to the presence of an unfamiliar human', *Applied Animal Behaviour Science*, 116(1), pp. 74-81.

Moore, A. (2001) 'Of mice and Mendel', *EMBO reports*, 2(7), pp. 554-558.

More, S.J., McKenzie, K., O'Flaherty, J., Doherty, M.L., Cromie, A.R. and Magan, M.J. (2010) 'Setting priorities for non-regulatory animal health in Ireland: results from an expert Policy Delphi study and a farmer priority identification survey', *Preventive veterinary medicine*, 95(3), pp. 198-207.

Morton, D.B. (1998) *The importance of non-statistical design in refining animal experiments*. ANZCCART.

Morton, D.B. (1999) 'Humane endpoints in animal experimentation for biomedical research: ethical, legal and practical aspects'. London: Royal Society of Medicine Press, pp. 5-12.

Morton, D.B., Berghardt, G.M. and Smith, J.A. (1989) 'Animals, science, and ethics--Section III. Critical anthropomorphism, animal suffering, and the ecological context', *The Hastings Center Report*, 20(3), pp. S13-9.

Morton, D.B. and Griffiths, P.H. (1985) 'Guidelines on the recognition of pain, distress and discomfort in experimental animals and an hypothesis for assessment', *The Veterinary Record*, 116(16), pp. 431-436.

Mukerjee, M. (1997) 'Trends in animal research', *Scientific American*, 276(2), pp. 86-93.

Muri, K. and Stubsjøen, S.M. (2017) 'Inter-observer reliability of Qualitative Behavioural Assessments (QBA) of housed sheep in Norway using fixed lists of descriptors', *Animal Welfare*, 26(4), pp. 427-435.

Murphy, E., Nordquist, R.E. and van der Staay, F.J. (2014) 'A review of behavioural methods to study emotion and mood in pigs, Sus scrofa', *Applied Animal Behaviour Science*, 159(0), pp. 9-28.

Nagy, T.R., Krzywanski, D., Li, J., Meleth, S. and Desmond, R. (2002) 'Effect of group vs. single housing on phenotypic variance in C57BL/6J mice', *Obesity research*, 10(5), pp. 412-415.

Napolitano, F., De Rosa, G., Grasso, F. and Wemelsfelder, F. (2012) 'Qualitative behaviour assessment of dairy buffaloes (Bubalus bubalis)', *Applied Animal Behaviour Science*, 141(3-4), pp. 91-100.

Nelson, R.J., Badura, L.L. and Goldman, B.D. (1990) 'Mechanisms of seasonal cycles of behavior', *Annual review of psychology*, 41(1), pp. 81-108.

Neubert, J.K., Widmer, C.G., Malphurs, W., Rossi, H.L., Vierck, C.J. and Caudle, R.M. (2005) 'Use of a novel thermal operant behavioral assay for characterization of orofacial pain sensitivity', *Pain*, 116(3), pp. 386-395.

Nevison, C.M., Hurst, J.L. and Barnard, C.J. (1999) 'Strain-Specific Effects of Cage Enrichment in Male Laboratory Mice (Mus Musculus)', *Animal Welfare*, 8(4), pp. 361-379.

Nevo, B. (1985) 'Face Validity Revisited', *Journal of Educational Measurement*, 22(4), pp. 287-293.

Nollet, M., Guisquet, A.M.L. and Belzung, C. (2013) 'Models of depression: unpredictable chronic mild stress in mice', *Current protocols in pharmacology*, 61(1), pp. 5.65. 1-5.65. 17.

Ohl, F., Arndt, S.S. and van der Staay, F.J. (2008) 'Pathological anxiety in animals', *The Veterinary Journal*, 175(1), pp. 18-26.

Olfert, E.D. (1995) 'Defining an acceptable endpoint in invasive experiments'.

Olsson, I.A.S. and Dahlborn, K. (2002) 'Improving housing conditions for laboratory mice: a review of'environmental enrichment'', *Laboratory Animals*, 36(3), pp. 243-270.

Ormandy, E.H., Schuppli, C.A. and Weary, D.M. (2009) 'Worldwide trends in the use of animals in research: the contribution of genetically-modified animal models', *Alternatives to laboratory animals: ATLA*, 37(1), pp. 63-68.

Palmer, C. (2010) *Animal ethics in context*. Columbia University Press.

Panksepp, J. (1994) 'Evolution constructed the potential for subjective experience within the neurodynamics of the mammalian brain'.

Panksepp, J. (2004) *Affective neuroscience: The foundations of human and animal emotions*. Oxford university press.

Paster, E.V., Villines, K.A. and Hickman, D.L. (2009) 'Endpoints for mouse abdominal tumor models: refinement of current criteria', *Comparative medicine*, 59(3), pp. 234-241.

Paul, E.S., Harding, E.J. and Mendl, M. (2005) 'Measuring emotional processes in animals: the utility of a cognitive approach', *Neurosci Biobehav Rev*, 29(3), pp. 469-91.

Pepperberg, I.M. (2006) 'Cognitive and communicative abilities of Grey parrots', *Applied Animal Behaviour Science*, 100(1), pp. 77-86.

Perry, P. (2007) 'The ethics of animal research: a UK perspective', *ILAR journal*, 48(1), pp. 42-46.

Phythian, C., Michalopoulou, E., Duncan, J. and Wemelsfelder, F. (2013) 'Inter-observer reliability of Qualitative Behavioural Assessments of sheep', *Applied Animal Behaviour Science*, 144(1–2), pp. 73-79.

Phythian, C.J., Michalopoulou, E., Cripps, P.J., Duncan, J.S. and Wemelsfelder, F. (2016) 'On-farm qualitative behaviour assessment in sheep: Repeated measurements across time, and association with physical indicators of flock health and welfare', *Applied Animal Behaviour Science*, 175, pp. 23-31.

Portfors, C.V. (2007) 'Types and functions of ultrasonic vocalizations in laboratory rats and mice', *Journal of the American Association for Laboratory Animal Science*, 46(1), pp. 28-34.

Portney, L. and Watkins, M. (1993) 'Foundations of clinical research: application to practice', *Stamford, USA: Appleton & Lange*.

Posner, J., Russell, J.A. and Peterson, B.S. (2005) 'The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology', *Development and psychopathology*, 17(3), pp. 715-734.

Potter, M.J. and Broom, D.M. (1987) 'The behaviour and welfare of cows in relation to cubicle house design', in *Cattle housing systems, lameness and behaviour*. Martinus Nijhoff Dordrecht, pp. 129-147.

Proctor, H.S. and Carder, G. (2014) 'Can ear postures reliably measure the positive emotional state of cows?', *Applied Animal Behaviour Science*, 161(0), pp. 20-27.

Quality, W. (2009) *Assessment protocol for pigs*. Welfare Quality®.

Quinn, T.J., Ray, G., Atula, S., Walters, M.R., Dawson, J. and Lees, K.R. (2008) 'Deriving modified Rankin scores from medical case-records', *Stroke*, 39(12), pp. 3421-3423.

Reeb-Whitaker, C.K., Paigen, B., Beamer, W.G., Bronson, R.T., Churchill, G.A., Schweitzer, I.B. and Myers, D.D. (2001) 'The impact of reduced frequency of cage changes on the health of mice housed in ventilated cages', *Laboratory animals*, 35(1), pp. 58-73.

Reeb, C.K., Jones, R.B., Bearg, D.W., Bedigian, H., Myers, D.D. and Paigen, B. (1998) 'Microenvironment in ventilated animal cages with differing ventilation rates, mice populations, and frequency of bedding changes', *Journal of the American Association for Laboratory Animal Science*, 37(2), pp. 43-49.

Regan, T. (1987) *The case for animal rights*. Springer.

Rencher, A.C. (2003) *Methods of multivariate analysis*. John Wiley & Sons.

Richardson, C.A. and Flecknell, P.A. (2005) 'Anaesthesia and post-operative analgesia following experimental surgery in laboratory rodents: are we making progress', *Anesthesia and Analgesia*, 1(3).

Rikkonen, P. (2008) 'Scenarios for future agriculture in Finland: a Delphi study among agri-food sector stakeholders', *Agricultural and Food Science*, 14(3), pp. 205-223.

Rock, M.L., Karas, A.Z., Rodriguez, K.B.G., Gallo, M.S., Pritchett-Corning, K., Karas, R.H., Aronovitz, M. and Gaskill, B.N. (2014) 'The time-to-integrate-to-nest test as an indicator of wellbeing in laboratory mice', *Journal of the American Association for Laboratory Animal Science*, 53(1), pp. 24-28.

Rogers, D.C., Fisher, E.M.C., Brown, S.D.M., Peters, J., Hunter, A.J. and Martin, J.E. (1997) 'Behavioral and functional analysis of mouse phenotype: SHIRPA, a proposed protocol for comprehensive phenotype assessment', *Mammalian Genome*, 8(10), pp. 711-713.

Rogers, D.C., Jones, D.N.C., Nelson, P.R., Jones, C.M., Quilter, C.A., Robinson, T.L. and Hagan, J.J. (1999) 'Use of SHIRPA and discriminant analysis to characterise marked differences in the behavioural phenotype of six inbred mouse strains', *Behavioural Brain Research*, 105(2), pp. 207-217.

Roper, T. (1973) 'Nesting material as a reinforcer for female mice', *Animal Behaviour*, 21(4), pp. 733-740.

Rousing, T., Bonde, M. and Sørensen, J.T. (2001) 'Aggregating welfare indicators into an operational welfare assessment system: a bottom-up approach', *Acta Agriculturae Scandinavica, Section A-Animal Science*, 51(S30), pp. 53-57.

Rousing, T. and Wemelsfelder, F. (2006) 'Qualitative assessment of social behaviour of dairy cows housed in loose housing systems', *Applied Animal Behaviour Science*, 101(1–2), pp. 40-53.

Russell, J.A. (1980) 'A circumplex model of affect', *Journal of personality and social psychology*, 39(6), p. 1161.

Russell, J.A. (2003) 'Core affect and the psychological construction of emotion', *Psychological review*, 110(1), p. 145.

Russell, W.M.S., Burch, R.L. and Hume, C.W. (1959) 'The principles of humane experimental technique'.

Rutherford, K.M., Donald, R.D., Lawrence, A.B. and Wemelsfelder, F. (2012) 'Qualitative Behavioural Assessment of emotionality in pigs', *Appl Anim Behav Sci*, 139(3-4), pp. 218-224.

Scharmann, W. (1991) 'Improved housing of mice, rats and guinea-pigs: a contribution to the refinement of animal experiments', *Alternatives to laboratory animals: ATLA*.

Scherer, K.R. (2005) 'What are emotions? And how can they be measured?', *Social science information*, 44(4), pp. 695-729.

Schmidt, R.C. (1997) 'Managing Delphi Surveys Using Nonparametric Statistical Techniques*', *Decision Sciences*, 28(3), pp. 763-774.

Scott, K., Laws, D.M., Courboulay, V., Meunier-Salaün, M.-C. and Edwards, S.A. (2009) 'Comparison of methods to assess fear of humans in sows', *Applied Animal Behaviour Science*, 118(1–2), pp. 36-41.

Serpell, J.A. (2002) 'Anthropomorphism and Anthropomorphic Selection—Beyond the'' Cute Response''', *Society & Animals*, 10(4), pp. 437-454.

Serrapica, M., Boivin, X., Coulon, M., Braghieri, A. and Napolitano, F. (2017) 'Positive perception of human stroking by lambs: Qualitative behaviour assessment confirms previous interpretation of quantitative data', *Applied Animal Behaviour Science*, 187, pp. 31-37.

Sharp, J., Zammit, T., Azar, T. and Lawson, D. (2003) 'Stress-like Responses to Common Procedures in Individually and Group-Housed Female Rats', *Journal of the American Association for Laboratory Animal Science*, 42(1), pp. 9-18.

Sherwin, C.M. (1997) 'Observations on the prevalence of nest-building in non-breeding TO strain mice and their use of two nesting materials', *Laboratory Animals*, 31(2), pp. 125-132.

Sherwin, C.M. (2001) 'Can invertebrates suffer? Or, how robust is argument-by-analogy?', *Animal Welfare*, 10(Supplement 1), pp. 103-118.

Singer, P. (1995) *Animal liberation*. Random House.

Singer, P. (2011) *Practical ethics*. Cambridge university press.

Sireci, S.G. (1998) 'The construct of content validity', *Social indicators research*, 45(1-3), pp. 83-117.

Sørensen, J.T. and Fraser, D. (2010) 'On-farm welfare assessment for regulatory purposes: Issues and possible solutions', *Livestock Science*, 131(1), pp. 1-7.

Spangenberg, E.M.F. and Keeling, L.J. (2015) 'Assessing the welfare of laboratory mice in their home environment using animal-based measures – a benchmarking tool', *Laboratory Animals*.

Staats, J. (1966) 'The laboratory mouse', *Biology of the laboratory mouse*, 1.

Stafleu, F.R., Rivas, E., Rivas, T., Vorstenbosch, J., Heeger, F.R. and Beynen, A.C. (1992) 'The Use of Analogous Reasoning for Assessing Discomfort in Laboratory Animals', *Animal Welfare*, 1(2), pp. 77-84.

Stevenson-Hinde, J., Stillwell-Barnes, R. and Zunz, M. (1980) 'Subjective assessment of rhesus monkeys over four successive years', *Primates*, 21(1), pp. 66-82.

Stevenson-Hinde, J. and Zunz, M. (1978) 'Subjective assessment of individual rhesus monkeys', *Primates*, 19(3), pp. 473-482.

Stockman, C., Collins, T., Barnes, A., Miller, D., Wickham, S., Beatty, D., Blache, D., Wemelsfelder, F. and Fleming, P. (2011) 'Qualitative behavioural assessment and quantitative physiological measurement of cattle naïve and habituated to road transport', *Animal production science*, 51(3), pp. 240-249.

Stokes, E.L., Flecknell, P.A. and Richardson, C.A. (2009) 'Reported analgesic and anaesthetic administration to rodents undergoing experimental surgical procedures', *Laboratory Animals*, 43(2), pp. 149-154.

Stokes, W.S. (2002) 'Humane endpoints for laboratory animals used in regulatory testing', *ILAR Journal*, 43(Suppl 1), pp. S31-S38.

Strozik, E. and Festing, M. (1981) 'Whisker trimming in mice', *Laboratory Animals*, 15(4), pp. 309-312.

Tannenbaum, J. and Bennett, B.T. (2015) 'Russell and Burch's 3Rs then and now: the need for clarity in definition and purpose', *Journal of the American Association for Laboratory Animal Science*, 54(2), pp. 120-132.

Taylor, K., Gordon, N., Langley, G. and Higgins, W. (2008) 'Estimates for worldwide laboratory animal use in 2005'.

Temple, D., Manteca, X., Velarde, A. and Dalmau, A. (2011) 'Assessment of animal welfare through behavioural parameters in Iberian pigs in intensive and extensive conditions', *Applied Animal Behaviour Science*, 131(1-2), pp. 29-39.

Thagard, P. and Aubie, B. (2008) 'Emotional consciousness: A neural model of how cognitive appraisal and somatic perception interact to produce qualitative experience', *Consciousness and Cognition*, 17(3), pp. 811-834.

Torrance, G.W., Feeny, D. and Furlong, W. (2001) 'Visual analog scales: do they have a role in the measurement of preferences for health states?'. Sage Publications Sage CA: Thousand Oaks, CA.

Treit, D., Terlecki, L.J. and Pinel, J.P. (1980) 'Conditioned defensive burying in rodents: Organismic variables', *Bulletin of the Psychonomic Society*, 16(6), pp. 451-454.

Tuli, J., Smith, J. and Morton, D. (1995) 'Corticosterone, adrenal and spleen weight in mice after tail bleeding, and its effect on nearby animals', *Laboratory Animals*, 29(1), pp. 90-95.

Turner, C.A., Yang, M.C. and Lewis, M.H. (2002) 'Environmental enrichment: effects on stereotyped behavior and regional neuronal metabolic activity', *Brain Research*, 938(1), pp. 15-21.

Ullman-Culleré, M.H. and Foltz, C.J. (1999) 'Body condition scoring: a rapid and accurate method for assessing health status in mice', *Comparative Medicine*, 49(3), pp. 319-323.

Valentinuzzi, V.S., Scarbrough, K., Takahashi, J.S. and Turek, F.W. (1997) 'Effects of aging on the circadian rhythm of wheel-running activity in C57BL/6 mice', *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology*, 273(6), pp. R1957-R1964.

Van de Weerd, H.A. and Baumans, V. (1995) 'Environmental enrichment in rodents', *Environmental Enrichment Information Resources for Laboratory Animals. AWIC Resource Series*, (2), pp. 145-9.

Van de Weerd, H.A., Van Loo, P.L.P., Van Zutphen, L.F.M., Koolhaas, J.M. and Baumans, V. (1997a) 'Nesting material as environmental enrichment has no adverse effects on behavior and physiology of laboratory mice', *Physiology & Behavior*, 62(5), pp. 1019-1028.

Van de Weerd, H.A., Van Loo, P.L.P., Van Zutphen, L.F.M., Koolhaas, J.M. and Baumans, V. (1997b) 'Preferences for nesting material as environmental enrichment for laboratory mice', *Laboratory Animals*, 31(2), pp. 133-143.

Van de Weerd, H.A., Van Loo, P.L.P., Van Zutphen, L.F.M., Koolhaas, J.M. and Baumans, V. (1998a) 'Strength of preference for nesting material as environmental enrichment for laboratory mice', *Applied Animal Behaviour Science*, 55(3–4), pp. 369-382.

Van de Weerd, H.A., Van Loo, P.L.P., Van Zutphen, L.F.M., Koolhaus, J.M. and Baumans, V. (1998b) 'Preferences for nest boxes as environmental enrichment for laboratory mice', *ANIMAL WELFARE-POTTERS BAR-*, 7, pp. 11-26.

Van der Meer, E., Van Loo, P.L.P. and Baumans, V. (2004) 'Short-term effects of a disturbed light–dark cycle and environmental enrichment on aggression and stress-related parameters in male mice', *Laboratory Animals*, 38(4), pp. 376-383.

Van der Meer, M., Baumans, V., Olivier, B., Kruitwagen, C.L.J.J., Van Dijk, J.E. and Van Zutphen, L.F.M. (2001a) 'Behavioral and physiological effects of biotechnology procedures used for gene targeting in mice', *Physiology & Behavior*, 73(5), pp. 719-730.

van der Meer, M., Rolls, A., Baumans, V., Olivier, B. and van Zutphen, L.F.M. (2001b) 'Use of score sheets for welfare assessment of transgenic mice', *Laboratory Animals*, 35(4), pp. 379-389.

Van Loo, P.L.P., Kruitwagen, C., Koolhaas, J.M., Van de Weerd, H.A., Van Zutphen, L.F.M. and Baumans, V. (2002) 'Influence of cage enrichment on aggressive behaviour and physiological parameters in male mice', *Applied Animal Behaviour Science*, 76(1), pp. 65-81.

Van Loo, P.L.P., Kruitwagen, C.L.J.J., Van Zutphen, L.F.M., Koolhaas, J.M. and Baumans, V. (2000) 'Modulation of Aggression in Male Mice: Influence of Cage Cleaning Regime and Scent Marks', *Animal Welfare*, 9(3), pp. 281-295.

Van Praag, H., Kempermann, G. and Gage, F.H. (2000) 'Neural consequences of enviromental enrichment', *Nature Reviews Neuroscience*, 1(3), pp. 191-198.

Van Swieten, J., Koudstaal, P., Visser, M., Schouten, H. and Van Gijn, J. (1988) 'Interobserver agreement for the assessment of handicap in stroke patients', *stroke*, 19(5), pp. 604-607.

Veissier, I., Jensen, K.K., Botrea, R. and Sandøe, P. (2011) 'Highlighting ethical decisions underlying the scoring of animal welfare in the Welfare Quality® scheme', *Animal Welfare*, 20(1), p. 89.

Velarde, A. and Dalmau, A. (2012) 'Animal welfare assessment at slaughter in Europe: Moving from inputs to outputs', *Meat Science*, 92(3), pp. 244-251.

Vieira de Castro, A. and Olsson, I.A. (2015) 'Does the Goal Justify the Methods? Harm and Benefit in Neuroscience Research Using Animals', in Lee, G., Illes, J. and Ohl, F. (eds.) *Ethical Issues in Behavioral Neuroscience*. Springer Berlin Heidelberg, pp. 47-78.

Vogel, B., Wagner, H., Gmoser, J., Wörner, A., Löschberger, A., Peters, L., Frey, A., Hofmann, U. and Frantz, S. (2016) 'Touch-free measurement of body temperature using close-up thermography of the ocular surface', *MethodsX*, 3, pp. 407-416.

Voikar, V., Polus, A., Vasar, E. and Rauvala, H. (2005) 'Long-term individual housing in C57BL/6J and DBA/2 mice: assessment of behavioral consequences', *Genes, Brain and Behavior*, 4(4), pp. 240-252.

Walker, J., Dale, A., Waran, N., Clarke, N., Farnworth, M. and Wemelsfelder, F. (2010) 'The assessment of emotional expression in dogs using a Free Choice Profiling methodology', *Animal Welfare*, 19(1), pp. 75-84.

Wallace, V.C.J., Norbury, T.A. and Rice, A.S.C. (2005) 'Ultrasound vocalisation by rodents does not correlate with behavioural measures of persistent pain', *European Journal of Pain*, 9(4), pp. 445-452.

Weary, D.M., Niel, L., Flower, F.C. and Fraser, D. (2006) 'Identifying and preventing pain in animals', *Applied Animal Behaviour Science*, 100(1–2), pp. 64-76.

Webster, J. (2016) 'Animal welfare: Freedoms, dominions and "a life worth living"', *Animals*, 6(6), p. 35.

Welfare Quality® (2009a) *Assessment protocol for pigs*. Welfare Quality®.

Welfare Quality® (2009b) *Assessment protocol for poultry*. Welfare Quality®.

Welfare Quality® (2009c) 'Welfare Quality® assessment protocol for cattle', *Lelystad, The Netherlands*, p. 182.

Wells, D.J. (2011) 'Animal welfare and the 3Rs in European biomedical research', *Annals of the New York Academy of Sciences*, 1245(1), pp. 14-16.

Wells, D.J., Playle, L.C., Enser, W.E.J., Flecknell, P.A., Gardiner, M.A., Holland, J., Howard, B.R., Hubrecht, R., Humphreys, K.R. and Jackson, I.J. (2006a) 'Assessing the welfare of genetically altered mice', *Laboratory animals*, 40(2), pp. 111-114.

Wells, D.J., Playle, L.C., Enser, W.E.J., Flecknell, P.A., Gardiner, M.A., Holland, J., Howard, B.R., Hubrecht, R., Humphreys, K.R., Jackson, I.J., Lane, N., Maconochie, M., Mason, G., Morton, D.B., Raymond, R., Robinson, V., Smith, J.A. and Watt, N. (2006b) 'Assessing the welfare of genetically altered mice', *Laboratory Animals*, 40(2), pp. 111-114.

Wemelsfelder, F. (1997) 'The scientific validity of subjective concepts in models of animal welfare', *Applied Animal Behaviour Science*, 53(1–2), pp. 75-88.

Wemelsfelder, F. (2007) 'How animals communicate quality of life: the qualitative assessment of behaviour', *Animal Welfare*, 16(2), pp. 25-31.

Wemelsfelder, F., Hunter, A.E., Paul, E.S. and Lawrence, A.B. (2012) 'Assessing pig body language: Agreement and consistency between pig farmers, veterinarians, and animal activists', *Journal of Animal Science*, 90(10), pp. 3652-3665.

Wemelsfelder, F., Hunter, E.A., Mendl, M.T. and Lawrence, A.B. (2000) 'The spontaneous qualitative assessment of behavioural expressions in pigs: first explorations of a novel methodology for integrative animal welfare measurement', *Applied Animal Behaviour Science*, 67(3), pp. 193-215.

Wemelsfelder, F., Hunter, T.E.A., Mendl, M.T. and Lawrence, A.B. (2001) 'Assessing the ‘whole animal’: a free choice profiling approach', *Animal Behaviour*, 62(2), pp. 209-220.

Wemelsfelder, F., Nevison, I. and Lawrence, A.B. (2009) 'The effect of perceived environmental background on qualitative assessments of pig behaviour', *Animal Behaviour*, 78(2), pp. 477-484.

Wentholt, M.T.A., Cardoen, S., Imberechts, H., Van Huffel, X., Ooms, B.W. and Frewer, L.J. (2012) 'Defining European preparedness and research needs regarding emerging infectious animal diseases: Results from a Delphi expert consultation', *Preventive Veterinary Medicine*, 103(2–3), pp. 81-92.

Whay, H., Main, D., Green, L. and Webster, A. (2003) 'Assessment of the welfare of dairy caftle using animal-based measurements: direct observations and investigation of farm records', *Veterinary record*, 153(7), pp. 197-202.

Whaytt, H.R., Main, D.C.J., Greent, L.E. and Webster, A.J.F. (2003) 'Animal-based measures for the assessment of welfare state of dairy cattle, pigs and laying hens: consensus of expert opinion', *Animal Welfare*, 12(2), pp. 205-217.

Wickham, S.L., Collins, T., Barnes, A.L., Miller, D.W., Beatty, D.T., Stockman, C., Blache, D., Wemelsfelder, F. and Fleming, P.A. (2012) 'Qualitative behavioral assessment of transport-naïve and transport-habituated sheep', *Journal of Animal Science*, 90(12), pp. 4523-4535.

Wiepkema, P.R. and Koolhaas, J.M. (1993) 'Stress and animal welfare', *Animal welfare*, 2(3), pp. 195-218.

Wolfensohn, S. and Lloyd, M. (2008) *Handbook of laboratory animal management and welfare*. John Wiley & Sons.

Wolfer, D.P., Litvin, O., Morf, S., Nitsch, R.M., Lipp, H.-P. and Würbel, H. (2004) 'Cage enrichment and mouse behaviour', *Nature*, 432, p. 821.

Workman, P., Aboagye, E., Balkwill, F., Balmain, A., Bruder, G., Chaplin, D., Double, J., Everitt, J., Farningham, D. and Glennie, M. (2010) 'Guidelines for the welfare and use of animals in cancer research', *British journal of cancer*, 102(11), p. 1555.

Wright-Williams, S.L., Courade, J.-P., Richardson, C.A., Roughan, J.V. and Flecknell, P.A. (2007) 'Effects of vasectomy surgery and meloxicam treatment on faecal corticosterone levels and behaviour in two strains of laboratory mouse', *Pain*, 130(1-2), pp. 108-118.

Würbel, H. (2001) 'Ideal homes? Housing effects on rodent brain and behaviour', *Trends in neurosciences*, 24(4), pp. 207-211.

Würbel, H., Stauffacher, M. and Holst, D. (1996) 'Stereotypies in Laboratory Mice—Quantitative and Qualitative Description of the Ontogeny of 'Wire-gnawing'and 'Jumping'in Zur: ICR and Zur: ICR nu', *Ethology*, 102(3), pp. 371-385.

Xiong, R., Blot, K., Meullenet, J.F. and Dessirier, J.M. (2008) 'Permutation tests for Generalized Procrustes Analysis', *Food Quality and Preference*, 19(2), pp. 146-155.

Yeates, J.W. and Main, D.C.J. (2008) 'Assessment of positive welfare: A review', *The Veterinary Journal*, 175(3), pp. 293-300.