**Assessing animal welfare: a triangulation of preference, judgement bias and other candidate welfare indicators**

**Elizabeth S. Paul1,\*, William Browne2, Michael T. Mendl1, Gina Caplen1, Anna Trevarthen1, Suzanne Held1, Christine J. Nicol3,\***

1Bristol Veterinary School, University of Bristol, Langford, U.K.

2School of Education, University of Bristol, Bristol, U.K.

3Royal Veterinary College, Hatfield, U.K.

Received 21 October 2021

Initial acceptance 16 February 2021

Final acceptance 20 December 2021

MS number 20-00776R

\*Corresponding author.

E-mail addresses: e.paul@bris.ac.uk (E. Paul); cnicol@rvc.ac.uk (C. Nicol).

To assess the welfare of captive animals, validated measures, so-called ‘welfare indicators’, are required. We used a triangulation approach to investigate the extent to which different measures converged to provide corroborating evidence of welfare. Laying hens were exposed to living conditions designed to be generally preferred (GP) or generally nonpreferred (GNP), using previous studies of chickens’ majority preferences for resources and environments. The hens were also tested at the end of the study to identify their individual preferences for these living conditions, assigned to groups that showed an individual preference, or nonpreference, for their own experimental housing, regardless of whether it was generally preferred or not (IP and INP). Both GP and IP birds showed more ground-foraging behaviour, and lower pulse rates during handling, than GNP and INP birds. Individual preference was associated with more optimistic-like judgement biases when birds were tested after 6 weeks of exposure to these living conditions, but not after 24 weeks. Serum blood glucose levels were also lower in hens showing individual preferences for their experimental living conditions. General preferences were associated with a number of measures, including higher rates of ground-foraging behaviour and lower faecal water content (after both 6 and 24 weeks), lower pulse rate during handling and greater tibia strength and stiffness post mortem. There were no associations between judgement bias and other candidate welfare indicators, but it is not clear whether this represents evidence of absence or merely absence of evidence. Overall, the different approaches did not converge to identify a precise state of animal welfare, although some measures (preference, stress indicators: pulse rate, faecal water and blood glucose, foraging behaviour) were aligned consistently across timescales. We conclude that further work is needed to establish which alternative measures of affective state might be more appropriate indicators of animal welfare.

***Keywords***

*animal welfare; chicken; cognitive bias; environmental enrichment; judgement bias; preference*

How can we assess good and bad welfare in animals? This is a key question for animal welfare science, and a variety of answers have been offered across the relatively short history of the discipline (Dawkins 2006). Currently, a range of different theoretical approaches are employed, yet there has been little examination of their levels of convergence. Here, we consider the welfare assessment of farmed chickens, *Gallus gallus domesticus*, the most populous domestic species on the planet (chickens numbered an estimated 23.7 billion individuals in 2018; FAO, 2019). In a longitudinal study, we made use of a novel triangulation approach to investigate the relationship between judgement bias measures, other putative welfare measures (behavioural and physiological ‘candidate welfare indicators’), and two alternative conceptions of preference. Our starting point was an approach to animal welfare that considers the affective state of the animal to be its central concern: the so-called ‘subjective’ or ‘affective’ view (Dawkins 2015; Duncan 1996; Mason & Mendl 1993; Mendl et al. 2017; Browning 2020). We made use of the established preferences of hens, based on previously published research, to construct living conditions for our experimental birds that could be classified as producing generally ‘good’ and ‘bad’ welfare (i.e. more and less positive affective states for most hens). That is, we made use of preference data to generate affect-focused, operational classifications of welfare. We also measured the birds’ own preferences for these living conditions, establishing an additional, individual preference-based classification of good and bad welfare, to provide a related, yet alternative comparison.

Previously, we have taken animal preferences to be a ‘ground truth’ reference point, and examined how these are associated with other behavioural, physiological and physical measures that are proposed to be relevant to welfare (Nicol et al. 2009; 2011). Newer measures of affective state, such as judgement bias (Harding et al. 2004; Paul et al. 2005; Mendl et al. 2009), potentially provide alternative reference points for assessing welfare, but can also be validated against preference and/or ‘willingness to work’ measures of affect (Mendl et al. 2017; Mendl & Paul 2020). However, all measures of preference, work and affective state are subject to error, and it could be argued that no pure ‘ground truth’ is practically available in the assessment of animal welfare. The aim of the study was therefore to establish whether we could identify an animal’s welfare state in the absence of any single ground-truth measure. Using a navigational analogy, we adopted a triangulation approach, which we describe in further detail below. In its loosest sense, triangulation has become an important focus of research in other conceptually difficult areas of social, health and behavioural science (e.g. Ostlund et al., 2011; Caillaud et al., 2019) where it refers to the practice of using multiple data sources and methods to look for convergent patterns in results. Triangulation studies can thus offer insights into the weaknesses, biases and sometimes contradictory findings that have arisen previously from single-method or single-measure studies.

*<H2>Defining and Assessing Animal Welfare*

Prior to the onset of mass public concern for animal welfare, physical health and productivity were considered by many as ‘good-enough’ confirmations of the wellbeing of most captive animals (e.g. Sainsbury, 1986; Taylor 1972; see also Fraser et al. 1997; McGlone 1993). But the sometimes-severe disruptions of behaviour caused by captivity in general, and by some intensive farming practices in particular, led to the increasing involvement of ethologists in the study of animal welfare, and to conceptions of welfare that combined behavioural and physical components (e.g. Brambell 1965; Broom, 1986). Within this expanded view, ‘welfare indicators’ could be measures of physical health (including disease, injury, body condition, productivity), of behavioural wellbeing (e.g. the presence of a range of activities including feeding, drinking, sleeping, playing, etc.) and of the absence of abnormal or deleterious behaviour patterns (including stereotypic pacing, bar biting, feather pecking, etc.). A number of behaviour-focused welfare researchers have emphasized the importance of naturalness in animal welfare, arguing that the opportunity to experience natural environments and to perform natural behaviour patterns is key (e.g. Barnard & Hurst, 1996; Duncan and Fraser 1997; Kiley-Worthington 1989; see also Browning 2020 for a contemporary critique). But more influential still has been the development of approaches to welfare that take an essentially subjective view (Dawkins 1990; Duncan 1993; Duncan 1996; Fraser & Duncan 1998; Mason & Veasey 2010; Mason & Mendl 1993; see also Fraser et al. 1997; Browning 2020). These are ultimately predicated on the presumption of animal sentience. If an animal is sentient (i.e. has the capacity to experience feelings and to suffer), then its welfare is inextricable from its subjective experience, and its experience of valenced affective states in particular (i.e. Dawkins 2015; Mendl et al., 2010; Mendl et al. 2020). Within this approach, the central goal of animal welfare assessment is to establish, as reliably as possible, the individual’s affective state, even if it is not possible to directly access or even confirm objectively the presence of conscious affective experiences (see Paul et al. 2020; Panksepp 1998 for discussions of affect and consciousness in animals).

In recent years, theory and research in animal affect have developed considerably, especially within neuroscientific disciplines (for reviews see Anderson & Adolphs 2014; de Waal 2011; LeDoux 2012; Paul & Mendl 2018; Raoult et al. 2017; Rolls 2014). In the context of animal welfare, a range of new methodologies aimed at assessing affective states have been developed; these have considered both short- and long-term (mood-like) affective states. A growing emphasis on validation has emerged, with a number of theorists and experimentalists highlighting the value of defining animal welfare explicitly and/or operationally when designing validation studies (e.g. Mason & Veasey 2010; Ahloy-Dallaire 2018). Of these, preference and ‘willingness to work’ (i.e. motivation) tasks, with their close links to reinforcement-based definitions of affective state, have been shown to have an important role in confirming the validity of putative welfare indicators (e.g. Diez-Leon et al. 2016; Nicol et al. 2009; Reijgwart et al. 2018).

*<H2>Preference, Work, Affect and Welfare*

Measuring an animal’s relative preferences for, or willingness to work for, particular resources, environments or stimuli provides a key source of information on situational valence, that is, whether something is perceived by the animal as relatively positive (attractive) or negative (aversive) (e.g. see Dawkins 1990, Mason et al. 2001). Across many years, a number of reinforcement learning-based, operational definitions of emotions (or ‘affective states’ or ‘positive and negative states’, the exact terminologies vary between authors) have been proposed (e.g. Amsel 1958, Mowrer 1960, Millenson 1967). Most recently, Rolls (2014) offered a two-stage definition of emotions as ‘states elicited by rewards and punishers’ where ‘a reward is anything for which an animal will work’ and ‘a punisher is anything that an animal will work to escape or avoid’ (Rolls 2014, p14; see Paul & Mendl 2018 for discussion of this definition and Mendl & Paul 2020 for an updated version). The logic is simple: animals have evolved to select (and strive for access to) situations that are beneficial to their fitness, and to avoid (and strive to escape from) those that are harmful. In humans, such situations are associated with valenced, subjective, affective states, that is, felt emotions or moods. So, in species that have a capacity for consciousness, the states defined above may be experienced as something akin to human experiences of ‘pleasant’ or ‘unpleasant’ (Paul et al., 2020). Within this view, preference and ‘willingness to work’ tests can be used to assign relatively ‘positive’ and ‘negative’ affective states in an operational manner (Mendl et al. 2017; Mendl & Paul 2020). In this way, animal preferences and willingness to work decisions (e.g. see Hughes and Black, 1973; Barnard, 2007; Dawkins, 1990; Fraser and Nicol, 2018; Mason et al. 2001) can offer researchers the most direct insight into animal affect, and can be used to validate measures of welfare when it is conceptualized according to the ‘subjective’ tradition.

*<H2>General and Individual Preference in Chickens*

Preference tests for resources and environments have been widely employed to develop an understanding of chicken welfare (for review see Weeks & Nicol 2006), to make improvements to the housing of farmed birds and to guide legislation. These applications of welfare research have ultimately relied on the subjective view of welfare outlined above, in which preference and willingness to work are taken as markers of (relative) affective valence, and therefore welfare. But while such preference experiments have proven extremely valuable, some potential problems remain. Most notably, it is inevitable that not all individual animals agree on what resources or environments they favour. This may be the result of minor idiosyncrasies of preference, or developmentally determined differences in reward or punishment valuation (e.g. early life experience of particular coloured or flavoured foods can determine later preferences, Gentle 1972). But preference for certain types of environment may also be stratified according to differing individual experiences of those environments, determined by social experience and factors such as ‘personality type’ (which may have a considerable influence on an individual’s wellbeing within that environment; see Nicol et al. 2009).

In previous studies of domestic chickens (Nicol et al., 2009; Nicol, Caplen, Statham et al., 2011), we have adopted a systematic and quantitative approach to validate welfare indicators against measures of bird preferences. We have also assessed variability in preference. We found that birds made consistent, transitive choices when tested repeatedly across different living conditions and over relatively long periods of time (Browne et al., 2010). But these birds did not universally agree on which environment they preferred; some persistently diverged from the general preference profile (i.e. the one favoured by the majority of birds). Moreover, we found that, irrespective of the nature of the living conditions selected by an individual bird as preferred (i.e. whether it was generally preferred (GP) or not), individual preference (IP) was associated with a variety of behavioural and physiological indicators (e.g. head shaking and self-scratching, lower faecal water content, lower corticosterone and lower heterophil:lymphocyte (H:L) ratios, more self-preening; Nicol et al., 2009; Nicol, Caplen, Statham et al., 2011). In addition, a number of measures previously believed to be valid indicators of chicken welfare were not associated with either general or individual preference (Nicol et al., 2009; Nicol, Caplen, Statham et al., 2011). In sum, these studies indicated that measures thought to assess good and bad welfare (i.e. welfare indicators) have not always been well validated.

*<H2>Judgement Bias*

Welfare research has employed a wide variety of measures, including HPA-axis activation, immunological defence mechanisms, body repair mechanisms, behavioural time budgets and behavioural responses to a variety of tests and challenges. Such measures have been used to assess welfare in all areas of human use and interaction with animals, including animal production (Dawkins et al., 2004; Ferrante et al., 2019; Pfeiffer et al., 2019), laboratory animal science (Littin et al., 2008; Hager et al., 2018), biological conservation (Chaves et al., 2019), zoo biology (Salas et al., 2016) and food safety (Iannetti et al., 2020). Thus, a number of potentially valuable welfare indicators exist. These are possible to measure through behavioural observations, behavioural tests, physiological and physical tests, post mortem tests, etc., and a few have been validated against the (general) preference or willingness to work classifications of affect and welfare outlined above (e.g. see Ross 2019). However, problems remain. For example, Alm et al. (2016) found that the pattern of the stress response in hens varied between measures, and correlations were generally few and inconsistent, highlighting the complexity of the relationship that exists among potential welfare indicators. In sum, it seems likely that many existing measures of welfare in fact target several different facets of the physical and psychological wellbeing of animals.

In the present study, we were keen also to include consideration of a marker of welfare that integrates, and summarizes across inputs, the different aspects of an animal’s life experience and thereby, perhaps, the differing facets of the welfare construct. This is the key hypothesized feature of affective states themselves, as integrative cognitive summaries of valenced (reward and punishment) experience (Mendl & Paul 2020; Paul et al. 2020). We decided to use the judgement bias test, which some of the present authors originally developed (Harding et al. 2004; Paul et al. 2005; Mendl et al. 2009), and which is now commonly used to assess animal affect and welfare. If the core goal of welfare measurement is affective state measurement (Browning 2020), then a tool such as the judgement bias task, which explicitly aims to measure affective state, is an important one to investigate. Other potentially integrative indicators of affective state in chickens have also been developed (e.g. startle responses; Ross et al. 2019), but were not available to us when this study commenced.

The judgement bias task is based on a discrimination learning task in which animals are trained to respond differentially to cues that predict rewarding or punishing/less-rewarding outcomes (e.g. tones varying in pitch or coloured images varying in hue or colour saturation). Relatively ‘optimistic’ or ‘pessimistic’ responses to ambiguous probe cues (i.e. where probe cues, intermediate in tone/colour/etc., are responded to as if they predicted reward or punishment) are interpreted as indicating more or less positive affective states (Harding et al. 2004; for recent reviews, see Lagisz 2020; Neville 2020; Roelofs 2016; see also Appendix 1 for further details about the judgement bias task used here and its validation). In the past decade, a number of judgement bias studies in chickens have been published, but findings have been inconsistent. Deakin et al. (2016), for example, found that chickens exposed to a preferred, warm (ca. 29 oC) environment showed more ‘optimistic’ judgement biases than when the same birds were exposed to a colder (ca. 20 oC), less preferred environment and Iyasere et al. (2017) found that corticosterone-treated broiler chicks (an established method of stress/anxiety induction) were more likely than controls to show ‘pessimistic’ judgement biases (see also Salmeto et al. 2011 for a finding of pessimistic-like responses in anxiety-induced chicks). Wichman et al. (2012), however, found no judgement bias differences between laying hens housed in enriched and basic pens, and Hernandez et al. (2015) found no judgement bias differences between stressed (socially isolated in a novel environment) and nonstressed hens (see also de Haas et al. 2017 and Ross et al. 2019 for equivocal findings). In sum, while judgement bias tests have proven to be valuable and potentially informative welfare indicators across a wide range of species (especially mammals), their role in assessing chicken welfare remains unclear, perhaps in part because of methodological differences between studies, and lack of consistent validation of particular methodologies. A key aim of the present study, therefore, was to investigate whether judgement bias can be validated as an indicator of affect and welfare in domestic chickens, using both types of living conditions preference described above (GP and IP), as well as investigating how it can contribute to a triangulation of the welfare state itself (see below).

*<H2>Other Candidate Welfare Indicators*

The selection of other potential welfare indicators for the current study was based on three principles. First, we included indicators used in our previous studies (Nicol et al., 2009; 2011), as an important exercise in reproducibility. We selected, in particular, indicators that were quick to obtain from large numbers of birds and/or that had shown previous relationships with measures of preference. For these reasons we included (1) measures of body, plumage, foot and comb condition which have also been used extensively in other applied studies of poultry welfare and in on-farm audits (e.g. Welfare Quality, 2019) and (2) ‘classic’ indicators of poultry stress such as blood corticosterone, osmolality, glucose and heterophil:lymphocyte ratio (Gross and Siegel, 1983;) as well as faecal dry matter (Puvadolripod and Thaxton, 2000) which had shown a strong previous relationship with measures of preference (Nicol et al., 2009; 2011). However, we omitted some expensive or time-consuming blood parameters that we had used previously, such as creatine kinase (an indicator of muscle damage, Mitchell et al., 1992), lactate and the acute phase protein alpha-1 acid glycoprotein (an indicator of immunological stress in poultry, Tuyttens et al., 2008), as well as measures that had shown little variation between birds in previous studies such as eye temperature measured using thermography.

Second, we trialled some additional blood parameters including triglycerides and gamma-glutamyl transferase (markers of fatty liver disease, a common condition when hens have reduced opportunities for exercise; Yang et al., 2014; Ronchen et al., 2008) and alkaline phosphatase (ALP; a marker of bone quality; Tarlton et al., 2013). Third, given the focus of the current study on the assessment of welfare over the long term (see below), we devised some novel measures of long-term health and wellbeing, including some based on data from humans. These included novel measures of resting heart rate and blood pressure in hens (Blanchflower and Oswald, 2008) and capacity to deal with physical challenge (Marrocco et al., 2020). We also took post mortem measures of bone quality and keel bone damage, reflecting the importance of these aspects for the welfare of laying hens (e.g. Tarlton et al., 2013; Eusemann et al., 2018). We also predicted that some of these potential indicators might differ between experimental groups after just a few weeks, but others would only differ after several months of cumulative experience (see below).

*<H2>Triangulation of Welfare: Preference, Judgement Bias and Other Candidate Welfare Indicators*

When navigating in unknown territory, back bearings from (at least) three different mapped features can be used to identify a precise physical location (Ordnance Survey, 2021). However, if the features used are not precisely fixed, are inaccurately portrayed on the map, or if bearings are subject to error, then only an approximate location can be identified. Using this navigational analogy, each welfare assessment method or data stream can be viewed as a potential bearing helping to ‘locate’ an animal’s welfare state (Fig. 1a, b). The approach illustrated here takes an alternative view to the traditional one in which welfare is defined in an a priori fashion; in essence, welfare is ‘found’ rather than assigned. But whether welfare can indeed be found in this way remains an empirical question. In the scenario illustrated in Fig. 1a, results obtained using general and individual preference are identical, and corroborating evidence of welfare state is obtained from judgement bias and other candidate welfare indicators. As a result, the ‘location’ of welfare state can be pinpointed with accuracy. In Fig. 1b, the results obtained from the different types of preference are similar but not identical, while results from judgement bias and other candidate indicators suggest the same approximate location but with far greater uncertainty. Any candidate measure that does not contribute to the triangulation can be discarded as a welfare indicator. But in this scenario, the question remains as to whether a single welfare construct should still be regarded as a meaningful (locatable) goal.

Our first aim was to improve our understanding of chicken welfare assessment by investigating whether judgement bias is associated with the birds’ general preferences regarding their living conditions and/or their individual preferences. Our second aim was to investigate whether the other potential welfare indicators measured were also significantly associated with chickens’ general preferences regarding their living conditions and/or their individual preferences. Because individuals may vary in their true preferences, we propose that individual preferences will be better aligned with other candidate welfare indicators than general preferences. However, we regard this as a weak prediction because individual preferences measured by point sample testing are subject to errors arising due to misidentification, shifts in preference over time or the intermittent imperative to check that conditions have not changed in the less preferred environment (Nicol, 1986). Such sources of ‘error’ will be reduced for general preferences based on previous studies, where variation across multiple individuals and experiments will be averaged. For practical purposes, it is important to know the extent to which individual and general preferences align with one another (and with welfare indicators), as it is harder to measure individual preferences and plan environments that take account of them. It is also important for our understanding of welfare as a whole to appreciate potential differences between individual and general preference; for example, it is possible that individual preferences have a more direct association with subjective (affective) welfare while general ones may have closer links with physical welfare.

Our third aim was to investigate potential associations between judgement biases in hens and other potential welfare indicators. As far as we are aware, no comprehensive assessment of the associations between judgement bias and a broad range of other measures has been conducted. In the present study, we hypothesized that some, but not all of the welfare indicators considered would correlate with judgement bias performance. For example, it is possible that judgement bias may be more closely associated with social behaviour welfare indicators (e.g. receiving aggressive attacks from other hens; see Wichman et al. 2012) than with some more traditionally employed physical or physiological indicators. Indeed, welfare may not be easily conceptualized as a unitary construct at all (i.e. with a single group of correlated indicators); it is possible that certain indicators cluster together, perhaps pointing to largely separate facets of the welfare construct.

*<H2>Longitudinal Study Design*

The present study employed a longitudinal study design in which the chickens’ judgement biases and other potential welfare indicators were measured on three occasions: prior to experiencing preferred and nonpreferred living conditions (baseline) and following short-term (6 weeks) and long-term (24 weeks) exposure to these conditions. By including a premanipulation baseline phase, it was possible to statistically control for existing variation in the measures taken. By including short- and long-term assessment phases, we also aimed to differentiate welfare indicators that are relevant to relatively brief experiences of preferred/nonpreferred conditions from those that pertain to much longer experiences. Many studies of welfare assessment are relatively short term, sometimes because the lives of the production animals in question are also short (e.g. broiler chickens), and sometimes simply because of practical and cost limitations of longer studies. Laying hens are farmed for considerably longer than broilers, and we hypothesized that preferred and nonpreferred living conditions may have different effects on the potential welfare indicators of such birds, depending on the length of period considered. In sum, the triangulation of welfare among these birds might shift and change across time.

**<H1>METHODS**

*<H2>Ethical Note*

All work was approved by the University of Bristol Animal Welfare and Ethical Review Body and conducted under U.K. Home Office Licences (PPL: 30/2779 and 30/3392). Animal use and care was in accordance with the Animals (Scientific Procedures) Act 1986, EU directive 2010/63/EU and UK Home Office code of practice for the housing and care of animals bred, supplied or used for scientific purposes. All husbandry and experimental procedures were designed to avoid and minimize distress, with procedures in place to terminate any task or measurement in the event of a hen becoming agitated or fearful. One bird was treated for crop infection/impaction during the study; she was temporarily withdrawn from the study but did not survive.

*<H2>Birds and Husbandry*

The birds were 60 medium-brown commercial laying hens sourced from a local commercial breeder in two separate batches (30 birds per batch; approximately 1 year apart), at 18 weeks of age. On arrival at the laboratory, they were immediately housed in groups of five birds per pen and remained in these groups for the duration of the study. All birds were wormed, treated prophylactically for mites, individually marked with coloured leg bands and coloured stock-marker spray paint. Throughout the study, they all received the same ad libitumfeeding regime (Farmgate Layers’ Mash, BOCM Pauls, Ipswich, Suffolk, U.K.), were kept on a 12:12 h light:dark cycle (light period 0700–1900 hours) and in an ambient temperature of 19–21 °C. Eggs were collected daily, pens were cleaned weekly, and the birds’ general health and wellbeing were monitored daily by professional animal caretakers.

*<H2>Experimental Design*

The study took place across three phases. In Phase 1, all birds were kept in identical, ‘intermediate’ living conditions (so called because they were intermediate between the ‘majority preferred’ and ‘majority nonpreferred’ experimental living conditions of the later part of the study; see Fig. 2) for training and baseline data collection. They were then moved to generally preferred (Group GP) or generally nonpreferred (Group GNP) living conditions for Phase 2 (short-term exposure to the experimental living conditions, 6 weeks plus 3 weeks testing) and Phase 3 (long-term exposure, a further 15 weeks plus 5 weeks testing). The GP and GNP living conditions were designed according to previously published experimental evidence regarding chickens’ preferences for environments, resources and events (see Appendix 2 for full details).

*<H2>Phase 1. Training and Baseline Data Collection*

During the initial 10 weeks of the experiment, a subset of the birds (*N*=24; 12 per living condition, two per pen) were trained to perform the judgement bias task. This was a Go/No go screen peck task developed specifically for chickens (see Appendix 1 for full training details and preliminary analyses). In the following 2-week period, behavioural and physiological ‘welfare indicator’ data were collected, and trained birds were given their first set of judgement bias tests.

*<H2>Phase 2. Short-term Exposure to Experimental Living Conditions and Data Collection*

This took place across the following 9-week period. Half of the birds were randomly allocated to the GP living conditions (Group GP; *N*=30; six pens of five hens) and the other half were allocated to the GNP living conditions (Group GNP; *N*=30; six pens of five hens). Welfare indicator data were collected for each bird across the final 3-week period of this phase. Trained birds’ second set of judgement bias tests were also conducted during this period. These were carried out following a brief period of refresher training (maximum 3 days; criterion 90% or more correct responses to training stimuli; see Appendix 1).

*<H2>Phase 3. Long-term Exposure to Experimental Living Conditions and Data Collection*

This took place across the following 20-week period, during which time the birds stayed in their same experimental conditions (GP or GNP). Welfare indicator data were collected for each bird across the final 5-week period of this phase, and a third set of judgement bias tests were also conducted during this time, again following a brief period of refresher training (maximum 3 days; criterion 90% or more correct responses to training stimuli; see Appendix 1). All birds’ blood pressures were measured during this period, and preference tests were conducted to establish birds’ individual preferences for their experimental living conditions (as compared with their initial, ‘intermediate’ living conditions). The study was then terminated, and all birds were humanely killed (using a Home Office approved Schedule 1 neck dislocation method) and post mortemexaminations were carried out.

*<H2>Living Conditions: GP and GNP Birds*

All pens were 1.8 m tall and 1.22 m wide, with concrete walls at the back and chicken wire fencing and door at the front (see Fig. 3). The GP living condition pens were 3.06 m long, had 100 mm deep wood shavings on the floor, a dustbath filled with a peat substitute/sand mix (0.8 m x 1.22 m), double perches (1 m long x 300 and 600 mm high), and a double-tier, roofed nestbox with wooden sides, rubber matting bases, a perch at the front and backing made from artificial grass (300 x 300 mm and 830 mm high). These pens also had an area of suspended black fabric cover fixed horizontally 1 m above the length of the perch, to provide a shaded area. Birds in the GP living conditions were exposed to two rewarding stimuli per day (5 days per week, at random times), from the following list: heated resting pad, 50 g portions of sweetcorn, spaghetti, chopped cheese.

The GNP living condition pens were 1.52 m long, had wire floor, no wood shavings or other floor substrate, no dustbath, a single-tier perch (400 mm long x 50 mm high), and a single-tier, roofed nestbox with wire sides and no furnishings (300 x 300 x 830 mm). Birds in the GNP pens were exposed to white noise (80 dB) for 4 h/day. They were also exposed to two mildly aversive or threatening stimuli per day (5 days per week, at random times), from the following list: small plastic inflatable toy lowered into the pen via string from above; 3 min of taped conspecific alarm calls; 3 min of taped dog barks; brief jet of water spray directed at tail feathers.

The intermediate pens were 2.25 m long, had 20 mm deep wood shavings on the floor, a dustbath filled with wood shavings (0.8 m x 1.22 m), a single-tier perch (1 m long x 150 mm high), and a double-tier, roofed nestbox with solid wooden sides but no furnishings (300 x 300 mm and 830 mm high). While in the intermediate pens (Phase 1 of the experiment) birds were not exposed to any rewarding or aversive stimuli.

The data collected at the end of Phases 1, 2 and 3 were divided according to the three approaches to welfare measurement outlined above: Preference, Judgement Bias and Candidate Welfare Indicators.

*<H2>General Preference and Individual Preference*

Birds’ preference groups were classified in two ways. First, as outlined above, birds were randomly assigned to Group GP and Group GNP: these were exposed to living conditions made up of components previously found to be more and less preferred, respectively, by the majority of chickens tested (see Appendix 2). Second, they were also classified as having individual preferences on the basis of T-maze preference tests conducted at the end of Phase 3 of the experiment: Groups IP and INP. These tests required the birds to make a series of seven individual choices between entering a pen resembling their own experimental living condition (GP or GNP) and entering one resembling the intermediate living condition they experienced during Phase 1 of the study. Each bird’s individually preferred pen type was the one that it chose across the majority of choices made (for full methods and detailed results, see Appendix 2).

*<H2>Judgement Bias*

Two hens per pen (totalling *N*=12 birds per experimental housing group) were trained and tested using the judgement bias apparatus and methodologies developed by Deakin et al. (2016). Hens were required to peck a screen (Go) in response to S+ colour cues to obtain a single mealworm reward, and to refrain from pecking the screen (No go) in response to S- colour cues to avoid receiving a single air puff punishment. The S+ and S- cues were centrally presented 35 mm diameter orange circles of differing levels of colour saturation. For half of the birds, the S+ cue was of high saturation orange and for the other half it was of low saturation orange; the S- cue in either case was of the opposite saturation (low or high).

During testing, three unrewarded, ambiguous probe cues of intermediate saturation were presented in addition to the S+ and S- anchor cues. These cues were of three types: ‘near negative’ (nS-); ‘middle’ (M); ‘near positive’ (nS+). Test sessions (one per day) were performed at the end of each experimental phase. Each test session comprised 40 stimuli presentations (17x S+ cues; 17x S- cues; 6x ambiguous cues, two of each type).

Two judgement bias summary scores, ‘Proportion of ambiguous probe cues pecked’ during test sessions (out of a total of 24 probe cues presented) and ‘Mean latency to peck ambiguous probe cues’ were calculated for each bird at each data collection point. Because these two scores were highly correlated, only results for the former are reported here (see Appendix 1 for full details of judgement bias test methodology and analyses).

*<H2>Candidate Welfare Indicators*

*<H3>Physical examination, blood and faecal tests*

All hens were weighed and the length of the middle claw on the left foot was measured. Body condition was scored using the five-point system described by Gregory and Robins (1998); because of the relatively little variation observed, this was collapsed to a binary variable (keel bone palpable/not palpable). Other measures included: presence or absence of foot lesions and comb lesions; comb size (0 <3.5 cm2; 1 > 3.5–5.5 cm2); comb elevation (0/1; upright/low-floppy); comb colour (0/1; bright/pale).

For the blood samples, 4 ml of blood was collected from a wing vein using a 23-gauge 16 mm needle, split between a Serum Collection Vacutainer Tube and an EDTA Blood Collection Vacutainer Tube (BD Vacutainer) and immediately put into ice storage. Two blood smears were made on glass slides and counts of heterophils (H) and lymphocytes (L) taken, to calculate the H:L ratio (Gross & Siegel, 1983). Samples of whole blood were analysed for levels of glucose, triglycerides, alkaline phosphatase (ALP) and gamma-glutamyl transferase (GGT) (Langford Diagnostics Laboratories, Bristol, U.K.). Blood samples collected for serum extraction were packed in ice overnight and then centrifuged at 1200 rpm for 15 min and the serum stored at -18 °C until analysis (corticosterone: Cambridge Specialist Laboratories, Cambridge, U.K.). All hens were visually assessed by an operator up to 2 h following blood sampling to ensure that they were showing no signs of distress.

To obtain a faecal sample, each hen was placed in an individual wire cage with a plastic sheet floor, with at least one other hen visible in an adjacent cage. Hens were left quietly until they had produced a faecal deposit (no longer than 60 min). Faecal samples were frozen in 50 ml plastic tubes at -20 °C. When all samples had been collected, they were desiccated for 3 days in a freeze drier to obtain a measure of bulk water content.

*<H3>Behavioural observations*

Video footage was taken of all pens for 1 h at two time points (starting: 0700 and 1500 hours) across 3 days during each data collection period (Phases 1, 2 and 3; i.e. 6 h of video for each phase). It was retrospectively coded by three observers (assigned 1 day each per bird, to avoid systematic observer bias) using a detailed ethogram (derived from Nicol et al., 2009) and specialist software (Observer XT 10, Noldus, Wageningen, Netherlands).

Two types of behavioural observations were conducted: Scan observations and Focal observations. Scan observations were used to provide an overview of the birds’ time budgets for all common behaviours (1% or more of all hens’ time observed: Feed from Hopper, Ground Forage (Ground Peck and Ground Scratch), Stand Alert, Preen, Walk, Nest, Drink). The behaviour patterns of each bird were recorded instantaneously at 10 min intervals (7 scans/h) for all 6 h of video recordings per phase. Because the initial scan of each recorded day was at the moment that lights came on (0700), it was deleted from the analysis, leaving a total of 13 scan points per day (39 per bird per phase). Scan data reported here are expressed as a percentage of all scans (per bird per phase) in which a behaviour pattern was observed.

Continuous focal observations of each bird made use of the middle 30 min of the same set of videos (0715 - 0745 hours and 1515 - 1545 hours); these continuously recorded observations were used to measure the frequencies and durations of social, emotional and aggressive behaviours not commonly observed within the scan sampling regime (Aggressive attack given, Aggressive attack received; Severe feather peck given, Severe feather peck received; Mild feather peck given, Mild feather peck received; Beak peck given, Beak peck received; Shake head; Dustbathe/Sham dustbathe). For each behaviour pattern witnessed, the frequency and length of each individual bout was recorded, and the mean frequency and mean bout length were calculated (per 30 min recording). To enable a greater accuracy in assessing bout length, the observation period was terminated at the end of the last behaviour scored (i.e. beyond the 30 min period). Focal data were recorded in the final data set as averages of all 30 min observation periods per day.

*<H3>Physical challenge tests*

These two tests were designed as measures of hens’ responses to physical challenges, in essence, decision tasks that required the hens to make stay/go choices based on assessments of their own physical capacity. We assessed physical challenge in two contexts, the first positive and the second negative. The High Perch test considered hens’ willingness to jump down to the floor, from a 1 m high perch, for a six-mealworm reward. It has been found that the ability of hens to jump for food is compromised if they are in poor physical condition or are in pain (Nasr et al., 2012; 2015). Mean latencies to jump across three successive tests were calculated (maximum duration of test: 180 s). The Water Box test assessed the mean latency of hens to jump out of a 6 cm deep plastic box containing shallow water (ca. 1 cm), across three successive tests (maximum duration of test: 180 s) Using similar tests, it has been found that broiler chickens are slower to avoid or escape aversive contact with water if they experience painful leg conditions (Weeks et al., 2002).

*<H4>High Perch Test.* A perch (wooden doweling: 1 m length x 3.5 cm diameter) was positioned in an experimental test room at a height of 1 m. A visible food reward (a ceramic bowl containing six mealworms) was placed on the floor 2 m from the perch. Six mealworms were used in order for the reward to be clearly visible from the perch, but not so large as to risk satiating the hens (in the mealworm eating test, hens would consume up to 25 g of mealworms in 1 min, without showing signs of satiation). Each hen was placed individually on the mid-point of the perch, facing the bowl and the latency to fly down to access the reward was recorded over three consecutive trials. Once mealworms were consumed, birds were immediately replaced on the perch to begin the next test. Each test had a maximum latency of 180 s. Birds that failed to leave the perch within 180 s were placed on the floor next to the bowl (within visual but not physical contact with the mealworms), and immediately replaced on the perch to start the next test. Birds’ mean latencies to jump down across the three tests were calculated.

*<H4>Water Box Test.* A white plastic box (35 × 27 cm and 25 cm high) was filled to a depth of 4 cm with room temperature water in the experimental test room. Each bird was placed individually in the box, and the latency to fly out (to stand either on the side of the box or on the floor) was recorded over three consecutive trials. Birds were immediately replaced in the water to begin the next test. Each test had a maximum latency of 180 s. Birds that failed to leave the box within 180 s were removed from the water, placed upon the floor next to the box, and immediately replaced in the water to start the next test. Birds’ mean latencies to leave the box across the three tests were calculated.

*<H3>Blood pressure tests (Phase 3)*

A novel method for measuring blood pressure in chickens was developed using equipment originally designed for blood pressure monitoring of companion animals in the veterinary setting. A specially designed cuff was applied to the tarsometatarsus of the right leg of each hen and high-definition oscillometry was used to record pulse rate, systolic, diastolic and mean arterial blood pressure from each hen as they were gently restrained on the lap of an experimenter (equipment manufacturer: Vet MD PRO). This was repeated five times in quick succession and the mean of each measure calculated.

*<H3>Post mortem measures (Phase 3)*

Using external examinations, we measured each bird’s feet, legs and wings to assess physical symmetry (middle toe length, middle claw length, hock joint diameter, tail feather follicle length). These were calculated as the unvalenced difference between left and right (mm); L/R differences of >2 mm were classified as asymmetric (middle claw length >1 mm). Following dissection, the lengths of both tibia and radius were also measured to assess symmetry. The body cavity fat, liver and spleen were weighed (g) and the keel bone was assessed for damage and scored using a five-point scale: (0 = no damage to 4 = severe break; Tarlton et al., 2013).

*<H4>Bone mineral density.* Both the tibia and the keel were stored at -20 °C for later bone mineral analysis and biomechanical strength testing. Quantification of bone mineral density (BMD) was performed on the keel bone and right tibia using dual energy X-ray absorptiometry (DEXA, Lunar PIXImus densitometer, Lunar Corp, Aarhus, Denmark). Following the marking of the midsection a 0.08 cm2 region of interest (ROI) was measured adjacent to the midsection of each bone and the BMD recorded.

*<H4>Bone biomechanical properties.* The left tibia underwent three-point bending to failure using an Instron 6022 materials testing apparatus (Instron, High Wycombe, U.K.) as previously reported (Knott et al., 1995; Tarlton et al., 2013). Tibia bones were mounted across a supporting bridge with a gap of 4 cm, and perpendicular load applied to the centre point at a speed of 2.5 mm/min. The strength (ultimate stress) at breakage and stiffness (Young’s modulus of elasticity) of tibia exposed to three-point bending were recorded. Keel bones were sectioned with a band saw and positioned on a custom support within the same apparatus. They were then loaded to failure at a constant speed of 50 mm/min using a 6.8 mm blunt probe at the manubrial spine and the strength and stiffness at breakage were recorded.

<H2> Analyses

Following preliminary analyses of preference and judgement bias data, three main sets of analyses were conducted: associations between preference and judgement bias, associations between preference and other candidate welfare indicators and associations between judgement bias and other candidate welfare indicators.

*<H3>Associations between preference and judgement bias*

Multilevel models were used to investigate links between the hens’ judgement bias scores (proportion of ambiguous probe cues pecked) and their preference groupings (GP or GNP; IP or INP). These models nested hens within pens and adjusted for baseline judgement bias data collected at the end of Phase 1 of the study (see Fig. 2). Full details of the models used are given in the next section.

*<H3>Associations between preference and other candidate welfare indicators*

Welfare indicator data were divided into continuous and binary (categorical) variables. To compare the preference groups (GP versus GNP and IP versus INP) in terms of their values for each continuous welfare indicator (physical tests, observations, physical challenge tests, blood pressure, post mortem), we used multilevel models, with hens nested in pens, and adjusting for baseline data (Phase 1). For continuous welfare indicators (with normal distributions), a model of the following form was used:

$$WI\_{ij}=α+βGroup\_{ij}+γWIbase\_{ij}+u\_{j}+e\_{ij}$$

Here, $WI\_{ij}$ is the value of the welfare indicator for hen *i* in pen *j,* during one of the later data collection periods (at the end of Phases 2 or 3). $Group\_{ij}$ indicates the preference group for hen *i* in pen *j* (GP or GNP; IP or INP),while $WIbase\_{ij}$ is the value of the welfare indicator for hen *i* in pen *j* at the first data collection period (end of Phase 1; baseline). The model has random effects $u\_{j}$ and $e\_{ij}$ for pen and hen, respectively, that are assumed to be normally distributed and capture variations due to individual pens and hens. The term of interest is $β,$ which captures the difference between the two preference groups having controlled for baseline measures and the clustering due to pens.

When the welfare indicator is a binary variable then an equivalent multilevel logistic regression model is fitted:

$$WI\_{ij }\~ Bernouilli \left(π\_{ij}\right), logit(π\_{ij})=α+βGroup\_{ij}+γWIbase\_{ij}+u\_{j}$$

where again interest lies in $β$ which captures the differences between the two preference groups having controlled for baseline measures and the clustering due to pens.

 To present the results of these analyses, we list in the text comparisons where β is significant, i.e. where there is a significant difference between the groups. In the tables giving more details of these results (Tables 1 - 4), we show the means (or proportions for binary variables) of the welfare indicators for the preference groups at all three time points, rather than giving two sets of α, βand γ parameter values. This is because β indicates the difference in increase (of the welfare indicator) between the groups over the two time points (Phases 1 - 2 and 1 - 3 depending on comparison). We are principally interested in situations where this results in a difference occurring at the later time points, due to the influence of the experimental living conditions (i.e. where the preference groups start to diverge) rather than a difference at Phase 1 (baseline) which then results in a coming together at the later time point; this is easier to see from the group means/proportions than the model parameters.

*<H3>Associations between judgement bias and other candidate welfare indicators*

For these analyses, hens’ judgement bias scores ‘proportion of ambiguous probes pecked’ were used in a standardized form, to control for any effects arising from the relative saturations of the S+ and S- training cues. We found that those birds trained to associate high saturation orange-coloured cues with a reward responded with a significantly greater proportion of ambiguous probe cues pecked (see Appendix 1). Because the S+ cue colour was balanced evenly across the experimental preference groups (GP and GNP), this did not have a confounding effect on the models described above. However, in the present analyses, with no group comparisons, steps were taken to standardize judgement bias scores accordingly, using *z* scores based on each training saturation group. *Z* scores standardize data to allow comparisons to be made between different samples; they are calculated as the number of standard deviations that a given raw score is above (+ve scores) or below (-ve scores) the mean of the sample (*z*=(*x*-mean(*x*))/SD(*x*) for a variable *x*).

Analyses were conducted for all judgement bias-trained birds together (i.e. regardless of preference group), on Phase 1 data only. During Phase 1, all birds lived in the same experimental living conditions (‘Intermediate’), so no experimental group effects would influence or interact with any associations found. Because two judgement bias-trained hens were used per pen, multilevel models were used (in which hens (level 1) were nested within pen (level 2)), as in the analyses described above. For these analyses, we used ‘proportion of ambiguous probes pecked (standardized)’ as the response variable, $y\_{ij}$ and the other candidate welfare indicators as predictors individually, with a structure of 24 birds within 12 pens:

$$y\_{ij}=α+βWI\_{ij}+u\_{j}+e\_{ij}$$

Here the estimated effect of the welfare indicator, $β$, gives an indication of the association between the specific welfare indicator and chosen cognitive bias measure.

*<H3>Note on multiple comparisons*

A study of this kind inevitably involves multiple statistical tests and comparisons, increasing the probability of Type 1 errors. This is an important problem, and there are many techniques to try to deal with it, from the ultraconservative Bonferroni correction to the more recent work of Benjamini and colleagues on the use of the false discovery rate to adjust for multiple comparisons (Benjamini & Hochberg 1995; Benjamini & Liu 1999). A particular challenge for the present paper is that potential welfare indicators from previous literature are considered, where on many occasions no such multiple comparisons adjustments were made, or where small studies were used for which no adjustments were required. In a large-scale study such as this, any adjustments made may remove the ‘level playing field’ required for such comparisons across studies. Nevertheless, we acknowledge this limitation and have taken a two-stage approach to avoiding the possibility of over-interpretation as a result of Type 1 errors. First, within the text, we list results significant at *P*<0.05 or below, and across all of these results, we present regression coefficients, which we suggest are as valuable (if not more valuable) than *P* values alone.In addition we perform a Benjamini false discovery rate calculation which returns an alternative threshold below which results are considered significant.

**<H1>RESULTS**

Of the 60 hens included in the experiment (Group GP (*N*=30); Group GNP (*N*=30)), all but one completed the study: a hen from Group GNP died of natural causes (sour crop) during Phase 3 of the experiment. Her data were retained for Phases 1 and 2. Another hen from Group GNP failed to complete the preference test conducted at the end of Phase 3, leaving 29 birds from each experimental group (GP and GNP) completing the preference tests and being assigned as group IP or INP birds (i.e. ones that showed a preference for their own experimental condition or not). There was a small reduction in sample size for some of the blood physiology measures due to technical constraints (e.g. insufficient blood volumes for some tests). The statistical model adjusted for sample size throughout, and any reductions in sample size from the numbers given above are indicated in the text.

In the preference tests conducted at the end of Phase 3, birds from experimental Group GP were significantly more likely than birds from Group GNP to show individual preferences for their own experimental living conditions (73.3% and 25.0%, respectively; 2 13.533, *P*<0.001; for further details, see Appendix 2). On the basis of these tests, 29 birds (22 from Group GP and seven from Group GNP) were assigned to individual preference Group IP and 29 to Group INP (eight from Group GP and 21 from Group GNP).

Analyses of (baseline) judgement bias tests conducted at the end of Phase 1 (12 Group GP and 12 Group GNP birds) confirmed that the birds successfully showed the expected discrimination of S+ and S- cues, pecking the coloured disc that resembled the S+ significantly more frequently than the disc that resembled the S- (*t*23=169.33, *P*<0.001). These birds also showed the expected generalization across the three intermediate, ambiguous probe cues (near S+, middle, near S-), with the middle probe being pecked significantly more than the S- (*t*23=7.859, *P*<0.001) and less than the S+ (*t*23=6.842, *P*<0.001). It can be seen from Fig. 4 that these generalized responses to the probe cues changed little across the course of the study, although the hens did show a small reduction in the number of pecks they made to ambiguous probes across the three test phases (Middle Probe, Phases 1 to 2: β=0.547 (SE 0.138), *P*<0.001; Middle probe, Phases 1 to 3: β=0.733 (SE 0.157), *P*<0.001; *N*=24).

Additional analyses were conducted to check whether there were any associations between feeding motivation and performance in the judgement bias task; from these, it was concluded that there was no evidence of such an association (see Appendix 1 for full analyses and discussion regarding possible associations between feeding motivation and judgement bias tests).

*<H2>Associations Between Preference and Judgement Bias*

*<H3>Group GP versus Group GNP birds*

Judgement bias-trained GP and GNP birds (*N*=12 per group) did not differ in the proportion of ambiguous probes pecked at either Phase 2 or 3. However, GP birds pecked a lower proportion of ambiguous ‘near S-’cues than GNP birds at Phase 3 (β=0.122 (SE 0.058), *P*=0.043; Fig. 5c). GP birds also made a significantly greater proportion of pecks to the ‘anchor’ S+ cues than GNP birds at Phase 2 (β =66.81 (SE 18.28), *P*=0.011; Fig 5b); this difference, however, was very small.

*<H3>Group IP versus Group INP birds*

At Phase 2, IP birds (*N*=9) pecked a significantly greater proportion of ambiguous probe cues than INP birds (*N*=15; β =0.133 (SE 0.065), *P*=0.048; see Fig. 6b). Further analyses revealed that these findings were largely dependent on the IP group pecking the middle ambiguous probe more frequently (β =0.224 (SE 0.084), *P*=0.013). Although a similar trend appeared to remain at Phase 3 (see Fig. 6c), there were no significant differences between the groups at this test point.

*<H2>Associations Between Preference and Candidate Welfare Indicators*

*<H3>Group GP versus Group GNP birds*

*<H4>Physical examination, blood and faecal tests.* Table 1 shows means and SDs for the physiological measures taken across the three phases of the study to help establish the meaning of the significant values of β given in the text below. Table 2 shows the percentages of GP and GNP hens showing a range of physical characteristics.

By the end of Phase 2, Group GP birds had greater body mass than Group GNP birds (β=40.72 (SE 18.8), *P*=0.034). Group GP birds had shorter claws than Group GNP birds (β=3.266 (SE 0.363), *P*<0.001). Faecal water content was lower in Group GP birds (β=2.467 (SE 0.758), *P*=0.002). Blood triglycerides were lower in Group GP birds (log transformed: β=0.142 (SE 0.069), *P*=0.044). Blood glucose levels, corticosterone levels, heterophil:lymphocyte ratios, gamma-glutamyl transferase levels, blood alkaline phosphatase and measures of comb, foot and body condition did not differ between the groups.

By the end of Phase 3, Group GP birds again had greater body mass than group GNP (β=51.32 (SE 23.76), *P*=0.034), shorter claws (β=6.14 (SE 0.602), *P*<0.001) and lower faecal water content (β =4.544 (SE 0.930), *P*<0.001). The proportion of birds with small combs was greater in Group GP birds (β=2.39 (SE 1.165), *P*=0.04). In contrast with Phase 2, blood triglycerides were higher in Group GP birds than Group GNP birds (log transformed: β=0.149 (SE 0.063), *P*=0.042). Blood alkaline phosphatase (log transformed) levels were lower in Group GP birds than Group GNP (β=0.265 (SE 0.091), *P*=0.013). Blood glucose, corticosterone and gamma-glutamyl transferase levels and heterophil:lymphocyte ratio, comb lesions, foot and body conditions did not differ between the groups.

*<H4>Behavioural observations.* By Phase 2, significantly lower levels of ground foraging (pecking and scratching the ground with feet) were seen in Group GNP birds than GP birds (β =13.880 (SE 1.761), *P*<0.001), and by Phase 3, this effect continued to be seen (β =9.648 (SE 2.466), *P*<0.001; see Table 3). At Phase 3, Group GNP birds also showed significantly less sitting behaviour than Group GP birds (β =1.959 (SE 0.818), *P*=0.017). The behaviours feed from hopper, stand alert, walk, preen, nest and drink did not differ between the groups at either data collection point. The focal behaviour observations of social, emotional and aggressive behaviours between individuals also failed to show any significant differences between the groups.

*<H4>Physical challenge tests*. No significant differences were found between the groups in performance in the Water Box test (percentage of birds whose mean latency to exit the water box was 10 s or less (the median latency in Phase 1): Group GP: Phase 1: 60.0%; Phase 2: 56.6%; Phase 3: 60.0%; Group GNP: Phase 1: 43.3%; Phase 2: 53.3%: Phase 3: 62.1%). There were, however, significant group differences in responses to the High Perch test at Phase 2, with GP birds being less likely to jump (β =-1.595 (SE 0.688), *P*=0.020; percentage of birds that failed to jump down from the perch across the 180 s test: Group GP: Phase 1: 53.3%; Phase 2: 76.6%; Phase 3: 80.0%; Group GNP: Phase 1: 53.3%: Phase 2: 50.0%; Phase 3: 69.0%).

*<H4>Blood pressure tests.* When measured at the end of Phase 3, Group GP (*N*=30) and Group GNP (*N*=29) birds did not differ in their systolic, diastolic or mean arterial blood pressures. Group GP birds did, however, have significantly lower mean pulse rates during blood pressure testing (Group GP mean 309 (SD25.71*)* beats/min, Group GNP mean 330 (SD 37.54*)* beats/min; β=-21.366 (SE 8.208), *P*=0.009).

*<H4>Post mortem measures.* Post mortems were conducted on all birds at the end of Phase 3 (GP, *N*=30; GNP, *N*=29). GP birds were less likely to show middle claw length asymmetry (37% of GP birds and 72% of GNP birds; β=1.512 (SE 0.562), *P*=0.007) and tail feather follicle length asymmetry (23% of GP birds and 52% of GNP; β=-1.259 (SE 0.569), *P*=0.027). Body fat composition, spleen weight and liver weight did not differ between groups when body mass was controlled for.

Keel bone biomechanical measurements of birds did not differ between the groups, but the tibias of GP birds had higher bone mineral density (GP: 270.3 mg/cm2; GNP: 248.9 mg/cm2; β=21.430 (SE 8.846), *P*=0.015). The strength (ultimate stress) at breakage and stiffness (Young’s modulus of elasticity) of tibia exposed to three-point bending also showed a significant group effect: GP birds had stiffer (GP: 902.7 MPa; GNP: 796.6 MPa; β=106.139 (SE 48.051), *P*=0.027) leg bones.

*<H3>Group IP versus Group INP birds*

*<H4>Physical examination, blood and faecal tests.* Table 4 shows the means and SDs of physiological measures taken across the three phases of the study for IP and INP birds. Table 5 shows the percentages of IP and INP hens showing a range of physical characteristics.

The mass of IP and INP birds did not differ during Phase 2 or Phase 3. IP birds had shorter claws than INP birds by the end of Phase 2 (β =0.846 (SE 0.388), *P*=0.047). A greater proportion of IP birds had comb lesions in Phase 2 (β =1.574 (SE 0.622), *P*=0.011), but this difference was not sustained into Phase 3.

In Phase 2, IP birds (*N*= 27) had lower blood glucose levels than INP birds (*N*=28; β =0.656 (SE 0.181), *P*<0.001), but there was no significant difference between the groups at the end of Phase 3. Other physical and physiological measures, including faecal water content, did not differ between IP and INP birds across any phase.

*<H4>Behavioural observations.* By Phase 2, significantly lower levels of ground foraging (pecking and scratching the ground with the feet) were seen in Group INP birds than IP birds (β =6.457 (SE 2.082), *P*=0.002), but by Phase 3, this effect was no longer significant (Table 6). The behaviours feed from hopper, stand alert, walk, preen, nest and drink did not differ between the groups at either Phase 2 or 3. The focal behaviour observations of social, emotional and aggressive behaviours between individuals also failed to show any significant differences between the groups.

*<H4>Physical challenge tests.* No significant differences were found between the groups in performance in the Water Box test (percentage of birds that left a box containing shallow water within a mean of 10 s of the start of the test: IP: Phase 1: 55.2%; Phase 2: 51.7%: Phase 3: 62.1%; INP: Phase 1: 51.7%; Phase 2: 55.2%; Phase 3: 58.6%). There were also no significant group differences in responses to the High Perch test, although the IP birds showed a nonsignificant tendency to be less likely to jump down from the high perch (β=-1.154 (SE 0.651), *P*=0.076; percentage of birds that failed to jump down from the perch across the 180 s test: IP: Phase 1: 51.7%; Phase 2: 72.4%; Phase 3: 79.3%; INP: Phase 1: 51.7%; Phase 2: 51.7%; Phase 3: 69.0%).

*<H4>Blood pressure tests.* IP and INP birds did not differ in their systolic, diastolic or mean arterial blood pressures, although these were all nonsignificantly lower in IP birds (mean systolic blood pressure: IP 130.74; INP 138.86; mean diastolic: IP 79.47; INP 80.26; mean arterial pressure: IP 97.92; INP 101.11). IP birds had significantly lower mean pulse rates during blood pressure testing (IP mean 309 (SD 28.64) beats/min; INP mean 329 (SD 35.49) beats/min; β=19.920 (SE 8.270), *P*=0.016).

*<H4>Post mortem measures.* INP birds were more likely to show middle claw length asymmetry than IP birds (34.5% of IP birds and 72.4% of INP birds had asymmetric claws; β=1.653 (SE 0.568), *P*=0.004), while IP birds showed more tibia asymmetry (55.2% of IP birds and 27.6% of INP birds had asymmetric tibias; β=1.219 (SE 0.557), *P*=0.028). Body fat composition, spleen weight, liver weight, measures of tibia and keel bone mineral density and other biomechanical measures did not differ between the groups.

*<H3>Associations Between Judgement Bias and Candidate Welfare Indicators*

*<H4>All birds*. No significant associations were found between birds’ summary judgement bias *z* scores (proportion of ambiguous probes pecked, standardized for reward colour training), and any of the other candidate welfare indicators, when controlling for the two-level structure of hens within pens. Table 7 lists all the analyses conducted, showing regression coefficients and *P* values.

*<H2>Results Overview and Benjamini Corrections*

In this paper we have performed a large number of statistical tests and comparisons and, as is well known, this raises the probability of making Type 1 errors. There are several procedures for adjusting for this including the use of false discovery rate (FDR) techniques as developed by Benjamini and Hochberg (1995). The idea here is that perceived significant results will be a mixture of genuine cases where the null hypothesis can be rejected and so-called false discoveries where the null hypothesis is true, but the sample still suggests we should reject it. It is impossible to distinguish between these two cases and so FDR techniques instead aim to control the proportion of the significant results that are in fact false discoveries. This means that for a specific FDR we adjust the threshold that is assumed to be significant and thus attribute any previously significant results above this new threshold as false discoveries. For the hypotheses considering links between preference groups and other candidate welfare indicators of which there are 168, if we consider an FDR of 0.2 then our significance threshold shifts from 0.05 to 0.0179. If instead we consider a lower FDR of 0.05 the significance threshold reduces further to 0.0024. Table 8 highlights which of the *P* values are still perceived significant at these levels of FDR.

 At an FDR of 0.2 none of the associations found between hens’ judgement bias scores and preference groupings are now below the FDR threshold and the third set of analyses reported above (associations between judgement bias and candidate welfare indicators) are not listed here as no significant associations were found, even prior to Benjamini adjustments.

**<H1>DISCUSSION**

Animal welfare has been studied scientifically for more than 50 years; prompted initially by public concern for the wellbeing of captive and industrially farmed animals (Harrison 1964, Bramble 1965), concepts and definitions of welfare have changed and developed considerably across this time. Today, the so-called ‘subjective’ or ‘affective’ conception dominates: a view in which welfare assessment aims to gauge the ongoing affective states of animals (e.g. Dawkins 2015; Duncan 1996; Mason & Mendl 1993; Mendl et al. 2017). Within this, a range of measures, putative welfare indicators, have been used to assess welfare in a variety of species, and a growing emphasis has been placed on empirically establishing the validity of these (e.g. Nicol et al., 2009; Nicol, Caplen, Edgar et al., 2011; Diez-Leon et al., 2016). While a number of different approaches have been used (e.g. see Mendl et al. 2017), establishing a ‘ground truth’ of good and poor welfare on the basis of animals’ preferences and/or willingness to work for resources and environments is rapidly becoming a key approach for contemporary validation research (e.g. Mason & Veasey 2010; Mendl & Paul, 2020).

While a few studies have explicitly set out to examine associations between different potential welfare measurements (e.g. Mason et al. 2001; Nicol et al. 2009; Colson et al. 2019), our present study was highly novel in encompassing three dominant approaches: preference assessments, judgement bias tests and other candidate welfare indicator measurements. We conducted a triangulation study to investigate relationships between these in an investigation of domestic chickens, focusing on two types of preference measure (general and individual) and measuring welfare across short (6-week) and long (24-week) durations. The product of this type of investigation might be expected, at its simplest, to be straightforward, with the different candidate welfare measures aligning with both types of preference-based welfare classification and also correlating with one another. But on the basis of findings from a number of previous studies, in which apparent inconsistencies have been found, we anticipated that these relationships may in fact be more complex (e.g. Ross et al., 2019; Nicol, et al. 2009; Nicol, Caplen, Edgar et al., 2011). Our findings here confirmed this second possibility.

*<H2>Associations Between Preference and Judgement Bias*

In judgement bias studies, responding to ambiguous probe cues as if they are reward-predicting S+ cues can be viewed as a more positive, optimistic-like (or risky) pattern of choice, often associated with positive welfare and/or rewarding affect manipulations (Lagisz 2020; Mendl et al. 2009; Mendl and Paul 2020; Roelofs et al., 2016, Nguyen et al., 2020). Although it is not yet clear what mechanisms underlie judgement biases in nonhuman animals, it is hypothesized that affective states are key. Specifically, the task is thought to tap background affective states as information, guiding decisions about the likely reward/punishment values of ambiguous (probe cue) stimuli, for which no other data are available (see Appendix 1 for further details about the judgement bias task; see also Harding et al. 2004; Paul et al. 2005; Mendl et al. 2010; Mendl et al. 2020). In other words, judgement biases are hypothesized to be indicators of current affective state. In turn, current affective states are thought to function as integrative, summary indicators of an animal’s ongoing welfare or wellbeing, combining information regarding immediate and present threats and opportunities with past experiences of reward and punishment (Mendl et al. 2020; Paul et al. 2011).

Here, the proportion of ambiguous probe cues pecked by hens was used as our key measure of judgement bias, with a significant difference detected between IP and INP hens, at Phase 2 testing (see Fig. 6b). We were surprised to find no significant difference between GP and GNP hens, although it is possible that a larger sample size of judgement bias-trained hens would have also revealed a group difference at the Phase 2 testing point (see Fig. 5b). The difference in findings between the two types of preference grouping corresponds with the possibility that individual preference measures may be more likely to be aligned with the actual affective states of animals than general preference ones. These findings also fit with the idea that individual preference is particularly likely to be associated with judgement biases, as both are associated with ‘subjective’ (affective) conceptions of welfare and both are revealed by tests of decision making and choice (Mendl et al. 2020). However, we must be cautious in our conclusions here. Following Benjamini corrections for multiple testing, no significant differences between the groups remained. A key aim of the present study was to investigate whether judgement bias can be validated as an indicator of affect and welfare in domestic chickens, using the two types of living condition preference described above (GP and IP), and our present findings do not provide this clear validation. The differences noted above are tentative; all they can do for now is provide grounds for future hypothesis testing, and point to the need for larger scale studies with greater statistical power.

Our finding that the association between individual preference and judgement bias was only significant after relatively short-term exposure to the different living conditions (Phase 2) was also a tentative one, but nevertheless intriguing. In humans, prolonged exposure to both positive and negative events or environments can result in ‘hedonic adaptation’ (e.g. see Gilbert 2009, Kettlewell et al. 2020). Essentially, this is the observation that even after dramatic and affectively powerful occurrences, like lottery wins or becoming severely disabled, many people return eventually to near pre-event affective norms. Hedonic adaptation has not been explicitly investigated in animals. However, there is evidence from a number of studies that changes in the quality of animals’ environments can influence their judgement biases. For example, in a study of spatial judgement biases, rats, *Rattus norvegicus*, that had experienced cage enrichment removal exhibited ‘pessimistic’ biases, while Brydges (2010) found that rats that had experienced a living condition shift from unenriched to enriched showed ‘optimistic’ biases. In birds (starlings, *Sturnus vulgaris*), Bateson and Matheson (2007) found that a move from enriched to standard (less enriched) cages was associated with more ‘pessimistic’ biases, when compared with birds that had stayed in standard cages throughout. We conclude that our current findings offer a novel suggestion of hedonic adaptation in hens, indicating the possibility of considerable evolutionary continuity in this affective phenomenon.

*<H2>Associations Between Preference and Other Candidate Welfare Indicators*

Some of the differences observed between birds in the generally preferred and nonpreferred living conditions are likely to reflect direct influences of structural differences between the environments, as well as the additional (food) rewards presented to GP hens. These birds were heavier at Phases 2 and 3, even though both groups had access to adlibitum commercial layer food (and despite controlling for baseline weights at Phase 1). They also had shorter claws in Phases 2 and 3, a potentially important welfare benefit (i.e. physical comfort), probably resulting from the litter-based floor substrate (as opposed to bare wire). This flooring is also likely to have been important in allowing significantly higher levels of ground foraging (also in Phases 2 and 3), a strongly motivated set of behaviours of ancestral importance (incorporating ground scratching and ground pecking; see ethogram in Appendix 3, Table A1.). This finding links with our previous work in which longer durations of highly motivated behaviours such as preening (Nicol et al., 2009) and foraging (Nicol, Caplen, Statham et al., 2011) were associated with preference. It perhaps also indicates a more settled behavioural profile in GP birds (see also increased levels of sitting behaviour in this group in Phase 3 observations, as also found in Nicol, Caplen, Statham et al., 2011). Lower physiological arousal levels in GP birds in Phase 3 were also indicated by a lower pulse rate during handling (for blood pressure testing), despite the birds in both experimental groups being extremely well habituated to people. Blood pressure was a novel measure for use in chickens; while blood pressure itself did not differ between the preference groupings, nonsignificant trends were observed in the expected direction (i.e. lower pressures in more preferred environments), pointing to the possibility that such measures may prove useful in future research.

Faecal water was significantly lower in GP birds than GNP birds at Phases 2 and 3. This is relevant, as water metabolism is strongly related to stress. For example, chicks treated with ACTH have been found to increase their water intake and excreta output and show a 187% increase in faecal water output as a result (Puvadolpirod and Thaxton, 2000). In addition, this result replicates our previous findings linking low faecal water content in adult chickens with experience of preferred environments (Nicol et al., 2009; Nicol, Caplen, Statham et al., 2011).

It seems likely that some of the candidate welfare indicator measures identified here as being significantly related to preference had causal relationships with other measures. For example, the relatively higher levels of foraging seen in GP birds, facilitated by the wood-shavings floor substrate, may well have contributed to increased tibia strength (higher bone mineral density and stiffness), shorter claws and reduced claw length asymmetry. It may also have been associated with the substantially lower alkaline phosphatase (ALP) levels seen in these birds at Phase 3 (due to practical limitations, ALP was not measured in Phase 2). Increased plasma ALP activity has previously been linked with osteoblastic activity (Harr, 2002).

Some of the findings of relationships between preference groupings and potential welfare indicators may be less easy to explain. For example, the shift between phases in the direction of an association between preference group and triglyceride levels may reflect complex effects of stress and diet on lipid metabolism that could require future elucidation, and GP birds were more likely to have small combs than Group GNP birds by the end of the study, possibly an effect of lower social competition for resources in the GP environment (if comb size is an indicator of social rank and/or competitive status).

Given that the hens that showed individual preferences for their experimental living conditions (IP birds) were drawn primarily (but not exclusively) from GP birds, it is not surprising that shared significant associations with potential welfare indicators were found. Other differences between Groups GP and GNP, however, were not mirrored in Groups IP and INP, and one unique association was identified: IP birds had lower blood glucose levels than INP birds during Phase 2. These groups also differed in their judgement biases in Phase 2 (see Fig. 6b). Blood glucose levels in the plasma can be greatly increased by the actions of glucocorticoid stress hormones (cortisol, corticosterone), and we have previously found that the downstream integrated measure of blood glucose has a stronger relationship with environmental preference than more variable and less stable measures of glucocorticoids (which are subject to high variation due to pulsatile release patterns, circadian rhythms and other influencing factors; Nicol et al., 2009). The access of GP birds to additional food rewards had the potential to obscure any relationship between stress and lower blood glucose levels, which may explain why this relationship became apparent only in IP birds (where the influence of food rewards was somewhat diluted). However, a possible alternative explanation is that the IP birds coped better than INP birds in their respective environments, resulting in lower levels of HPA activation between Phases 1 and 2 of the study.

In sum, a number of anticipated associations were found between experimental living conditions (GP, GNP) and candidate welfare indicators, and these were roughly mirrored in comparisons of birds that did and did not show individual preferences for these conditions (IP, INP). A number of the relationships that were found between preference and potential welfare indicators were detected at both short-term (Phase 1, after 6 weeks) and long-term (Phase 3, after 24 weeks) intervals, indicating that, unlike judgement bias, these behavioural, physiological and physical measures were not potentially susceptible to hedonic adaptation (mass, claw length, ground foraging, faecal water content). We examined a wide range of potential welfare indicators, including a number that we had not assessed in previous studies (e.g. blood pressure, a novel measure in chickens; physical challenge tests; etc.) so it was perhaps surprising that not more differences were found between preference groups. One reason may be that the generally nonpreferred living conditions were not as aversive as in some previous studies. In Nicol et al. (2009), for example, hens were housed for up to 35 days in wire-floored pens that, unlike the current experiment, did not contain even rudimentary perches or a nestbox, and therefore may have had more negative affective consequences (indicated in measures such as H/L ratio elevation and head-shaking behaviour).

*<H2>Associations Between Judgement Bias and Candidate Welfare Indicators*

We anticipated that the hens’ judgement bias summary measure (proportion of ambiguous probe cues pecked) would show some associations with the other candidate welfare indicators studied here. Phase 1, during which all the hens were being kept in the same, ‘intermediate’ living conditions, provided the best opportunity to investigate this without the additional influences of differing experimental living conditions (which it was not possible to statistically control for within our relatively small sample of judgement bias-trained hens). In principle, if judgement biases reflect a summary, integrative measure of the ongoing welfare of an animal, then worse living conditions (less rewarding and more punishing, such as that experienced by the GNP group) might be expected to be associated with a greater negative (more ‘pessimistic’) bias in responding to ambiguous cues (Mendl et al. 2010; Mendl et al. 2020; see discussion above of our findings here relating to preference groups). In addition, however, individual hens faring less well in uniform living conditions might also be expected to show more negatively biased, less ‘optimistic’ judgements regarding ambiguous cue stimuli. In particular, we hypothesized that social interactions involving aggression and/or feather pecking might be associated with judgement bias, with hens that were observed receiving aggressive attacks or severe feather pecks being most likely to show a reduction in optimistic-like responses (e.g. see Wichman et al. 2012). In fact, very little aggressive behaviour was seen across the entire study, and there were no links between judgement biases and social/aggressive behaviours, or indeed any of our other candidate welfare indicators, in Phase 1. A possible reason for this is that the intermediate living conditions of Phase 1 did not place many stressors on birds, and so, few socially deleterious behaviours or other welfare diminutions occurred. This may be in contrast to some contemporary farming systems (and perhaps some previous experimental studies), in which a variety of welfare hazards, including social conflicts between hens, may be much higher, with the resulting possibility of considerably greater variation in welfare, affective state and judgement biases.

In sum, no associations were found between hens’ judgement bias scores and other potential measures of welfare, but it is not possible to conclude whether this represents ‘evidence of absence’ or merely ‘absence of evidence’ (see Study Limitations, below). However, this lack of association also resonates with the recent finding of Ross et al. (2019), who found judgement bias to be unrelated to another potential welfare measure in hens (startle magnitude) and concluded that judgement bias may be controlled by different underlying mechanisms and predominant affective states within tests. While both of these (absence) findings may be explained by low statistical power, (i.e. the relatively small numbers of judgement bias-trained birds), the possibility remains that the judgement bias test, by its nature a cognitive and integrative measure, may reflect a different component or facet of welfare than many of the more traditional welfare measures used with hens, as well as the new avian startle measure of affect developed by Ross et al. (2019). It is also possible that, given the absence of findings relating to judgement bias in some other studies of chickens (Hernandez et al. 2015; Wichman et al. 2012), there are certain problems associated with measuring or validating the task with this particular species. This would be surprising, given that we have argued that a key advantage of the judgement bias paradigm is its translatability to a wide range of species (Mendl & Paul 2020). But although this possibility could cast doubt on the usefulness of judgement bias for future studies of chickens, it could also raise some potentially interesting questions about the features of affective states that the task measures, and the features or facets of affective states that arise in this (and perhaps other related) species.

*<H2>Study Limitations*

We have already noted that the generally nonpreferred environment employed here may not have been as aversive as those used in some other studies of potential welfare indicators under conditions of greater/lesser enrichment. Although a key finding was that subject hens differed significantly in their individual preferences for the two environments (i.e. confirming the living condition manipulation was detectable for these hens, and that generally nonpreferred living conditions were individually preferred significantly less than generally preferred ones), the relative aversiveness of the nonpreferred living conditions may have simply not been great enough for some potentially valid welfare indicators to be revealed. In addition, the relatively high levels of human handling these hens experienced (as a result of general management of the study and training for judgement bias and other tests) meant that one particular and well-established welfare indicator, the tonic immobility test (a measure of variation in fearfulness among chickens; e.g. see Forkman et al. 2007), could not be used, as none of the hens showed the immobility response. The present study may have failed, therefore, to detect (and provide evidence of the interrelationships between) some potentially useful welfare indicators. This highlights an ongoing scientific and ethical challenge for animal welfare research: how, whether and to what degree animals’ welfare should be disrupted in the search for candidate welfare measures and potential welfare solutions.

The statistical analyses performed here made use of multilevel models to control for individual hens’ baseline (Phase 1) performance and to adjust for the nonindependence of hens living within the same pens (a common problem in animal science research; Knowles & Green 2002; Iason & Elston, 2002). However, the nature of the triangulation approach meant that a large number of tests and comparisons needed to be made. Multiple testing gives rise to a heightened risk of Type 1 errors (indicating an association where none exists), a problem that has a number of potential statistical solutions (e.g. Benjamini & Hochberg 1995; Benjamini & Liu 1999). However, these in turn can increase Type 2 errors (omission of actual associations), and also make comparisons with earlier studies difficult (where no adjustment may have occurred). We have addressed this issue, and our response to it, in detail in the Analyses section above. We acknowledge the risk of Type 1 errors having occurred within these analyses and as a consequence take a cautious view of the results found. Future replications will be required to confirm or disconfirm some of the findings reported here.

An additional limitation was set by the practical constraints of training and testing judgement biases on multiple chickens. Because it was only possible for us to train 24 (of a total of 60) hens, statistical power was lower for this part of the triangulation analyses. While these sample sizes were similar to those used in previous judgement bias studies (e.g. Deakin et al., 2016; de Haas et al., 2017; Hernandez et al. 2015), and some findings emerged in the links between preference and judgement bias test results (see above), analyses of associations between judgement bias (proportion of ambiguous probes pecked) and other potential welfare indicator measures have been hampered by these relatively small sample sizes.

*<H2>Conclusions*

Previous attempts to formally examine the relationships between different approaches to welfare assessment have been limited (e.g. see Mason et al., 2001; Nicol et al., 2009; Colson et al., 2019); our current study is highly novel in triangulating preference assessment (both general and individual), judgement bias tests, and other candidate welfare indicator measurements. Our proposition was that we would find convergent evidence of differing welfare states in domestic chickens kept in living conditions that previous experiments have shown to be generally preferred or nonpreferred by most hens (GP and GNP), but that we may also discover complexity in these associations. We have confirmed that, among domestic chickens, animal welfare is a construct well defined by preference and a relatively small set of robustly corresponding ‘welfare indicator’ measures. Moreover, this is likely to depend on the degree of challenge faced by the animals, with more of these potential indicators becoming apparent if living conditions decline further (Nicol et al., 2009).

We found that some welfare indicators were associated with both general and individual preferences, while others were linked with individual preferences only (or are more clearly linked with these). As in many other studies (e.g. Browne et al., 2010; Maia et al., 2017), a minority of subjects expressed living condition preferences that diverged from the majority. Minority preferences can sometimes be explained as artefacts (error, continued monitoring or information seeking; Nicol, 1986) but may also reflect a genuine preference away from the norm. There could be a number of reasons for this, including the possibility that generally improved (e.g. larger, enriched) environments can increase social stratification, leading to greater differentiation between competitively successful and unsuccessful individuals (and also between individuals of differing ‘personality types’; Asher et al., 2009). The difference between general and individual preference classifications in the present study was particularly notable in judgement bias, which was primarily linked (during Phase 2) with individual preferences, rather than general ones. We suggest that future studies using judgement biases in animals might benefit from considering individual as well as general preference comparisons. This may help to reduce heterogeneity in judgement bias findings (see Neville et al. 2020), although we also acknowledge that the practical demands associated with measurement of individual preference can be considerable.

We also found that the timescale of exposure to these different living conditions was important. While some of our candidate welfare indicators were affected similarly by short-term and long-term exposure to preferred and nonpreferred living conditions (e.g. body mass, claw length, faecal water), others were restricted to either shorter- or longer-term effects. Judgement biases, for example, showed most group differences after short-term (6-week) exposure to the experimental living conditions, pointing to the possibility that affective tasks such as this may be most likely to be affected by environmental changes (i.e. improvements or declines that are registered by animals as increases or decreases of preference); a possible example of hedonic adaptation in birds.

**Author Contributions**

**Elizabeth S. Paul**: Conceptualization; Data curation; Formal analysis; Investigation; Methodology; Project administration; Supervision; Validation; Writing – original draft; Writing – review & editing. **William Browne**: Formal analysis; Methodology, Software; Writing -review and editing. **Michael T. Mendl**: Conceptualization; Funding acquisition; Methodology; Writing – review and editing. **Gina Caplen**: Investigation; Project administration; Writing – review and editing. **Anna Trevarthen**: Investigation; Project administration; Writing – review and editing. **Suzanne Held**: Conceptualization; Methodology; Writing–review and editing. **Christine J. Nicol**: Conceptualization; Funding acquisition; Methodology; Project administration; Resources; Supervision; Writing – original draft; Writing – review & editing.

**Acknowledgments**

The work was funded by BBSRC grant [BB/K00042X/1](https://bbsrc.ukri.org/research/grants-search/AwardDetails.aspx?FundingReference=BB/K00042X/1). We thank Fran Booth for bone biomechanics testing, and an anonymous referee for their very detailed and helpful comments. We also thank Georgia Mason in particular for her extensive, insightful and invaluable comments on two early drafts of the manuscript. We believe that the paper is much improved as a result of her input and are very grateful indeed for her thoughtfulness and rigour.

**References**

Ahloy-Dallaire, J., Espinosa, J., & Mason, G. (2018). Play and optimal welfare: Does play indicate the presence of positive affective states. *Behavioural Processes, 156*, 3-15. <https://doi.org/10.1016/j.beproc.2017.11.011>

Alm, M., Tauson, R., Holm, L., Wichman, A., Kalliokoski, O., & Wall, H. (2016). Welfare indicators in laying hens in relation to nest exclusion. *Poultry Science, 95*, 1238-1247. <https://doi.org/10.3382/ps/pew100>

Amsel, A. (1958). The role of frustrative nonreward in noncontinuous reward situations. *Psychological Bulletin*, *55***,** 102-119. <https://doi.org/10.1037/h0043125>

Anderson, D.J., & Adolphs, R. (2014). A framework for studying emotions across species. *Cell,* *157***,** 187-200. <https://doi.org/10.1016/j.cell.2014.03.003>

Appleby, M.C., & McRae, H.E. (1986). The individual nest box as a super-stimulus for domestic hens. *Applied Animal Behaviour Science, 15*(2), 169-176. [https://doi:10.1016/0168-1591(86)90062-6](https://doi:10.1016/0168-1591%2886%2990062-6)

Appleby, M.C., & Hughes, B.O. (1995). The Edinburgh modified cage for laying hens. *British Poultry Science, 36*(5), 707-71. <https://doi:10.1080/00071669508417815>.

Bar-Haim, Y., Lamy, D., Pergamin, L., Bakermans-Kranenburg, M.J., & van Ijzendoorn, M.H. (2007). Threat-related attentional bias in anxious and nonanxious individuals: A meta-analytic study. *Psychological Bulletin, 133*(1), 1-24. <https://doi:10.1037/0033-2909.133.1.1>.

Barnard, C.J. (2007). Ethical regulation and animal science: why animal behaviour is special. *Animal Behaviour, 74,* 5-13. <https://doi.org/10.1016/j.anbehav.2007.04.002>.

Barnard, C.J., & Hurst, J.L. (1996). Welfare by design: The natural selection of welfare criteria. *Animal Alternatives, Welfare, and Ethics. Developments in Animal and Veterinary Sciences, 27,* 209-214.

Bateson, M., & Matheson, S.M. (2007) Performance on a categorization task suggests that removal of enrichment induces ‘pessimism’ in captive European starlings (*Sturnus vulgaris*). *Animal Welfare, 16(Supplement 1),* 33-36.

Bateson, M., Desire, S., Gartside, S.E., & Wright, G.A. (2011). Agitated honey bees exhibit cognitive biases. *Current Biology, 21*(12), 1070-1073. <https://doi.org/10.1016/j.cub.2011.05.017>

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate – a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B: Statistical Methodology, 57(*1*),* 289-300. <https://doi.org/10.1111/j.2517-6161.1995.tb02031.x>

Benjamini, Y., & Liu, W. (1999). A step-down multiple hypotheses testing procedure that controls the false discovery rate under independence. *Journal of Statistical Planning and Inference, 82(*1-2*),* 163-170. [https://doi.org/10.1016/S0378-3758(99)00040-3](https://doi.org/10.1016/S0378-3758%2899%2900040-3)

Blanchflower, D.G. & Oswald, A.J. (2008). Hypertension and happiness across nations. *Journal of Health Economics, 27*, 218-233. Doi: [10.1016/j.jhealeco.2007.06.002](https://doi.org/10.1016/j.jhealeco.2007.06.002)

Bouvarel, I., Chagneau, A. M., Lecuelle, S., Lescoat, P., Ferreira, G., Duvaux-Ponter, C., & Leterrier, C. (2009). Feed composition and hardness interact in preference and intake in chickens. *Applied Animal Behaviour Science, 118*(1-2), 62-68. <https://doi.org/10.1016/j.applanim.2009.01.008>

Brambell, F.W.R. (1965). *Report of the technical committee to enquire into the welfare of animals kept under intensive livestock husbandry systems*. Her Majesty’s Stationery Office.

Bright, A., Brass, D., Clachan, J., Drake, K. A., & Joret, A. D. (2011). Canopy cover is correlated with reduced injurious feather pecking in commercial flocks of free-range laying hens. *Animal Welfare, 20*(3), 329-338.

Brilot, B.O., Asher, L., & Bateson, M., (2010). Stereotyping starlings are more ‘pessimistic’. *Animal Cognition, 13* (5), 721-731. <https://doi:10.1007/s10071-010-0323-z>

Broom, D.M. (1986). Indicators of poor welfare. *British Veterinary Journal*, *142*, 524–526. [http://dx.doi.org/10.1016/0007-1935(86)90109-0](http://dx.doi.org/10.1016/0007-1935%2886%2990109-0)

Browne, W.J., Caplen, G., Edgar, J., Wilson L.R., & Nicol, C.J. (2010). Consistency, transitivity and inter-relationships between measures of choice in environmental preference tests with chickens. *Behavioural Processes, 83,* 72-78. <https://doi.org/10.1016/j.beproc.2009.10.004>

Browning, H. (2020). The natural behaviour debate: Two conceptions of animal welfare. *Journal of Applied Animal Welfare Science, 23*(3)*,* 325-337. <https://doi.org/10.1080/10888705.2019.1672552>

Bruce, E.H., Prescott, N.B., & Wathes, C.M. (2003). Preferred food rewards for laying hens in behavioural experiments. *British Poultry Science, 44*(3), 345-349. <https://doi:10.1080/0007166031000085490>

Brydges, N.M., Leach, M., Nicol, K., Wright, R., & Bateson, M. (2011). Environmental enrichment induces optimistic cognitive bias in rats. *Animal Behaviour, 81,* 169-175. <https://doi.org/10.1016/j.anbehav.2010.09.030>

Caillaud, C., Doumergue, M., Preau, M., Haas, V. & Kalampalikis, N. (2019). The past and present of triangulation and social representations theory: a crossed history. *Qualitative Research in Psychology, 16*, 375-391. <https://doi.org/10.1080/14780887.2019.1605272>

Chaves, O.M., Fernandes, F.A., Oliveira, G.T., & Bicca-Marques, J.C. (2019). Assessing the influence of biotic, abiotic and social factors on the physiological stress of a large Neotropical primate in Atlantic forest fragments. *Science of the Total Environment, 690*, 705-716. <https://doi.org/10.1016/j.scitotenv.2019.07.033>

Chen, D.H., Bao, J., Meng, F.Y., & Wei, C.B. (2014). Choice of perch characteristics by laying hens in cages with different group size and perching behaviours. *Applied Animal Behaviour Science, 150*, 37-43. <https://doi.org/10.1016/j.applanim.2013.10.007>

Colson, V., Mure, A., Vaoltaire, C., Le Calvez, J.M., Goardon, L., Labbe, l., Leguen, I., & Prunet, P. (2019). A novel emotional and cognitive approach to welfare phenotyping in rainbow trout exposed to poor water quality. *Applied Animal Behaviour Science, 210*, 103-112. <https://doi.org/10.1016/j.applanim.2018.10.010>

Cooper, J.J., & Appleby, M. C. (1996). Demand for nest boxes in laying hens. *Behavioural Processes, 36*(2), 171-182. [https://doi.org/10.1016/0376-6357(95)00027-5](https://doi.org/10.1016/0376-6357%2895%2900027-5)

Cooper, J.J., & Appleby, M. C. (2003). The value of environmental resources to domestic hens: A comparison of the work-rate for food and for nests as a function of time. *Animal Welfare, 12*(1), 39-52.

Davis, N.J., Prescott, N. B., Savory, C. J., & Wathes, C. M. (1999). Preferences of growing fowls for different light intensities in relation to age, strain and behaviour. *Animal Welfare, 8*(3), 193-203.

Dawkins, M.S., & Beardsley, T. (1986). Reinforcing properties of access to litter in hens. *Applied Animal Behaviour Science, 15*(4), 351-364. [https://doi.org/10.1016/0168-1591(86)90127-9](https://doi.org/10.1016/0168-1591%2886%2990127-9)

Dawkins, M.S. (1990). From an animal’s point of view: Motivation, fitness and animal welfare. *Behavioral and Brain Sciences, 13*, 1-9. <https://doi.org/10.1017/S0140525X00077104>

Dawkins, M.S., Donnelly, C.A., & Jones, T.A. (2004). Chicken welfare is influenced more by housing conditions than by stocking density. *Nature, 427,* 342-344. <https://doi.org/10.1038/nature02226>

Dawkins, M.S. (2006). A user’s guide to animal welfare science. *Trends in Ecology and Evolution, 21*(2), 77-82. <https://doi.org/10.1016/j.tree.2005.10.017>

Dawkins, M. (2015). Animal welfare and the paradox of animal consciousness. *Advances in the Study of Behavior, 47*, 5-38. <https://doi.org/10.1016/bs.asb.2014.11.001>

Deakin, A., Browne, W.J., Hodge, J.J.L., Paul, E.S., & Mendl, M. (2016). A screen-peck task for investigating cognitive bias in laying hens. *PLoS ONE 11*(7), Article e0158222. <https://doi.org/10.1371/journal.pone.0158222>

de Haas, E.N., Lee, C., & Rodenburg, T.B. (2017). Learning and judgement can be affected by predisposed fearfulness in laying hens. *Frontiers in Veterinary Science, 4*, Article e113. <https://doi.org/10.3389/fvets.2017.00113>

de Waal, F.B.M. (2011). What is an animal emotion? *Annals of the New York Academy of Sciences, 1224*(1), 91-206*.* [*https://*doi.org/10.1111/j.1749-6632.2010.05912.x](https://doi.org/10.1111/j.1749-6632.2010.05912.x)

Diez-Leon, M.,Bursian, S., Galicia, D., Napolitano, A., Palme, R., & Mason, G. (2016). Environmentally enriching American mink (Neovision vision) increases lymphoid organ weight and skeletal symmetry, and reveals differences between two sub-types of stereotypic behavior. *Applied Animal Behaviour Science, 177*, 59-69. <https://doi.org/10.1016/j.applanim.2015.12.002>

Duncan, I.J.H. (1993). Welfare is to do with what animals feel. *Journal of Agricultural and Environmental Ethics*, *6(*Suppl. 2*),* 8–14.

Duncan, I.J.H. (1996). Animal welfare defined in terms of feelings. *Acta Agriculturae Scandinavica, Section A. Animal Science Supplement, 27,* 29-35.

Duncan, I.J.H., & Fraser, D. (1997). Understanding animal welfare. In Appleby, M.C. and Hughes, B.O. (Eds) *Animal welfare*, pp19-31. CAB International.

Eusemann, B.K., Baulain, U., Schrader, L., Thone-Reineke, C., Patt, A., & Petwo, S. (2020). Radiographic examination of keel bone damage in living hens of different strains kept in two housing systems. *PLoS One, 13*(5), Article e0194974. <https://doi.org/10.1371/journal.pone.0194974>

FAO (2019). FAOSTAT Live Animals, 2019. Accessed May 2021. http://www.fao.org/faostat/en/#data/QA/visualize

Fawcett, T., Fallenstein, B., Higginson, A.D., Houston, A.I., Mallpress, D.E.W, Trimmer, P.C., & McNamara, J.M. (2014) The evolution of decision rules in complex environments. *Trends in Cognitive Science, 18,* 153-161. <https://doi.org/10.1016/j.tics.2013.12.012>

Ferrante, V., Lolli, S., Ferrari, L., Watanabe, T.T.N., Tremolada, C., Marchewka, J., & Estevez, I. (2019). Differences in prevalence of welfare indicators in male and female turkey flocks (*Meleagris gallopavo*). *Poultry Science, 98,* 1568-1574 <https://doi.org/10.3382/ps/pey534>

Forkman, B.,Boissy, A., Meunier-Salauen, M.C., Canali, E., & Jones, R.B. (2007). A critical review of fear tests used on cattle, pigs, sheep, poultry and horses. *Physiology and Behavior, 92(*3)*,* 340-374. <https://doi.org/10.1016/j.physbeh.2007.03.016>

Fraser, D., & Nicol, C.J. (2018). Preference and motivation research. In Appleby, M.C., Olsson, I.A.S., Galindo, F *Animal welfare* (3rd edn, pp. 213-231). Doi:[[10.1079/9781786390202.0213](https://doi.org/10.1079/9781786390202.0212)](http://dx.doi.org/10.1079/9781786390202.0213)

Fraser, D., Weary, D.M., Pajor, E.A., & Milligan, B.N. (1997). A scientific conception of animal welfare that reflects ethical concerns. *Animal Welfare, 6*, 187-205.

Fraser, D., & Duncan, I.J.H., (1998). Pleasures, pains and animal welfare: Toward a natural history of affect. *Animal Welfare, 7,* 383-396.

Freire, R., Appleby, M.C., & Hughes, B.O. (1997). Assessment of pre-laying motivation in the domestic hen using social interaction. *Animal Behaviour, 54*, 313-319. <https://doi.org/10.1006/anbe.1996.0463>

Gentle, M.J. (1972). Taste preference in chicken (*Gallus domesticus*). *British Poultry Science 13*(2), 141-155. <https://doi.org/10.1080/00071667208415928>

Gilbert, D. (2009). *Stumbling on happiness.* Vintage.

Gregory, N.G. & Robins J.K. (1998). A body condition scoring system for laying hens. *New Zealand Journal of Agricultural Research, 41,* 555-559. <https://doi.org/10.1080/00288233.1998.9513338>

Gross, W.B., & Siegel, H.S. (1983). Evaluation of the heterophil-lymphocyte ratio as a measure of stress in chickens. *Avian Diseases, 27*, 972-979. <https://doi.org/10.2307/1590198>

Hager, C., Keubler, L.M., Talbot, S.R., Biernot, S., Weegh, N., Buchheister, S., Buettne, M., Glage, S., & Bleich, A. (2018). Running in the wheel: Defining individual severity levels in mice. *PLOS Biology, 16*, Article e2006159. <https://doi.org/10.1371/journal.pbio.2006159>

Harding, E.J., Paul, E.S., & Mendl, M. (2004). Cognitive bias and affective state. *Nature, 427*, 312-312. <https://doi.org/10.1038/427312a>

Harr, K.E. (2002). Clinical chemistry of companion avian species: A review. *Veterinary Clinical Pathology, 31(*3), 140-151. <https://doi.org/10.1111/j.1939-165X.2002.tb00295.x>

Harrison, R. (1964). *Animal machines*. Vincent Stuart.

Hernandez, C.E., Hinch, G., Lea, J., Ferguson, D., & Lee, C. (2015). Acute stress enhances sensitivity to a highly attractive food reward without affecting judgement bias in laying hens. *Applied Animal Behaviour Science, 163*, 135-143. <https://doi.org/10.1016/j.applanim.2014.12.002>

Hughes, B.O., & Black, A.J. (1976). Preference of domestic hens for different types of battery cage floor. *British Poultry Science, 14,* 615-618. <https://doi.org/10.1080/00071667308416071>

Hughes, B.O. (1993). Choice between artificial turf and wire floor as nest sites in individually caged laying hens. *Applied Animal Behaviour Science, 36*(4), 327-335. [https://doi.org/10.1016/0168-1591(93)90130-h](https://doi.org/10.1016/0168-1591%2893%2990130-h)

Iannetti, L., Neri, D., Santarelli, G.A., Cotturone, G., Vulpiani, M.P., Salini, R., Antoci, S., Di Serafino, G., Di Giannatale, E., Pomilio, F., & Messori, S. (2020). Animal welfare and microbiological safety of poultry meat: Impact of different at-farm animal welfare levels on at-slaughterhouse Campylobacter and Salmonella contamination. *Food Control, 109,* Article 106921. <https://doi.org/10.1016/j.foodcont.2019.106921>

Iason, G.R., & Elston, D.A. (2002). Groups, individuals, efficiency and validity of statistical analyses. *Applied Animal Behaviour Science, 75*(3)*,* 261-265. [https://doi.org/10.1016/S0168-1591(01)00130-7](https://doi.org/10.1016/S0168-1591%2801%2900130-7)

Iyasere, O.S., Beard, A.P., Guy, J.H., & Bateson, M. (2017). Elevated levels of the stress hormone, corticosterone, cause ‘pessimistic’ judgement bias in broiler chickens. *Scientific Reports, 7.* <https://doi.org/10.1038/s41598-017-07040-y>

Jones, A R., Bizo, L.A., & Foster, T.M. (2012). Domestic hen chicks' conditioned place preferences for sound. *Behavioural Processes, 89*(1), 30-35. <https://doi.org/10.1016/j.beproc.2011.10.007>

Jones, R.B. (1986a). Background auditory-stimulation and tonic immobility in the domestic-fowl. *Ircs Medical Science-Biochemistry, 14*(4), 337-338.

Jones, R.B. (1986b). Conspecific vocalization, tonic immobility and fearfulness in the domestic-fowl. *Behavioural Processes, 13*(3), 217-225. [https://doi.org/10.1016/0376-6357(86)90085-9](https://doi.org/10.1016/0376-6357%2886%2990085-9)

Jones, R.B., & Carmichael, N.L. (1998). Pecking at string by individually caged, adult laying hens: colour preferences and their stability. *Applied Animal Behaviour Science, 60*(1), 11-23. [https://doi.org/10.1016/s0168-1591(98)00151-8](https://doi.org/10.1016/s0168-1591%2898%2900151-8)

Kettlewell N., Morris R.W., Hi, N., Cobb-Clark, D.A., Cripps, S., & Glozier, N. (2020). The differential impact of major life events on cognitive and affective well-being. *SSM – Population Health, 10,* Article 100533. <https://doi.org/10.1016/j.ssmph.2019.100533>

Kiley-Worthington, M. (1989). Ecological, ethological, and ethically sound environments for animals: Toward symbiosis. *Journal of Agricultural Ethics*, *2*, 323–347. <http://doi.org/10.1007/BF01826810>

Knott, L., Whitehead, C.C., Fleming R.H., & Bailey A.J. (1995). Biochemical changes in the collagenous matrix of osteoporotic avian bone. *Biochemical Journal, 310,* 1045–1051. <https://doi.org/10.1042/bj3101045>

Knowles, T.G., & Green, L.E. (2002). Multilevel statistical models allow simultaneous consideration of both individual and group effects. *Applied Animal Behaviour Science, 77(*4)*,* 335-336. [https://doi.org/10.1016/S0168-1591(02)00065-5](https://doi.org/10.1016/S0168-1591%2802%2900065-5)

Koch, M. (1999). The neurobiology of startle. *Progress in Neurobiology, 59*, 107–128. [https://doi.org/10.1016/s0301-0082(98)00098-7](https://doi.org/10.1016/s0301-0082%2898%2900098-7)

Kruschwitz, A., Zupan, M., Buchwalder, T., & Huber-Eicher, B. (2008). Nest preference of laying hens (*Gallus gallus domesticus*) and their motivation to exert themselves to gain nest access. *Applied Animal Behaviour Science, 112*(3-4), 321-330. <https://doi:10.1016/j.applanim.2007.08.005>

Lagisz, M., Zidar, J., Nakagawa, S., Neville, V., Sorato, E., Paul, E.S.,Bateson, M., Mendl, M., & L∅vlie, H. (2020). Optimism, pessimism and judgement bias in animals: A systematic review and meta-analysis. *Neuroscience and Biobehavioral Reviews, 108,* 269-286. <https://doi.org/10.1016/j.neubiorev.2020.07.012>

LeDoux, J. (2012). Rethinking the Emotional Brain. *Neuron, 73****,*** 653-676. <https://doi.org/10.1016/j.neuron.2012.02.004>

Littin, K., Acevedo, A., Browne, W., Edgar, J., Mendl, M., Owen, D., Sherwin, C., Würbel, H., & Nicol, C. (2008). Towards humane end points: behavioural changes precede clinical signs of disease in a Huntington’s disease model*. Proceedings of the Royal Society B – Biological Sciences, 275,* 1865-1874. <https://doi.org/10.1098/rspb.2008.0388>

Mackenzie, J.G., Foster, T.M., & Temple, W. (1993). Sound avoidance by hens. *Behavioural Processes, 30*(2), 143-156. [https://doi.org/10.1016/0376-6357(93)90004-b](https://doi.org/10.1016/0376-6357%2893%2990004-b)

MacLeod, A.K., & Byrne, A. (1996). Anxiety, depression, and the anticipation of future positive and negative experiences. *Journal of Abnormal Psychology, 105*(2), 286-9. <https://doi.org/10.1037//0021-843x.105.2.286>

Maia, C.M., Ferguson, B., Volpato, G.L., & Braithwaite, V.A. (2017). Physical and psychological motivation tests of individual preferences in rainbow trout. *Journal of Zoology, 302,* 108-118. <https://doi.org/10.1111/jzo.12438>

Marrocco, J., Verhaeghe, R., Bucci, D., Di Menna, L., Traficante, A., Bouwalerh, H., Van Camp, G., Ghiglieri, V., Picconi, B., Calabresi, P., Ravasi, L., Cisani, F., Bagheri, F., Pittaluga, A., Bruno, V., Battaglia, G., Morley-Fletcher, S., Nicoletti, F., & Maccari, S*.* (2020). Maternal stress programs accelerated aging of the basal ganglia motor system in offspring. *Neurobiology of Stress, 13,* Article 100265. <https://doi.org/10.1016/j.ynstr.2020.100265>

Mason, G., & Mendl, M. (1993). Why is there no simple way of measuring animal welfare? *Animal Welfare, 2*, 301-319.

Mason, G.J., Cooper, J.J., & Clarebrough, C. (2001). Frustrations of fur-farmed mink. *Nature, 410*, 35-36. <https://doi.org/10.1038/35065157>

Mason, G.J. & Veasey, J.S. (2010). How should the psychological well-being of zoo elephants be objectively investigated. *Zoo Biology, 29*, 237-255. <https://doi.org/10.1002/zoo.20256>

Matheson, S.M., Asher, L., & Bateson, M., (2008) Larger, enriched cages are associated with ‘optimistic’ response biases in captive European starlings (*Sturnus vulgaris*). *Applied Animal Behavior Science, 109*(2-4), 374-383. <https://doi:10.1016/j.applanim.2007.03.007>

McAdie, T.M., Foster, T.M., Temple, W., & Matthews, L.R. (1993). A method for measuring the aversiveness of sounds to domestic hens. *Applied Animal Behaviour Science, 37*(3), 223-238. [https://doi.org/10.1016/0168-1591(93)90113-4](https://doi.org/10.1016/0168-1591%2893%2990113-4)

McAdie, T. M., Keeling, L. J., Blokhuis, H. J., & Jones, R. B. (2005). Reduction in feather pecking and improvement of feather condition with the presentation of a string device to chickens. *Applied Animal Behaviour Science, 93*(1-2), 67-80. <https://doi.org/10.1016/j.applanim.2004.09.004>

McGlone, J. J. (1993). What is animal welfare? *Journal of Agricultural and Environmental Ethics*, *6*, 26–36.

Mendl, M., Burman, O.H.P., Parker, R.M.A., & Paul, E.S. (2009). Cognitive bias as an indicator of animal emotion and welfare: Emerging evidence and underlying mechanisms. *Applied Animal Behaviour Science, 118(3-4),* 161-81. <https://doi.org/10.1016/j.applanim.2009.02.023>

Mendl, M., Burman, O.H.P., & Paul, E.S. (2010). An integrative and functional framework for the study of animal emotion and mood. *Proceedings of the Royal Society B: Biological Sciences, 277,* 2895-2904. <https://doi.org/10.1098/rspb.2010.0303>

Mendl, M., Mason, G.J. & Paul, E.S. (2017). Animal welfare science. In J. Call, G. M. Burghardt, I. M. Pepperberg, C. T. Snowdon, & T. Zentall (Eds.), *APA handbook of comparative psychology: Perception, learning, and cognition* (pp. 793–811). American Psychological Association.

Mendl, M., & Paul, E.S. (2020). Animal affect and decision-making. *Neuroscience and Biobehavioral Reviews, 112,* 144-163. <https://doi.org/10.1016/j.neubiorev.2020.01.025>

Millensen, J. R. (1967). *Principles of behavioral analysis.* Macmillan.

Mitchell, M.A., Kettlewell, P.J., & Maxwell, M.H. (1992) Indicators of physiological stress in broiler chickens during road transportation. *Animal Welfare, 1*, 91-103.

Moe, R.O., Nordgreen, J., Janczak, A.M., Spruijt, B.M., & Bakken, M. (2013). Effects of signalled reward type, food status and a mu-opioid receptor antagonist on cue-induced anticipatory behaviour in laying hens (*Gallus domesticus*). *Applied Animal Behaviour Science, 148*(1-2), 46-53. <https://doi.org/10.1016/j.applanim.2013.08.001>

Mowrer, O.H. (1960). *Learning Theory and Behavior,* New York, John Wiley & Sons.

Nasr, M.A.F., Nicol, C.J., & Murrell, J.C. (2012). Do laying hens with keel bone fractures experience pain? *PLoS ONE, 7*, Article e42420 <https://doi.org/10.1371/journal.pone.0042420>

Nasr, M.A.F., Nicol, C.J., Wilkins, L.J., & Murrell, J.C. (2015). The effects of two non-steroidal anti-inflammatory drugs on the mobility of laying hens with keel bone fractures. *Veterinary Anaesthesia and Analgesia, 42,* 197-204. <https://doi.org/10.1111/vaa.12175>

Neville, V., Nakagawa, S., Zidar, J., Paul, E.S., Lagisz, M., Bateson, M., L∅vlie, H., & Mendl, M. (2020). Pharmacological manipulations of judgement bias: a systematic review and meta-analysis. *Neuroscience and Biobehavioral Reviews, 108,* 269-286. <https://doi.org/10.1016/j.neubiorev.2019.11.008>

Nicol, C.J. (1986). Nonexclusive spatial preference in the laying hen. *Applied Animal Behaviour Science, 15(4),* 337-350. [https://doi.org/10.1016/0168-1591(86)90126-7](https://doi.org/10.1016/0168-1591%2886%2990126-7)

Nicol, C.J., Caplen, G., Edgar, J., & Browne, W.J. (2009). Associations between welfare indicators and environmental choice in laying hens. *Animal Behaviour, 78,* 413-424. <https://doi.org/10.1016/j.anbehav.2009.05.016>

Nicol, C.J., Caplen, G., Statham, P., & Browne, W.J. (2011). Decisions about foraging and risk trade-offs in chickens are associated with individual somatic response profiles. *Animal Behaviour, 82,* 255-262. <https://doi.org/10.1016/j.anbehav.2011.04.022>

Nicol, C.J., Caplen, G., Edgar, J., Richards, G., & Browne, W.J. (2011). Relationships between multiple welfare indicators measured in individual chickens across different time periods and environments. *Animal Welfare, 20,* 133-143

Nguyen, H.A.T., Guo, C., & Homberg, J.R. (2020). Cognitive bias under adverse and rewarding conditions: A systematic review of rodent studies. *Frontiers in Behavioral Neuroscience, 14,* Article 14. <https://doi.org/10.3389/fnbeh.2020.00014>

Olsson, I. A. S., & Keeling, L. J. (2002). The push-door for measuring motivation in hens: Laying hens are motivated to perch at night. *Animal Welfare, 11*(1), 11-19.

Ordnance Survey (2021). <https://getoutside.ordnancesurvey.co.uk/guides/navigation-pinpointing-your-location/>

Ostlund, U., Kidd, L., Wengstrom, Y. & Rowa-Dewar, N. (2011). Combining qualitative and quantitative research within mixed method research designs: a methodological review. *International Journal of Nursing Studies, 48,* 369-383. <https://doi.org/10.1016/j.ijnurstu.2010.10.005>

Panksepp, J. (1998). *Affective neuroscience: The foundations of human and animal emotions*. Oxford University Press.

Paul, E. S., Harding, E. J., & Mendl, M. (2005). Measuring emotional processes in animals: the utility of a cognitive approach. *Neuroscience and Biobehavioral Reviews,* *29****,*** 469-491. <https://doi.org/10.1016/j.neubiorev.2005.01.002>

Paul, E.S.. & Mendl, M.T. (2018). Animal emotion: Descriptive and prescriptive definitions and their implications for a comparative perspective. *Applied Animal Behaviour Science, 205****,*** 202-209. <https://doi.org/10.1016/j.applanim.2018.01.008>

Paul, E.S., Sher, S., Tamietto, M, Winkelman, P., & Mendl, M.T. (2020). Towards a comparative science of emotion : Affect and consciousness in humans and animals. *Neuroscience and Biobehavioural Reviews, 108*, 749-770. <https://doi.org/10.1016/j.neubiorev.2019.11.014>

Perry, C.J., Baciadonna, L., & Chittka, L. (2016). Unexpected rewards induce dopamine-dependent positive emotion-like state changes in bumblebees. *Science, 353(*6307), 1529-1531. <https://doi.org/10.1126/science.aaf4454>

Petherick, J. C., & Duncan, I. J. H. (1989). Behavior of young domestic-fowl directed towards different substrates. *British Poultry Science, 30*(2), 229-238. <https://doi.org/10.1080/00071668908417143>

Pfeiffer, M., Eggemann, L., Kransmann, J., Schmitt, A.O., & Hessel, E.F. (2019). Inter- and intra-observer reliability of animal welfare indicators for the on-farm self-assessment of fattening pigs. *Animal, 13,* 1712-1720. <https://doi.org/10.1017/S1751731118003701>

Pickel, T., Scholz, B., & Schrader, L. (2010). Perch material and diameter affects particular perching behaviours in laying hens. *Applied Animal Behaviour Science, 127*(1-2), 37-42. <https://doi.org/10.1016/j.applanim.2010.08.005>

Puvadolpirod, S., & Thaxton, J.P. (2000). Model of physiological stress in chickens 4. Digestion and metabolism. *Poultry Science, 79,* 383-390. <https://doi.org/10.1093/ps/79.3.383>

Raoult, C.M.C., Moser, J., & Gygax, L. (2017). Mood as cumulative expectation mismatch: A test of theory based on data from non-verbal cognitive bias tests. *Frontiers in Psychology, 8*, 2197. <https://doi.org/10.3389/fpsyg.2017.02197>.

Reed, H. J., & Nicol, C. J. (1992). Effects of spatial allowance, group-size and perches on the behavior of hens in cages with nests. *British Veterinary Journal, 148*(6). [https://doi.org/10.1016/0007-1935(92)90008-o](https://doi.org/10.1016/0007-1935%2892%2990008-o)

Reijgwart, M.L., Vinke, C.M., Hendriksen, C.F.M., den Hoed, K.M., van den Meer, M.I., van der Meer, M.,Schoemaker, N.J., & van Zeeland, Y.R.A*.* (2018). An exploratory study on the effect of provision of preferred and non-preferred enrichment on behavioural and physiological parameters in laboratory ferrets (*Mustela putorius furo*). *Applied Animal Behaviour Science, 203*, 64-72. <https://doi.org/10.1016/j.applanim.2018.02.014>

Roelofs, S., Boleij, H., Nordquist, R.E., & van der Staay, F.J. (2016) Making decisions under ambiguity: Judgement bias tasks for assessing emotional state in animals. *Frontiers in Behavioural Neuroscience, 10,* Article 119. <https://doi.org/10.3389/fnbeh.2016.00119>

Rolls, E.T. (2014). *Emotion and decision making explained.* Oxford University Press.

Rönchen, S., Scholz, B., Hamann, H., & Distl, O. (2008). Fat status in Lohmann Silver and Lohmann Tradition laying hens kept in modified small group housing systems, small group housing systems, furnished cages and an aviary system. *Berliner Und Munchener Tierarztliche Wochenschrift, 121*, 11-18. <https://doi.org/10.2376/0005-9366-121-11>

Ross, M., Garland, A., Harlander-Matauscheck, Kitchenham, L., & Mason, G. (2019). Welfare-improving enrichments greatly reduce hens’ startle responses, despite little change in judgement bias. *Scientific Reports, 9,* Article 11881. <https://doi.org/10.1038/s41598-019-48351-6>

Sainsbury, D. (1986). *Farm animal welfare. Cattle, pigs and poultry*. Collins.

Salmeto, A.L., Hymel, K.A., Carpenter, E.C., Brilot, B.O., Bateson, M., & Sufka, K.J. (2011). Cognitive bias in the chick anxiety-depression model. *Brain Research, 1373*, 124-130. <https://doi.org/10.1016/j.brainres.2010.12.007>

Salas, M., Temple, D., Abaigar, T., Cuadrado, M., Delclaux, M., & Ensenat, C. (2016). Aggressive behaviour and hair cortisol levels in captive Dorcas gazelles (*Gazella dorcas*) as animal-based welfare indicators. *Zoo Biology, 35*, 467-473. <https://doi.org/10.1002/zoo.21323>

Struelens, E., Tuyttens, F.A.M., Duchateau, L., Leroy, T., Cox, M., Vranken, E., Buyse, J., Zoons, J., Berkmans, D., Ödberg, F., & Sonck, B. (2008a). Perching behaviour and perch height preference of laying hens in furnished cages varying in height. *British Poultry Science, 49*(4), 381-389. <https://doi.org/10.1080/00071660802158332>

Struelens, E., Van Nuffel, A., Tuyttens, F. A. M., Audoorn, L., Vranken, E., Zoons, J., Berkmans, D., Ödberg, F., Van Dongen, S., & Sonck, B. (2008b). Influence of nest seclusion and nesting material on pre-laying behaviour of laying hens. *Applied Animal Behaviour Science, 112*(1-2), 106-119. <https://doi.org/10.1016/j.applanim.2007.07.010>

Tarlton, J.F., Wilkins, L.J., Toscano, M.J., Avery, N.C., & Knott, L. (2013). Reduced bone breakage and increased bone strength in free range laying hens fed omega-3 polyunsaturated fatty acid supplemented diets. *Bone, 52*, 578–586. <https://doi.org/10.1016/j.bone.2012.11.003>

Taylor, G.B. (1972). One man’s philosophy of welfare. *Veterinary Record, 91*, 426-428. https://doi:[10.1136/vr.91.18.426](https://doi.org/10.1136/vr.91.18.426)

Tuyttens, F., Heyndrick, M., De Boeck, M., Moreels, A., Van Nuffel, A., Van Poucke, E., Van Coillie, E., Van Dongen, S., & Lens, L. (2007). Broiler chicken health, welfare and fluctuating asymmetry in organic versus conventional production systems*. Livestock Science 113*, 123-132. <https://doi.org/10.1016/j.livsci.2007.02.019>

Vanliere, D.W., Kooijman, J., & Wiepkema, P.R. (1990). Dustbathing behavior of laying hens as related to quality of dustbathing material. *Applied Animal Behaviour Science, 26*(1-2), 127-141. [https://doi.org/10.1016/0168-1591(90)90093-s](https://doi.org/10.1016/0168-1591%2890%2990093-s)

Weeks, C.A., Knowles, T.G., Gordon, R.G., Kerr, A.E., Peyton, S.T., & Tilbrook, N.T. (2002). New method for objectively assessing lameness in broiler chickens. *Veterinary Record, 151*(25), 762-764. <https://doi:10.1136/vr.151.25.762>

Weeks, C.A., & Nicol, C.J. (2006). Behavioural needs, priorities and preferences of laying hens. *World Poultry Science Journal, 62*(2), 296-307. <https://doi.org/10.1079/WPS200598>

Welfare Quality version 2.0 (2019) http://www.welfarequality.net/media/1294/wq\_laying\_hen\_protocol\_20\_def-december-2019.pdf

Wichman, A., & Keeling, L.J. (2009). The influence of losing or gaining access to peat on the dustbathing behaviour of laying hens. *Animal Welfare, 18*(2), 149-157.

Wichman A., Keeling L.J., & Forkman, B. (2012). Cognitive bias and anticipatory behaviour of laying hens housed in basic and enriched pens. *Applied Animal Behaviour Science, 140*(1-2)*,* 62-69. <https://doi.org/10.1016/j.applanim.2012.05.006>

Widowski, T.M., & Duncan, I.J.H. (2000). Working for a dustbath: are hens increasing pleasure rather than reducing suffering? *Applied Animal Behaviour Science, 68*(1), 39-53. [https://doi.org/10.1016/s0168-1591(00)00088-5](https://doi.org/10.1016/s0168-1591%2800%2900088-5)

Yang, H. M., Yang, Z., Wang, W., Wang, Z.Y., Sun, H.N., Ju, X.J., & Qi. X.M. (2014). Effects of different housing systems on visceral organs, serum biochemical proportions, immune performance and egg quality of laying hens. *European Poultry Science 78*. <https://doi.org/10.1399/eps.2014.48>

**Appendix1. Judgement Bias Test**

*<H3>Judgement bias background*

In recent years there has been a dramatic upsurge in studies that make use of cognitive bias tasks as putative indicators of affective state and thus welfare. This approach was based originally on findings in humans that measurable cognitive biases in judgement, attention or memory arise from underlying affective states (see Paul et al., 2005 for a review). People experiencing negative affect (including anxiety and depression) tend to interpret ambiguous events and stimuli more negatively than other people and may pay more attention to potentially negative or threatening cues (Macleod and Byrne 1996, Bar-Haim et al. 2007). Mendl and Paul (2020) suggested that these findings reflect adaptive shifts in decision-making processes that are likely to be widely distributed across species. In the judgement bias task, animals are pretrained that a positive (S+) cue predicts the arrival of a reward, while a negative (S-) cue predicts no reward or a punishment (Harding et al. 2004). Once the animal responds appropriately to these cues using a conditional discrimination method, it is occasionally presented with ambiguous probe cues that are intermediate between the S+ and S- (i.e. sharing features of both). The expectation is that animals in a negative affective state will judge these ambiguous probes to share more features with the pretrained negative cues than will animals in a positive affective state (i.e. a ‘pessimistic’ or ‘pessimistic-like’ response). This is because these animals are hypothesized to have an increased tendency to expect negative events and a reduced tendency to expect positive ones (Mendl et al. 2009; Mendl et al. 2010; Mendl & Paul 2020). Such biases would be expected to be adaptive in environments in which rewards and punishers are autocorrelated across time (e.g. in reward-rich environments with plentiful food resources or threat/punishment-rich environments with many predators; Fawcett et al., 2013).

Judgement bias studies have become commonplace in animal welfare research as putative indicators of positive and negative affect, and recent meta-analyses of the effects of environmental and psychopharmacological manipulations have generally supported the hypotheses that ‘optimistic’ biases are associated with positive affect manipulations and ‘pessimistic’ biases with negative affect manipulations (Neville et al. 2020; Lagisz 2020). However, results have been mixed, reflecting wide-ranging methodological variations across studies, and some variably effective affect manipulations. While the majority of judgement bias studies have used mammals as subjects (including rats, dogs, cattle, horses, sheep, etc.), others have focused on birds (e.g. Brilot et al., 2010; Matheson et al., 2008) and even insects (Bateson et al., 2011; Perry et al., 2016).

*<H3>Judgement bias test methods*

The judgement bias task used in this study was a Go/No go task first designed and used by Deakin et al. (2016). Two hens per pen (total *N*=24; *N*=12 per experimental housing group) were trained and tested using the judgement bias apparatus. The birds were trained to peck a positive (S+) cue displayed on a computer screen in order to obtain a single mealworm reward, and to not peck a negative (S-) cue in order to avoid receiving a 1 s air puff. The cue stimuli were centrally presented, 3.5 cm diameter, orange-coloured circles of differing colour saturation (low saturation: hue: 19; luminance: 120; saturation: 50; high saturation: hue: 19; luminance: 120; saturation: 250); computer monitor: contrast: 70; brightness: 100). For half the birds, the S+ cue was of high-saturation orange and for the other half it was of low saturation; the S- cue in either case was of the opposite saturation (low or high).

For testing, three ambiguous probe cues of intermediate saturation were presented in addition to the S+ and S- anchor cues. These cues were of three types: near negative (‘near S-’), middle (‘M’) and near positive (‘near S+’). The birds’ responses to these cues were not reinforced (i.e. they were unrewarded and unpunished). Three test sessions, one per day, were performed at the end of each experimental phase and each test session comprised 40 stimuli presentations (17x S+ cues; 17x S- cues; 6x ambiguous cues, two of each type). The order in which the ambiguous probe cues were presented was the same for all birds but differed over three test sessions so that all birds saw three different sequences.

The two summary variables collected, ‘proportion of ambiguous probe cues pecked’ during test sessions (of a total of 24 probe cues presented) and ‘mean latency to peck ambiguous probe cues’, both showed largely normal distributions and were strongly inversely correlated (Phase 1: Pearson *r*=-0.989, *P*<0.01; Phase 2: Pearson *r*=-0.986, *P*<0.01; Phase 3: Pearson *r*=-0.987, *P*<0.01). In the analyses presented here, therefore, we only considered the former. Where significant results were found with this summary variable, additional analyses were conducted to identify which of the three ambiguous probes these differences were focused upon (near S-, M, near S+). A greater proportion of ambiguous probe cues pecked are generally thought of as representing more ‘optimistic’ and/or more ‘risky’ choices (Deakin et al., 2016).

*<H3>Judgement bias training and testing procedure*

*<H4>Reward training.* All birds were initially fed mealworms in their home pens for 10–15 min (four sessions). Following this habituation, hens were clicker trained in pairs, for 5 min at a time, to receive a mealworm reward when a clicker sounded (two sessions). This was conducted in a separate (experimental) room adjacent to the home pen room. Four of five birds per pen were then individually clicker trained during one additional session and the three most successfully trained birds from each pen were selected to continue to the next stage of training. This individual clicker training continued in the judgement testing box (within the experimental room), initially with the computer monitor present, but turned off (one 5 min session and one 10 min session/bird). Each bird was then given two 10 min clicker-training sessions with the computer screen turned on (but with a blank white screen). Training continued with 10 min sessions in which a black worm-shaped image (Powerpoint) was presented on the screen, which was animated to attract chickens to peck at it for a click and mealworm reward (between two seven sessions, depending on progression; Fig. A1). During each of these sessions, birds were scored using the following scoring system: NS = near screen; PS = pecking screen; APW = almost pecking worm; PW = pecking worm; to reach criterion each individual had to be scored as PW across two consecutive sessions. Following this, each hen was presented with a still (not moving) worm image on the screen for one 5 min session to ensure they would peck at the worm when it was not moving. Once each individual hen had completed this stage, the S+ (a coloured circle in the centre of the screen, either low saturation (hue: 19; luminance: 120; saturation: 50) or high saturation (hue: 19; luminance: 120; saturation: 250), was introduced (computer monitor contrast: 70; brightness: 100). The S+ was associated with obtaining a single mealworm reward. One bird from each pen was given a high-saturation S+ and the other was given a low-saturation S+. Chickens were trained using clicker training to peck at the S+ cue (some shaping was initially required to encourage birds towards the circle). Once they were reliably pecking at the circle (within 15 s of the circle being present) across three consecutive training sessions, the intertrial interval (ITI) was introduced. Initially a 5 s ITI (blank white screen) was used once the chicken pecked the S+ cue. This was achieved using a second experimenter who pressed the computer mouse to bring up the blank white screen. The second experimenter was also responsible for timing the 5 s ITI and pressing the mouse button again, to continue the Powerpoint presentation. The ITI was then increased to 10 s using the same procedure. When a bird successfully pecked the S+ cue, it received a click and a mealworm reward, and the screen went white for a 5 s ITI before reappearance of the S+ cue. If a bird did not peck the stimulus during a 15 s period, the screen went blank for the ITI and then the circle reappeared. When birds pecked the S+ cue in at least 90% of presentations to receive a reward over three consecutive 10 min sessions (two sessions per day), they moved on to discrimination training.

*<H4>Discrimination training.* In discrimination training, birds were presented with a sequence of S+ and S- cues. The S- cue was of the opposite saturation (high/low) to the S+ cue and 40 cues were presented in total per session (two sessions per day). A pseudorandom sequence of presentations was generated (Microsoft Excel) and modified to ensure that the first and last presentation was an S+ cue and there were no more than three consecutive S- cue presentations (by swapping S- cues to S+ cues at these points). Birds were clicked and rewarded for pecking at an S+ cue and were given a 1 s air puff for pecking an S- cue. Reinforcement rate for the S+ and S- cues remained at 100% for the duration of the task. Birds were given 15 s to decide whether to peck the screen, and a 5 s ITI (blank white screen) was shown between presentations. To move on to the testing phase, birds had to achieve a success rate of 90% or more on both S+ and S- cue presentations over three consecutive discrimination sessions.

*<H4>Testing.* In judgement bias tests (performed at the end of Phases 1, 2 and 3), three ambiguous probe cues (unreinforced) were interspersed between S+ and S- cue presentations. Ambiguous probe cues were intermediate between the reference cues (100, 150 and 200 saturation on the orange scale). For each bird, one ambiguous probe was ‘central’ between the S+ and S- (150; middle: ‘M’), one was central between M and S+ (100 or 200 depending on contingency; near the S+ cue: ‘near S+’) and one was central between M and S- (200 or 100 depending on contingency; near the S- cue: ‘near S-’). Each test session comprised 40 stimuli presentations. Each ambiguous probe cue was presented twice per session (six ambiguous cue presentations; 17 S+ cue presentations; 17 S- cue presentations). The order in which the ambiguous probe cues were presented was the same for all birds but differed over three test sessions so that all birds saw three sequences (session 1: M, near S+, near S-; near S+, M, near S-; session 2: near S-, M, near S+; M, near S-, near S+; session 3: near S+, near S-, M; near S-, near S+, M). The exact cue sequence in a session was randomized (MS Excel RAND() function), and then adjusted so that there were no more than three consecutive presentations of an S- cue by swapping S- cues for S+ cues where necessary. This meant that in most cases, birds were shown slightly more presentations of the S+ cue than the S- cue. On each trial in each session, birds had 15 s to decide whether to peck the cue. Pecking S+ and S- cues resulted in a mealworm or an air puff, respectively. If an ambiguous probe cue was pecked, the screen immediately went white for 5 s before the next presentation. If a cue (S+, S- or ambiguous probe cue) was not pecked, the screen went white for 5 s after the 15 s (i.e. responses to ambiguous probe cues were not reinforced). Probe cues were never presented consecutively to avoid birds becoming frustrated with the lack of reward or punishment after such cue presentation. Whether or not cues were pecked and the latencies to peck them were recorded for each trial using stop watches and a check sheet (held by a second experimenter). There was no evidence of birds learning that the ambiguous probe cues were unreinforced.

*<H4>Refresher training.* Before judgement bias testing was undertaken at the ends of Phases 2 and 3, refresher training took place to ensure that the hens were still able to complete the task. Using the same discrimination training sessions as outlined above (40 unambiguous cues per session), birds were again required to meet the criterion of 90% or more correct responses to S+ and S- cue presentations over three consecutive training sessions. For all birds, these criteria were met within 3 days of morning and afternoon sessions.

*<H3>Effects of cue colour and standardization of judgement bias scores*

Within the judgement bias procedure, hens were trained to peck orange-coloured discs on a computer screen that predicted the arrival of food and to refrain from pecking other discs that predicted the delivery of an air puff. For half the birds, the discs that predicted the reward were a high-saturation orange colour while the disc that predicted the air puff was a low-saturation orange colour. For the other half, the reverse contingency was used. Initial analyses demonstrated that this colour saturation contingency had a significant effect on the ‘proportion of ambiguous probes pecked’. Birds that had been trained to associate high-saturation orange discs with a reward (S+) pecked more ambiguous probes (Phase 1: *t*22=3.191, *P*<0.01) than those that had been trained to associate low-saturation orange discs with a reward. The most likely explanation for this is a perceptual asymmetry in the colour saturation of the probe cues used: if, for example, the middle probe cue is seen by all birds to be more similar to the high-saturation orange cue than the low-saturation one, then birds trained to associate high saturation with a food reward will peck the M cue more often (and more quickly), while birds trained to associate high saturation with the air puff will peck it less often (and less quickly). For most of the analyses presented in this paper, this training colour contingency was not considered a confounding factor, as it was counterbalanced across the two experimental groups, GP and GNP. For analyses of associations between judgement bias scores and candidate welfare indicators, however, scores standardized for colour saturation contingency were calculated. These *z* scores give a measure of an individual’s performance relative to others that were also trained on the same colour saturation contingency (i.e. they indicate how many standard deviations an individual’s score is from the mean of the population from which it is drawn; *z*=(*x*-mean *x*)/SD), thereby providing comparable scores for birds trained on different contingencies.

*<H3>Are judgement bias scores associated with motivation to feed?*

Checks were made to establish whether hens’ responses in the judgment bias task could be interpreted simply as a reflection of motivation to feed. Exploratory correlational analyses were conducted between the judgement bias summary variable ‘proportion of ambiguous probes pecked’ and a range of behavioural and physical measures that might be expected to be indicative of motivation to feed: body mass, blood glucose level, frequency of feeding and mean time spent feeding (from food hopper) during behavioural observations. These showed little evidence of an association with judgment bias; only frequency of feeding during Phase 3 focal observations was correlated with judgement bias during the same phase (Pearson *r*=0.531, *P*<0.0.01).

To experimentally address the possibility that any differences in judgement bias between the preference groups (GP and GNP) might simply be due to differences in food motivation, two additional ‘food motivation’ tasks were also conducted.

*<H4>Mealworm consumption test.* The Mealworm consumption test was used as a formalized measure of the feeding motivation of all hens. Each bird was placed on the floor of the test room and given a ceramic bowl containing 100 g of live mealworms to eat. The mass of mealworms remaining after 1 min was measured. Two tests were conducted on consecutive days, giving a measure of mean amount consumed per test; an additional score of mass consumed as proportion of the hen’s bodyweight was also calculated. The mass of mealworms eaten in this test was not associated with individuals’ performances in the judgement bias task across any of the three phases (Phase 1: *r*=0.032; Phase 2: *r*=0.210; Phase 3: *r*=0.130; all NS).

*<H4>Runway feeding test.* The runway feeding test was used to mirror the judgement bias task in an independent group of hens (two per pen). The aim was to establish, in these separate birds, whether feeding motivation was associated with preference group (GP and GNP). Because training was needed to take part in this task, it was not logistically possible to conduct it with the hens that were already being judgement bias trained (this training was also considered likely to disrupt judgement bias performance, as the same rewards were used for different contingencies). Analyses of the relationship between runway-trained birds’ mean running speed to reach a mealworm reward and their experimental group (GP versus GNP) were conducted with an independent group of birds (*N*= 20; Group GP: *N*=9; Group GNP: *N*=11). These revealed no significant association between this measure of feeding motivation and preference across any of the three phases (Phase 1: *t*17 =0.506; Phase 2: *t*17 =1.173; Phase 3: *t*17 =1.628; all NS).

 <H4>*Judgement bias sample with two trained hens per pen.* Our aim was to train two hens per pen in the judgement bias task and two hens per pen in the runway feeding task (see above). The fifth hen in each pen was used as a judgement-bias ‘back-up’. Judgement bias tests often require lengthy training and not all animals learn the task in the window of time required. Three hens were therefore trained to the reward-training level of judgement bias paradigm (see Methods, above), and then the poorest/slowest learning from each pen was discontinued.

**Appendix 2. Preference and Preference Testing**

*<H3>Published evidence used to design GP and GNP living conditions*

Evidence for designing the Generally Preferred (GP) and Generally nonpreferred (GNP) living conditions came from a range of published studies that demonstrated chickens’ preferences for particular environments or resources over others, and/or demonstrated that most chickens would work to obtain access to particular environments or resources. These studies are listed below:

Background noise and loud sounds (Jones, Bizo, & Foster, 2012; Mackenzie, Foster, & Temple, 1993; McAdie, Foster, Temple, & Matthews, 1993)

Dustbath and dustbathing substrate (Petherick & Duncan, 1989; Vanliere, Kooijman, & Wiepkema, 1990; Wichman & Keeling, 2009; Widowski & Duncan, 2000)

Discrete punishers (R. B. Jones, 1986a, 1986b)

Discrete rewards (Bouvarel et al., 2009; Bruce, Prescott, & Wathes, 2003; Moe, Nordgreen, Janczak, Spruijt, & Bakken, 2013)

Floor substrates (Dawkins & Beardsley, 1986)

High and low light intensity areas; Canopy (Bright, Brass, Clachan, Drake, & Joret, 2011; Davis, Prescott, Savory, & Wathes, 1999)

Manipulable enrichment objects (string) (Jones, 1986a, 1986b; Jones & Carmichael, 1998; McAdie, Keeling, Blokhuis, & Jones, 2005)

Nestbox design (Appleby & McRae, 1986; Cooper & Appleby, 1996, 2003; Freire, Appleby, & Hughes, 1997; Hughes, 1993; Kruschwitz, Zupan, Buchwalder, & Huber-Eicher, 2008; Reed & Nicol, 1992; Struelens, Van Nuffel, et al., 2008)

Perch design (Appleby & Hughes, 1995; Chen, Bao, Meng, & Wei, 2014; Olsson & Keeling, 2002; Pickel, Scholz, & Schrader, 2010; Reed & Nicol, 1992; Struelens, Tuyttens, et al., 2008)

Space allocation and stocking density (Reed & Nicol, 1992)

*<H3>Preference test methods*

The living condition preference tests conducted at the end of Phase 3 were used to assign individual preference groups IP and INP, by calculating the majority choice made by individual birds from groups GP and GNP after a total of seven T-maze choice tests (see Figs A2, A3).

Testing took place at the end of the Phase 3 data collection period. Two test pens were set up opposite one another on either side of an experimental room: an intermediate pen and an experimental pen (either GP or GNP living conditions, depending on which the bird to be tested had experienced during the experiment). Prior to testing, each group of five pen-mates was housed overnight in the intermediate pen, to refresh their experience of that environment (previously experienced during Phase 1 of the experiment). The following morning, approximately 10 min before tests commenced, the five hens were taken from this pen and placed together in a poultry transport crate. All faeces and eggs were removed from the intermediate pen and fresh food and water were placed in both pens. A T-maze tunnel was inserted between the intermediate pen and the experimental pen (generally preferred or generally nonpreferred), with the start box in a central position between the two. Birds were tested individually in a predetermined random order, and with a randomly allocated but individually constant T-maze start side. Initially, each bird was required to make two forced choices in pseudorandom order (one to each of the pens on offer, with 50% of birds accessing the intermediate pen first); this was achieved by opening just one of two sliding doors at the entrances to the pens. After entering a pen following a forced choice, each hen was allowed to explore the pen for 2 min, with no opportunity to re-enter the maze tunnel.

After both forced choices had been made, each hen was allowed to make seven free choices (again, each choice was followed by a 2 min period of contained exploration within the chosen pen). During these free-choice tests the latency to enter one of the two pens from being placed in the start box was recorded, together with the pen type chosen. If a hen failed to make a choice within 2 min, she was removed from the apparatus and ‘no choice’ was recorded for the test. She was then replaced within the start box to begin the next test. Those birds that showed a majority preference for their own experimental living condition (i.e. four or more of the seven choices were for their own living condition, GP or GNP) were labelled Group IP (individually preferred) and those that showed a majority preference for the intermediate living condition were labelled Group INP (individually nonpreferred).

*<H3>Preference test results*

Birds living in the generally preferred living conditions (GP) were more likely to show an individual preference (IP) for their own experimental condition than were birds living in the generally nonpreferred living conditions (Group GNP; 73.3% of Group GP birds and 25.0% of GNP birds preferred to access a pen resembling their own experimental living condition (2=13.533, *P*<0.001). On the basis of these tests, 29 birds (22 from Group GP and seven from Group GNP) were assigned to individual preference Group IP and 29 to Group INP (eight from Group GP and 21 from Group GNP). Of the seven preference tests conducted with each bird, the relative preference Group GP birds showed for their own experimental living conditions (ratio of home choice/total number of choices made) was significantly greater than that of Group GNP birds (Group GP ratio 0.711 (SD0.225); Group GNP ratio 0.374 (SD0.358); *t*56 =4.332, *P*<0.001). When choices for the birds’ own living conditions were made, Group GP birds did not decide more quickly (mean latency 16.656 s (SD 13.768)) than Group GNP birds (mean latency 24.722 s (SD 18.926); *t*50=-1.781, *P*=0.081).

For the subset of birds trained to participate in judgement bias tests, a similar but nonsignificant pattern of results was found. More birds living in the generally preferred living conditions (Group GP) showed an individual preference (IP) for their own experimental condition (50.0% of Group GP birds and 25.0% of GNP birds preferred to access a pen resembling their own experimental living conditions), but this trend was nonsignificant (2=1.60,NS). Of the seven tests, the relative preference Group GP birds showed for their own experimental living conditions (ratio of home choice/total number of choices made) was nonsignificantly greater than that of Group GNP birds (Group GP ratio 0.625 (SD 0.274); Group GNP ratio 0.373 (SD 0.329) (*t*22=2.034, *P*=0.054)).When choices for the birds’ own living conditions were made, Group GP birds did not decide more quickly (mean latency 16.291 s (SD 12.915)) than Group GNP birds (mean latency 30.043 s (SD 19.840); *t*20 =-1.959, *P*=0.064).

**Appendix 3. Other Candidate Welfare Indicators**

*<H3>Ethogram used in scan and focal observations*

Behaviours were recorded mutually exclusively, with the most fine-grained behaviour taking priority (e.g. if a hen was sitting on the floor and feather pecking another bird, this was recorded as ‘Feather peck given’ rather than ‘Sit’). The same ethogram was used for scan and focal observations and was based on Nicol et al. (2009). The full ethogram, with behaviour definitions, is listed in Table A1.

**TABLES**

|  |  |  |  |
| --- | --- | --- | --- |
| **Table 1.** Means (and SDs in parentheses) of physiological measures for Group GP and Group GNP hens, across three phases of study | Phase 1 | Phase 2 | Phase 3 |
|  | GP | GNP | GP | GNP | GP | GNP |
| Faecal water (%) | 74.10(2.32)*N*=30 | 74.39(2.41)*N*=30 | 73.12(2.98)*N*=30 | 75.64(3.00)*N*=29 | 72.93(2.69)*N*=30 | 77.49(2.92)*N*=29 |
|  |  | β=-2.467 (SE 0.758)*P*=0.002 | β=-4.544 (SE 0.930)*P*<0.001 |
| Glucose (mmol/litre) | 13.99(0.824)*N*=28 | 13.669(1.311)*N*=29 | 13.11(0.611)*N*=28 | 13.138(0.827)*N*=29 | 13.029(0.733)*N*=28 | 13.733(1.180)*N*=27 |
|  |  | β =-0.0655 (SE 0.189)*P*=0.729 | β =-0.807 (SE 0.401)*P*=0.058 |
| H/L ratio | 1.270(0.510)*N*=27 | 1.231(0.599)*N*=26 | 1.642(0.895)*N*=26 | 1.341(0.760)*N*=24 | 1.366(1.141)*N*=25 | 2.499(3.210)*N*=22 |
|  |  | Non-normal dataMWU=241*P*=0.347 | Non-normal dataMWU=329*P*=0.746 |
| Corticosterone (ng/ml) | 2.461(1.222)*N*=27 | 2.898(1.664)*N*=29 | 2.188(1.407)*N*=27 | 2.315(1.206)*N*=28 | 1.820(1.099)*N*=27 | 2.356(1.144)*N*=28 |
|  |  | β =0.0162 (SE 0.313) *P*=0.959 | β =-0.294 (SE 0.229)*P*=0.201 |
| GGT (mmol/litre) | 35.381(19.868)*N*=21 | 33.042(18.227)*N*=24 |  |  | 31.364(29.867)*N*=11 | 31.385(29.299)*N*=13 |
|  |  |  | β =0.540 (SE11.550)*P*=0.963 |
| ALP (mmol/litre)(log transformed for analyses) | 465.238(357.943)*N*=21 | 505.240(368.436)*N*=25 |  |  | 490.444(280.019)*N*=18 | 1292.174(1977.948)*N*=23 |
|  |  |  | β =0.265 (SE 0.091)*P*=0.013 |
| Triglycerides (mmol/litre) | 14.71(6.118)*N*=25 | 13.317(4.089)*N*=26 | 15.068(8.361)*N*=24 | 17.878(9.654)*N*=24 | 17.811(5.928)*N*=22 | 12.877(5.003)*N*=23 |
|  |  | β =-3.915 (SE 2.272)*P*=0.09 | β =4.233 (SE 1.888)*P*=0.039 |
| Mass (g) | 1740.47(135.02)*N*=30 | 1762.80(136.17)*N*=30 | 1815.53(145.61)*N*=30 | 1794.37(143.49)*N*=30 | 1830.17(143.47)*N*=30 | 1792.31(150.88)*N*=29 |
|  |  | β =40.720 (SE 18.800)*P*=0.034 | β =51.320 (SE 23.760)*P*=0.034 |
| Claw length (mm) | 14.03(1.87)*N*=30 | 13.73(1.46)*N*=30 | 13.83(1.18)*N*=30 | 17.03(1.56)*N*=30 | 12.93(1.55)*N*=30 | 19.00(2.22)*N*=29 |
|  |  | β =-3.266 (SE 0.363)*P*<0.001 | β =-6.14 (SE 0.602)*P*<0.001 |

There were 30 hens per group; *N*s are smaller for some measures when insufficient blood samples were available, etc.).

**Table 2.** Percentages of Group GP and Group GNP hens (*N*=30 per group) showing physical characteristics, across three phases of study

|  |  |  |  |
| --- | --- | --- | --- |
|  | Phase 1 | Phase 2 | Phase 3 |
|  | GP | GNP | GP | GNP | GP | GNP |
| Body condition low (keel bone detectable) | 63.3*N*=30 | 70.0*N*=30 | 60.0*N*=30 | 66.7*N*=30 | 56.7*N*=30 | 69.0*N*=29 |
|  |  | β=0.212 (SE 0.571)*P*=0.710 | β=0.500 (SE 0.564)*P*=0.375 |
| Comb lesions present | 30.0*N*=30 | 43.3*N*=30 | 46.7*N*=30 | 46.7*N*=30 | 26.7*N*=30 | 51.7*N*=29 |
|  |  | β =0.269 (SE 0.575)*P*=0.640 | β =-1.031 (SE 0.588)*P*=0.080 |
| Comb small | 20.0*N*=30 | 20.0*N*=30 | 43.3*N*=30 | 36.7*N*=30 | 30.0*N*=30 | 10.3*N*=29 |
|  |  | β=-0.288 (SE 0.538)*P*=0.595 | β =-2.390 (SE 1.165)*P*=0.040 |
| Comb elevation low | 10.0*N*=30 | 36.7*N*=30 | 26.7*N*=30 | 43.3*N*=30 | 26.7*N*=30 | 50.0*N*=29 |
|  |  | β =0.303 (SE 0.776)*P*=0.696 | β =0.128 (SE 0.803)*P*=0.874 |
| Comb colour pale | 53.3*N*=30 | 56.7*N*=30 | 56.7*N*=30 | 63.3*N*=30 | 43.3*N*=30 | 53.3*N*=29 |
|  |  | β =-0.268 (SE 0.583)*P*=0.645 | β=-0.341 (SE 0.542)*P*=0.529 |
| Foot lesions present | 16.7*N*=30 | 10.0*N*=30 | 16.7*N*=30 | 13.3*N*=30 | 16.7*N*=30 | 6.9*N*=29 |
|  |  | β=0.136 (SE 0.753)*P*=0.856 | β =0.997 (SE 0.884)*P*=0.260 |

**Table 3.** Percentage of scans in which behaviour patterns were observed in home pens during observations of Group GP and Group GNP hens (*N*=30 per group), across three phases of study (mean percentage of all scans)

|  |  |
| --- | --- |
|  |  |
|  | Phase 1 | Phase 2 | Phase 3 |
|  | GP | GNP | GP | GNP | GP | GNP |
| Feed (from hopper) (%) | 37.01*N*=30 | 32.99*N*=30 | 34.27*N*=30 | 39.02*N*=30 | 25.98*N*=30 | 36.78*N*=29 |
|  |  | β =11.714 *(*SE 6.572)NS | β=6.963 (SE 5.090)NS |
| Ground forage(%) | 21.54*N*=30 | 21.20*N*=30 | 19.66*N*=30 | 10.21*N*=30 | 27.61*N*=30 | 16.76*N*=29 |
|  |  | β =-13.880 (SE 1.761)*P*<0.001 | β=-9.648 (SE 2.466)*P*<0.001 |
| Stand alert(%) | 19.23*N=30* | 20.09*N=30* | 22.48*N=30* | 27.14*N=30* | 24.27*N=30* | 27.98*N=29* |
|  |  | β=-0.522 (SE 2.716)NS | β=3.655 (SE 2.620)NS |
| Preen(%) | 7.35*N*=30 | 7.78*N*=30 | 7.26*N*=30 | 8.29*N*=30 | 19.32*N*=30 | 6.98*N*=29 |
|  |  | β=1.029 (SE 1.865)NS | β=-2.931 (SE 2.238)NS |
| Walk(%) | 5.81*N*=30 | 5.30*N*=30 | 5.47*N*=30 | 4.32*N*=30 | 4.44*N*=30 | 2.74*N*=29 |
|  |  | β =-0.884 (SE 0.744)NS | β =0.662 (SE 0.933)NS |
| Nest(%) | 3.42*N*=30 | 5.98*N*=30 | 4.02*N*=30 | 3.29*N*=30 | 3.16*N*=30 | 1.33*N*=29 |
|  |  | β =-1.059 (SE 1.953)NS | β=-1.589 (SE 1.803)NS |
| Drink (%) | 2.05*N*=30 | 1.71*N*=30 | 1.97*N*=30 | 1.50*N*=30 | 1.11*N*=30 | 1.86*N*=29 |
|  |  | β=0.465 (SE 0.485)NS | β =0.183 (SE 0.455)NS |
| Dustbathe(%) | 0.85*N*=30 | 1.62*N*=30 | 0.26*N*=30 | 0*N*=30 | 0.85*N*=30 | 0*N*=29 |
|  |  | β =-0.554 (SE 0.439)NS | β =-1.704 (SE 1.529)NS |
| Sit(%) | 0.85*N*=30 | 0.77*N*=30 | 1.45*N*=30 | 1.03*N*=30 | 2.48*N*=30 | 0.53*N*=29 |
|  |  | β=0.088 (SE 0.666)NS | β =-1.959 (SE 0.818)*P*=0.017 |

**Table 4.** Means (and SDs in parentheses) of physiological measures for Group IP and Group INP hens (*N*=30 per group), across three phases of study

|  |  |  |  |
| --- | --- | --- | --- |
|  | Phase 1 | Phase 2 | Phase 3 |
|  | IP | INP | IP | INP | IP | INP |
| Faecal water (%) | 73.90(2.13)*N*=29 | 74.46(2.56)*N*=29 | 73.62(3.51)*N*=29 | 75.00(2.87)*N*=29 | 74.25(4.07)*N*=29 | 76.06(2.91)*N*=29 |
|  |  | β =-1.241 (SE 0.839)*P*=0.143 | β =-0.827 (SE 0.798)*P*=0.307 |
| Glucose (mmol/litre) | 14.137(1.117)*N*=27 | 13.554(0.898)*N*=28 | 12.848(0.718)*N*=27 | 13.378(0.662)*N*=28 | 13.244(1.001)*N*=27 | 13.448(1.048)*N*=27 |
|  |  | β =-0.656 (SE 0.181)*P*<0.001 | β =-0.0694 (SE 0.244)*P*=0.779 |
| H/L ratio | 1.190(0.438)*N*=25 | 1.334(0.630)*N*=27 | 1.446(0.747)*N*=25 | 1.553(0.928)*N*=25 | 1.729(1.406)*N*=24 | 2.109(3.069)*N*=24 |
|  |  | β =-0.125 (SE 0.236)*P*=0.596 | β =-0.398 (SE 0.685)*P*=0.562 |
| Corticosterone (ng/ml) | 2.366(0.877)*N*=26 | 3.003(1.849)*N*=28 | 1.981(1.191)*N*=26 | 2.534(1.837)*N*=27 | 1.757(0.930)*N*=26 | 2.384(1.267)*N*=28 |
|  |  | β =-0.344 (SE 0.325)*P*=0.292 | β =-0.315 (SE 0.235)*P*=0.183 |
| GGT (mmol/litre) | 33.521(19.273)*N*=23 | 35.35(19.637)*N*=20 |  |  | 22.417(27.985)*N*=12 | 36.455(25.874)*N*=11 |
|  |  |  | β =-13.25 (SE 11.11)*P*=0.240 |
| ALP (mmol/litre)(log transformedfor analyses) | 440.869(325.074)*N*=23 | 524.952(407.746)*N*=21 |  |  | 578.250(302.084)*N*=20 | 1300.5(2139.136)*N*=20 |
|  |  |  | β =-0.037 (SE 0.071)*P*=0.606 |
| Triglycerides (mmol/litre) | 13.736(5.826)*N*=26 | 14.024(4.527)*N*=23 | 16.016(8.106)*N*=24 | 16.422(9.946)*N*=22 | 16.870(6.689)*N*=22 | 13.973(4.857)*N*=22 |
|  |  | β =-0.0929 (SE2.301)*P*=0.968 | β =2.302 (SE 1.516)*P*=0.141 |
| Mass (g) | 1746.90(149.64)*N*=29 | 1744.10(118.15)*N*=29 | 1807.48(159.70)*N*=29 | 1797.40(128.55)*N*=29 | 1830.31(154.92) *N*=29 | 1793.43(139.38)*N*=29 |
|  |  | β =9.956 (SE 19.89)*P*=0.617 | β =38.69 (SE 24.52)*P*=0.118 |
| Claw length (mm) | 13.76(1.83)*N*=29 | 14.07(1.56)*N*=29 | 14.83(1.93)*N*=29 | 16.03(2.19)*N*=29 | 14.48(3.51)*N*=29 | 17.30(3.15)*N*=29 |
|  |  | β =0.846 (SE 0.388)*P*=0.047 | β =0.0327 (SE 0.568)*P*=0.955 |

**Table 5.** Percentages of Group IP and Group INP hens (*N*=30 per group) showing physical characteristics, across three phases of study

|  |  |  |  |
| --- | --- | --- | --- |
|  | Phase 1 | Phase 2 | Phase 3 |
|  | IP | INP | IP | INP | IP | INP |
| Body condition low (keel bone detectable) | 62.07*N*=29 | 72.41*N*=29 | 55.17*N*=29 | 75.86*N*=29 | 58.62*N*=29 | 68.97*N*=29 |
|  |  | β=0.878 (SE 0.606)*P*=0.147 | β=0.363 (SE 0.567)*P*=0.522 |
| Comb lesions present | 37.93*N*=29 | 31.03*N*=29 | 62.07*N*=29 | 27.59*N*=29 | 31.03*N*=29 | 44.83*N*=29 |
|  |  | β=1.574 (SE 0.622)*P*=0.011 | β=-0.795 (SE 0.597)*P*=0.183 |
| Comb size small | 27.59*N*=29 | 13.79*N*=29 | 41.38*N*=29 | 37.93*N*=29 | 31.03*N*=29 | 10.34*N*=29 |
|  |  | β=-0.009 (SE 0.556)*P*=0.987 | β=-1256 (SE 0.867)*P*=0.148 |
| Comb elevation low | 13.8*N*=29 | 31.0*N*=29 | 31.0*N*=29 | 37.9*N*=29 | 27.6*N*=29 | 44.8*N*=29 |
|  |  | β=0.491 (SE 0.772)*P*=0.525 | β=-0.272 (SE 0.781)*P*=0.728 |
| Comb colour pale | 51.7*N=29* | 58.6*N=29* | 65.5*N=29* | 58.6*N=29* | 44.8*N=29* | 51.7*N=29* |
|  |  | *β* =0.529 (SE 0.616)*P*=0.391 | *β* =-0.222 (SE 0.544)*P*=0.683 |
| Foot lesions present | 20.69*N*=29 | 6.90*N*=29 | 13.79*N*=29 | 17.24*N*=29 | 10.34*N*=29 | 10.34*N=29* |
|  |  | β=-0.624 (SE 0.803)*P*=0.437 | *β* =-0.039 (SE 0.883)*P*=0.965 |

**Table 6.** Percentage of scans in which behaviour patterns were observed in home pens during observations of Group IP and Group INP hens (n=30 per group), across three phases of study (mean percentage of all scans)

|  |  |
| --- | --- |
|  |  |
|  | **Phase 1** | **Phase 2** | **Phase 3** |
|  | **IP** | **INP** | **IP** | **INP** | **IP** | **INP** |
| **Feed (from hopper)** (%) | 35.01*N=29* | 35.37*N=29* | 33.60*N=29* | 40.14*N=29* | 30.86*N=29* | 31.39*N=29* |
|  |  | *β* =3.914 (*SE*2.736)*N.S.* | *β* =0.386(*SE*3.302)*N.S.* |
| **Ground forage**(%) | 23.43*N=29* | 19.71*N=29* | **18.74*****N=29*** | **11.76*****N=29*** | 25.55*N=29* | 19.32*N=29* |
|  |  | ***β* =-6.457 (*SE*2.802)*****P*=0.002** | *β* =-2.347(*SE*2.326)*N.S.* |
| **Stand alert**(%) | 19.10*N=29* | 20.25*N=29* | 24.31*N=29* | 25.29*N=29* | 23.70*N=29* | 28.34*N=29* |
|  |  | *β* =0.187 (*SE*2.099)*N.S.* | *β* =0.505(*SE*1.988)*N.S.* |
| **Preen**(%) | 7.69*N=29* | 7.16*N=29* | 7.87*N=29* | 7.25*N=29* | 8.58*N=29* | 7.78*N=29* |
|  |  | *β* =0.160 (*SE*1.369)*N.S.* | *β* =0.756(*SE*1.889)*N.S.* |
| **Walk**(%) | 4.95*N=29* | 5.92*N=29* | 4.86*N=29* | 5.04*N=29* | 4.16*N=29* | 3.01*N=29* |
|  |  | *β* =0.601 (*SE*0.571)*N.S.* | *β* =0.182(*SE*0.580)*N.S.* |
| **Nest**(%) | 3.45*N=30* | 5.57*N=30* | 3.71*N=30* | 2.82*N=30* | 2.48*N=30* | 2.12*N=29* |
|  |  | *β* =-1.328 (*SE*1.582)*N.S.* | *β* =-1.080(*SE*1.583)*N.S.* |
| **Drink** (%) | 1.95*N=29* | 1.77*N=29* | 1.42*N=29* | 2.03*N=29* | 1.24*N=29* | 1.77*N=29* |
|  |  | *β* =0.244 (*SE*0.330)*N.S.* | *β* =0.048(*SE*0.294)*N.S.* |
| **Dustbathe**(%) | 1.50*N=29* | 1.06*N=29* | 0.18*N=29* | 0.09*N=29* | 0.61*N=29* | 0.27*N=29* |
|  |  | *β* =0.169 (*SE*0.356)*N.S.* | *β* =-0.404(*SE*0.696)*N.S.* |
| **Sit**(%) | 0.97*N=29* | 0.71*N=29* | 1.33*N=29* | 1.24*N=29* | 1.33*N=29* | 1.77*N=29* |
|  |  | *β* =0.644 (*SE*0.516)*N.S.* | *β* =0.218(*SE*0.797)*N.S.* |

**Table 7.** Coefficients (β) and *P* values from regression analyses of associations between hens’ summary judgement bias scores (proportion of ambiguous probe cues pecked, standardized for reward colour used in training; see Appendix 1) and other candidate welfare indicators measured during Phase 1 data collection

|  |  |  |
| --- | --- | --- |
| Candidate welfare indicator | β (SE) | *P* |
| Sit  | 0.065 (0.130) | 0.614 |
| Stand | -0.004 (0.042) | 0.921 |
| Walk | 0.025 (0.110) | 0.819 |
| Drink | 0.125 (0.144) | 0.384 |
| Dustbathe | 0.003 (0.041) | 0.933 |
| Feed (from hopper) | -0.006 (0.015) | 0.685 |
| Nest | -0.006 (0.014) | 0.675 |
| Preen | 0.001 (0.025) | 0.980 |
| Forage | 0.012 (0.021) | 0.579 |
| Mass | 0.0016 (0.0017) | 0.330 |
| Claw length | -0.099 (0.128) | 0.600 |
| Triglycerides | 0.057 (0.034) | 0.088 |
| ALP | -0.0002 (0.0006) | 0.695 |
| GGT | -0.0061 (0.0182) | 0.738 |
| Glucose | -0.275 (0.169) | 0.104 |
| H/L ratio | -0.487 (0.356) | 0.171 |
| Corticosterone | -0.176 (0.133) | 0.187 |
| Faecal water | 0.085 (0.069) | 0.219 |
| High perch latency | 0.111 (0.396) | 0.778 |
| Water box latency  | 0.11 (0.389) | 0.642 |
| Comb lesions | -0.816 (0.452) | 0.071 |
| Comb colour | 0.151 (0.395) | 0.702 |
| Comb elevation | 0.230 (0.427) | 0.590 |

**Table 8.** Summary of candidate welfare indicators varying significantly between preference groups GP/GNP and IP/INP, across short-term (Phase 2) and long-term (Phase 3) timescales

|  |  |  |  |
| --- | --- | --- | --- |
| Preference groups compared | Candidate welfare indicators, sorted in order of significance level | *P* | Significant following Benjamini correction  |
| GP versus GNP (Phase 2) | Claw length | <0.001 | \*\* |
|  | Forage (behaviour) | <0.001 | \*\* |
|  | Faecal water | 0.002 | \*\* |
|  | High perch latency  | 0.02 |  |
|  | Mass | 0.034 |  |
| GP versus GNP (Phase 3) | Faecal water | <0.001 | \*\* |
|  | Claw length | <0.001 | \*\* |
|  | Forage (behaviour) | 0.005 | \* |
|  | Middle claw symmetry  | 0.007 | \* |
|  | Pulse rate | 0.009 | \* |
|  | ALP Tibia bone mineral density | 0.0130.015 | \*\* |
|  | Sit behaviour | 0.017 | \* |
|  | Tail feather Symmetry  | 0.027 |  |
|  | Tibia modulus  | 0.027 |  |
|  | Mass | 0.034 |  |
|  | Triglycerides | 0.039 |  |
|  | Comb size small | 0.04 |  |
| IP versus INP (Phase 2) | Blood glucose | <0.001 | \*\* |
|  | Forage (behaviour) | 0.002 | \*\* |
|  | Comb lesions\* | 0.011 | \* |
|  | Claw length | 0.047 |  |
| IP versus INP (Phase 3) | Middle claw symmetry  | 0.004 | \* |
|  | Pulse rate | 0.016 | \* |
|  | Tibia symmetry | 0.028 |  |

Applying a false discovery rate (FDR) of 0.2, the comparison *P* equivalent to *P*<0.05 was *P*<0.0179 while for an FDR of 0.05 it was *P*<0.0024; remaining significant associations following these Benjamini corrections are indicated in the rightmost column: \**P*<0.2; \*\* *P*<0.05.

**Table A1.** Behaviours recorded during both scan and focal observations from videos made in the home pens of all birds

| Behaviour | Description |
| --- | --- |
| Aggressive attack given\* | Bird directs threat or severe pecks towards head region of a conspecific, sometimes followed by a chase |
| Aggressive attack received\* | Receipt of the above behaviour |
| Beak peck given | Bird gently pecks at or around the beak of a conspecific |
| Beak peck received | Receipt of above behaviour |
| Beak wipe | Beak placed on pen floor or perch and rubbed from side to side |
| *Drink* | Intake of water from hanging drinkers in pen |
| *Dustbathe* (inc. sham dustbathe when no substrate available) | Lie on side, scratch at pen/dustbath floor, rub head and neck on floor, open wings |
| *Feed from hopper* | Ingestion of commercial feed ration from feeder |
| Flap wings | Bilateral movement of the wings including wing raising |
| Furniture peck | Pecks at any part of the pen walls or fixtures |
| *Ground peck* | Pecks at the ground for food, litter, etc. |
| *Ground scratch* *(*combined with Ground peck to make *‘Ground forage’)* | Scratching movements at the ground, of varying intensity performed in a slight crouching position |
| Mild feather peck given\*  | Gentle peck at plumage or particles on plumage, but no feather removal |
| Mild feather peck received\* | Receipt of the above behaviour |
| *Nest* | Suite of sitting, scratching, turning and pecking behaviours occurring prior to egg laying |
| *Preen* | Raise feathers and clean and realign them with beak |
| Scratch self | Leg brought upwards and forwards under wing to scratch the lowered head |
| Severe feather peck given\* | Severe peck at plumage or particles on plumage, with attempted and/or actual feather removal |
| Severe feather peck received\* | Receipt of the above behaviour |
| Shake head\* | Rapid rotatory movement of the head, accompanied by a slight raising of head and neck feathers. A less pronounced side-to-side flick occasionally observed |
| *Sit* (alert) | Sitting position, alert (sitting as a component of nesting or dustbathing not included) |
| *Stand alert* | Stand with alert body and upright neck and head |
| Stretch wing | Unilateral backward and downward stretching of wing and leg together |
| Tail wag | Rapid side-to-side movement of tail |
| *Walk* | Unhurried steps on pen floor |

Based on Nicol, Caplen, et al., 2009). Scan sample analyses were made on the most common behaviours observed (at least 1% of all hens’ time observed; shown in italics). Less frequent social, emotional and aggressive behaviours used in the analyses from focal observations are indicated with an asterisk (\*).

**FIGURE LEGENDS**

**Figure 1.** (a) Results obtained using general and individual preference are identical, and corroborating evidence of welfare state is obtained from judgement bias and other candidate welfare indicators. The ‘location’ of welfare state can be pinpointed with accuracy. (b) Results obtained using general and individual preference are similar but not identical. Results from judgement bias and other candidate indicators indicate the same approximate location but with far greater uncertainty.

**Figure 2.** Timeline of the experiment***.***

**Figure 3.** (a) Generally Preferred living condition pen (GP); (b) Generally Nonpreferred living condition pen (GNP).

**Figure 4**. Mean proportion of cues pecked in the Go/No go judgement bias task (± 1 SEM) for all judgement bias-trained hens (*N*=24), across Phases 1, 2 and 3.

**Figure 5.** Mean proportion of cues pecked in the Go/No go judgement bias task (± 1 SEM) for Groups GP (*N*=12) and GNP (*N*=12): (a) Phase 1, (b) Phase 2 and (c) Phase 3.

**Figure 6.** Mean proportion of cues pecked in the Go/No go judgement bias task (± 1 SEM) for Groups IP (*N*=9) and INP (*N*=15): (a) Phase 1, (b) Phase 2 and (c) Phase 3.

**Figure A1.** Judgement bias training, showing hen, experimenter, judgement bias testing box, animated ‘worm’ on computer screen and mealworm rewards delivery tube.

**Figure A2.** Preference testing apparatus for Group GP birds, which were given a choice between a replica of their experimental housing condition pen (generally preferred) and a replica of the intermediate pen, which was experienced by all birds during Phase 1.

**Figure A3.** Preference testing apparatus for Group GNP birds, which were given a choice between a replica of their experimental housing condition pen (generally nonpreferred) and a replica of the intermediate pen, which was experienced by all birds during Phase 1.