.

A review of current indicators of welfare in captive elephants (*Loxodonta africana* and *Elephas maximus*)

Ellen Williams[a*], Carly L. Chadwick[b], Lisa Yon[c+] and Lucy Asher[d+]

[a] School of Animal Rural and Environmental Sciences, Nottingham Trent University, Brackenhurst Campus, Southwell, Nottinghamshire, NG25 0QF, UK

[b] Ecosystems and Environment Research Centre, School of Environment and Life Sciences, University of Salford, The Crescent, Greater Manchester, M5 4WT, UK

[c] School of Veterinary Medicine and Science, University of Nottingham, Sutton Bonington Campus, Loughborough, Leicestershire, LE12 5RD, UK

[d] Centre for Behaviour and Evolution, Institute of Neuroscience, Newcastle University, Framlington Place, Newcastle NE2 4HH, UK

*Corresponding author

Ellen Williams

Nottingham Trent University, School of Animal Rural and Environmental Sciences, Brackenhurst Campus, NG25 0QF, UK

Tel: 07912 755482

Email: ellen.williams@ntu.ac.uk

[+]both authors had an equal contribution

Running title: Indicators of welfare in captive elephants

.

Abstract

Concerns over elephant welfare in UK zoos have implications for their future in captivity. To monitor improvements made to elephant welfare in UK zoos, non-invasive, valid and reliable indicators of welfare are needed. Using a rapid review strategy and critical appraisal tool, we aimed to appraise evidence from peer-reviewed literature on potential welfare indicators for captive elephants. Scopus, Web of Knowledge and Ovid were searched in January 2014 using terms relevant to captive elephants and welfare assessment. Inclusion and exclusion criteria were applied and remaining articles were critically appraised against a specially designed welfare indicator appraisal tool. Thirty-seven unique indicators of welfare were extracted from 30 peer-reviewed papers which met the inclusion criteria. Behavioural measures of welfare (n=21) were more common than either physical (n=11) or physiological (n=5) measures. Stereotypies were the most frequently used behavioural measure, glucocorticoids were the most frequently used physiological measure and body condition scores were the most frequently used physical measure. There was most support for the following indicators of improved welfare state: reduced stereotypies, reduced glucocorticoids and improved body condition scores. Additional measures which require further validation but had strong associations with the most supported measures, and thus have potential use in welfare assessment, were: increased lying rest and positive social interactions. Further validation of the described measures is needed, but this information forms a crucial part of knowledge required to efficiently monitor and improve the welfare of elephants in captivity.

Key words: animal behaviour; animal welfare; captivity; elephant; welfare assessment; welfare indicators

.

Introduction

Zoo elephant welfare across North America and Europe has been publicly criticised in influential reports (Clubb & Mason, 2002; Kiiru, 2007). These reports have led to a widespread response from animal welfare organisations and the UK Government (Zoos Forum, 2010; Born Free, 2015; PETA, 2015). In 2008, the UK government called for an independent study with the remit of providing 'objective, independent data on the welfare of elephants in the UK' (Harris *et al*., 2008). The results from the report by Harris and colleagues (2008) were reviewed by the Zoos' Forum (a government advisory committee) and then used to make recommendations to government. A range of areas of concern were highlighted in the report by Harris and colleagues (2008). In response to these concerns the Zoos' Forum stated that unless substantial improvements were shown in the health and welfare of captive elephants in the UK and unless there was a compelling reason to breed elephants in the UK, then UK zoos should take steps to stop keeping elephants (Zoos Forum, 2010). In order to document improvements in welfare in any species, including elephants, valid welfare indicators are needed. Here we use an evidence synthesis approach to identify a suite of welfare indicators for elephants.

For the purposes of this review, animal welfare is considered to be a concept which encompasses both mental and physical health, engagement with the physical or social environment and the opportunity to exhibit control or choice. This is purposefully similar to the definition by Dawkins (2008) who defined welfare as whether or not an animal is healthy and has what it wants. Maintaining a high level of welfare for animals in any captive environment is of paramount importance. The very nature of the captive environment usually means individuals are exposed to a range of situations which, in all likelihood, they would rarely, if ever, experience in the wild. However, the assessment of wild animal welfare in captive contexts can be difficult. There are typically few animals of each species in captivity, little standardisation in husbandry and housing (Mason, 2010), and limited scope to perform experimental rather than observational studies. Hill and Broom (2009) suggested that the ability of an individual animal to cope with challenges faced in

.

79 captivity is dependent upon their background and previous experiences. Particularly for elephants, a
80 long-lived species which in captivity have a wide variety of different backgrounds and experiences,
81 measuring individual welfare may be important. Tracking the response of each animal to changes in
82 their environment may allow for measurement of welfare on an individual level.

83 A number of welfare indicators have previously been identified and used for assessing
84 wellbeing in captive elephants (see Mason & Veasey, 2010 for a full review). The most validated of
85 which were expression of stereotypies (behaviours defined as 'repetitive, invariant behaviour
86 patterns with no obvious goal or function' (Mason, 1991)) and levels of glucocorticoids (GC) (Mason
87 & Veasey, 2010). However, it is widely agreed that the use of stereotypies as a sole indicator of
88 welfare must be treated with caution, as if they have become habitual it is likely they are not reliable
89 indicators of current welfare state as they can persist in circumstances that have improved welfare
90 (Mason & Latham, 2004). GC must also be interpreted with care as they are an indicator of arousal
91 and thus may be indicative of either positive or negative situations (Ralph & Tilbrook, 2016).
92 Furthermore, faecal glucocorticoid metabolites (FGM), which are increasing used as a non-invasive
93 measure of GC, are confounded by a number of factors, including those directly related to the sample
94 (e.g. age of sample, collection method) and biological factors (e.g. sex, age and reproductive status
95 of the animal), which can complicate interpretation (Millspaugh & Washburn, 2004). Further
96 suggested measures of welfare in zoo elephants have included skin and foot health, infant mortality
97 rates, signs of affective state and measures of preference or avoidance (Mason & Veasey, 2010).
98 When questioned about measures to assess elephant welfare, stakeholders advocated the use of a
99 range of behavioural, physical and physiological indicators of welfare (Chadwick *et al*., 2017). The
100 suggested behavioural indicators of good welfare included the presence of natural behaviours (such
101 as social interaction and environmental exploration), lying rest, positive social interactions and
102 behavioural synchrony within groups. Physical indicators of poor welfare which were suggested
103 included being overweight, having poor physical health or being physically unable to lie down.
104 Physiological indicators included GC and temporal gland secretion (Chadwick *et al*., 2017). Recent

.

work has investigated the relationship between ovarian cyclicity, prolactin, recumbence, musculoskeletal health, foot health, daily walking distance, BCS and stereotypies and the zoo environment, social life and management (Meehan *et al*., 2016). However, despite repeated use of some of these measures in the literature, not all of these measures have been validated for use in welfare assessment.

Hill and Broom (2009) recognised the importance of employing a suite of related measures to attempt to identify the welfare state of an individual animal. A number of papers have assessed one or more behavioural, physical or physiological measurement of elephants in captivity (including zoos, circuses, timber camps) both in their current environment or following changes to their environment or routine (e.g. loss of a conspecific, change in housing), although the term 'assessment of welfare' was rarely used. Links between the measurements used have occasionally been discussed; however, the reliability and validity of these indicators has never been assessed. Veasey (2006) suggested that documentation of baseline time budgets and comparison with time budgets in new environmental or social conditions, or comparison with wild elephant time budgets may also be a valid means of measuring welfare. Furthermore, being able to reliably predict how a measure of welfare may change following a change of circumstance forms a measure of validity (Meagher, 2009).

In order to accurately assess captive elephant welfare through non-invasive measures, it is essential to identify and describe those indicators which provide a reliable and valid assessment of the welfare state of the animal being observed, both at a given time and over a period of time. The indicators should differ between animals in different states of welfare, and results should be repeatable to allow assessment of change over time. In this manuscript, we review and appraise current indictors of welfare which have been applied to individual captive elephants and which have been published in the peer-reviewed literature.

Methods

Search methods – rapid review and critical appraisal

.

A rapid review (a systematic review which does not include grey literature – books and non-peer-reviewed journal articles in order to provide information in a timely manner) (Harker & Kleijnen, 2012) was undertaken in January 2014. Studies were identified and reviewed from searches of 'all years' on the following databases: Scopus, Web of Knowledge [Core Collection, Biosis Citation Index, Biosis Previews, Current Contents Connect, Data Citation Index, Derwent Innovations Index, Medline, Zoological Records (2007 – January 2014)] and Ovid [CAB Abstracts, Psycinfo, Zoological Records (1978-2007)]. Searches were made of titles, keywords and abstracts during January 2014 using a combination of terms relating to elephants; 'elephant', 'Elephantidae', 'Loxodonta', 'Elephas' and to welfare and husbandry in captivity e.g. 'welfare', 'quality of life', 'enrichment', 'husbandry', 'housing', 'behav*', 'stress', 'requirements', 'needs', , 'activity', 'movement', 'communication', 'health', 'anticipatory', 'handling', 'drinking', 'eating', 'functional responses', 'living conditions', 'grooming', 'rest', 'antagonis*', 'play', 'repetiti*', 'compulsion', 'self-stimulation', 'posture', 'temperament', 'traits', 'group size', 'psychology', 'learning', 'memory', 'intelligence', 'wellbeing'.

Inclusion criteria

Only publications which met all of the following criteria were included in the rapid review and subsequent critical appraisal: (1) Captive elephants (e.g. those in sanctuaries, zoos, timber camps, circuses, etc.) were the main subject of interest or the main focus of the investigation, (2) the publication contained at least two of the search terms in the abstract, (3) the publication was available to the authors in full, (4) the publication was in English, (5) the publication was in a peer-reviewed journal and (6) the publication assessed the welfare, behaviour, physiology or physical condition of an elephant at a point in time (or was a proxy for one of these, e.g. keeper assessment, questionnaires).

Exclusion criteria

Papers were excluded from the rapid review and subsequent critical appraisal if they did not meet all of the above inclusion criteria. Additionally, studies assessing the welfare of captive

.

elephants using methods which could not be applied to an individual (e.g. retrospective studies assessing population level reproduction or morbidity rates) were excluded from the review, as these did not fit with the aims of this review. Additionally measures which involved human interaction (e.g. keeper-elephant interaction) were not included, due to the complexity of analysis of such a measure. Whilst it is acknowledged that human interaction is an important aspect of welfare, individual differences in keeper-elephant relationships would mean this measure would require more complex analysis, and during this review we were seeking to identify standardised and objective measures which could be universally applied to assess welfare with relative ease.

Application of inclusion and exclusion criteria

A single author (EW) performed the initial database search and applied the inclusion and exclusion criteria to all identified publications. To ensure accuracy and consistency, a random sample of publications from the initial searches (50 papers) were independently assessed using the same inclusion and exclusion criteria by a second author (LA). Information to complete the critical appraisal (Table 1) was extracted by one author (EW) from all of the publications which met the inclusion criteria. All papers which met all of the inclusion criteria were critically appraised and included in the final review.

Critical appraisal

Each article which met all of the inclusion criteria was critically reviewed to ascertain further details about the study and to evaluate the reliability and validity of the work, prior to its inclusion in the review (Table 1). The critical appraisal tool consisted of a series of questions relating to the sample population, the study design, the reliability and validity of the paper, the sampling technique, the method of assessing welfare and the measures of welfare used (Figure 1).

Assessment of reliability and validity

Each complete article which met all of the inclusion criteria was assessed for reliability and validity, using questions as detailed in section 4 of the critical appraisal tool (Figure 1), and using predefined categories (Table 1). These were independently assessed and recorded for all papers

.

183 which met the inclusion criteria by two authors (EW, LA) (Table 2); there was no disparity between
184 the authors.

185 ***Indicators of welfare***

186 Welfare indicators were extracted from all of the critically reviewed papers and categorised
187 into three broad themes: behavioural, physiological and physical. Within each theme, welfare
188 indicators were grouped as far as possible. A note was made of whether the measures showed: (a)
189 Percentage change in concentration or frequency during the period of the study which may not have
190 been subjected to statistical analysis, (b) Statistically significant change in concentration or
191 frequency during the period of the study, (c) Correlation with any other measures of welfare
192 identified during the period of the study, (d) No change during the period of the study and/or no
193 correlation with any other measures of welfare.

194 **Results**

195 Initial searches yielded 21,000 records, of which 30 publications met all of the inclusion
196 criteria and were critically reviewed. Thirty-seven unique indicators of welfare were extracted from
197 the 30 peer-reviewed papers (see Table 2 for summary). A complete narrative review of these papers
198 is provided in supplementary material. The articles were published in 13 different journals, with the
199 majority of articles being found in Zoo Biology (11 papers), Animal Welfare (5 papers), Applied
200 Animal Behaviour Science (3 papers) and the Journal of Applied Animal Welfare Science (3
201 papers). Nearly half of these studies were conducted by researchers in the USA (14 papers).

202 ***Sample size***

203 A summary of the 30 reviewed papers is included in Table 3. The papers reviewed ranged
204 from small, single elephant, single institution studies to large, multi-institutional studies involving
205 over 100 individuals. The median number of elephants sampled was 7 (range 1 to 288). Thirty-one
206 percent of the studies assessed fewer than 4 elephants, and 9% studied a single elephant. The mean
207 number of institutions included was 8 (range 1 to 80). Only 28% of the reviewed papers were multi-
208 institutional studies.

.

Measures of welfare

Behavioural measures of welfare (21 measures identified) were far more common than either physical (11 measures identified) or physiological (5 measures identified) measures of welfare. Within behavioural measures of welfare, the most frequently used indicators were abnormal (17 papers), comfort (12 papers), feeding (10 papers), locomotion (10 papers), resting (10 papers) and social behaviours (7 papers) (Figure 2). Physical measures of welfare predominantly focused on body condition scoring (3 papers). All physiological measurements involved assessment of GC, in particular, faecal glucocorticoid metabolites (FGM) (4 papers), salivary cortisol (3 papers), and serum cortisol (2 papers). An overview of the identified measures of welfare, and whether observed changes were significant, is provided in Table 2 and Figure 2.

Behavioural measures of welfare

Behavioural indicators of welfare were broadly separated into nine categories: abnormal, sleep/rest, feeding, environmental interaction, comfort (self-maintenance), activity (walking/locomotion), inactive, social interactions and other (Table 2). Sample sizes ranged from 1 to 140 for behavioural measures of welfare (Table 2) but the majority of studies were based on 10 or fewer individuals. Correlation was observed between stereotypical behaviour and five other welfare measures: feeding (negative) (Koyama *et al*., 2012, Rees, 2009), walking (positive) (Koyama *et al*., 2012), resting (negative) (Koyama *et al*., 2012), foot health (negative) (Haspeslagh *et al*., 2013) and GC (positive) (Wilson *et al*., 2004). Sleep/rest and feeding behaviour were both correlated with walking (negative) and stereotypies (negative) (Koyama *et al*., 2012). Changes in frequency of social interactions and interactions with the environment were not correlated with any other potential welfare measures, however, associations were identified between increased environmental interaction, reduced stereotypies and increased social interactions in one paper. Frequency of walking or locomotion correlated with rest (negative), feeding (negative) and stereotypic pacing (positive) (Koyama *et al*. 2012). Frequency of comfort or self-maintenance behaviours such as dust bathing or mud wallowing, were frequently reported in the literature in

.

papers describing general activity budgets of elephants, however, despite being widely reported, comfort behaviours were not correlated with any other measures of welfare. Lesser used indicators of welfare included inactivity, play behaviour and vocalisations. Correlations between these indicators and more established indicators are yet to be reported.

Physiological indicators of welfare

Measurement of GC and FGM was carried out using various sample types: saliva (3 papers), faeces (3 papers), serum (three papers) and urine (two papers). Glucocorticoids were noted to correlate with stereotypies (positive) (Wilson *et al*., 2004) and specific personality traits (as identified using a keeper assessment of personality): 'fearful' (positive), 'effective' (described as 'gets its own way by controlling other elephants') (negative), 'sociable' (negative) and aggressive (negative) (Fanson *et al*., 2013; Grand *et al*., 2012). Glucocorticoid assessment was used in studies looking at 1 to 8 elephants (mean 5).

Physical indicators of welfare

The only reported physical indicators of welfare were body condition scores and foot health assessment. All of the study samples for assessment of physical welfare were comparatively large, generally multi-institution studies; body condition was assessed in 82 to 140 elephants (mean 114) in three studies and foot health was assessed in 87 elephants in a single study. Foot health correlated negatively with stereotypies (Haspeslagh *et al*., 2013), but otherwise visual assessment of body condition and foot health have not been validated against other behavioural or physiological indicators of welfare.

Reliability and validity of welfare indicators

The overall strength of each measure was assessed based on substantial and biologically meaningful statistical associations with other measures, whether the measure was statistically associated with a previously scientifically validated measure, and the results from the study using the assessed indicator (whether a statistically significant change in the indicator was reported, or a percentage change, and whether this change should be expected based on the conditions experienced

.

by the subject(s) in the study). Indicator strength was also assessed on an individual basis for each study, taking into account the level of validity used by the researchers in the assessment and the number of elephants assessed (Table 3). Due to paucity of information and inconsistency in reporting it was not always possible to garner enough information from the reviewed articles to assess the level of reliability. This information could therefore not be used to assess the strength of the indicator of welfare. Where available, details of test reliability are provided in Table 3. In 15 of the 30 reviewed papers, no assessment of reliability was reported, in five instances measures were taken to increase the reliability of the assessment (e.g. use of a single observer throughout all observations) but there was no formal statistical assessment, and in ten papers, statistical analysis was undertaken. Level of validity was either explicitly stated or could be ascertained from the information provided, so this information is provided in Table 3; validity reached the construct or criterion level (Table 1) in 26 of the reviewed papers. Levels of reliability were not clear in all of the papers, but in 58% of the reviewed papers there was some form of reliability test detailed (Table 3).

The reviewed papers assessed welfare over a range of time periods, using a variety of methods of welfare assessment. Time scales ranged from observations on a small selection of days spread over months or years, to observations in a block of continuous days over a period of days or months. Approaches used to assess welfare included, but were not limited to, monitoring change over time, monitoring change following presumed stressful events, assessing differences between two or more situations, and comparison of the same measures with elephants in the wild. Resting, stereotypies, environmental interaction, feeding, social interactions, self-maintenance behaviours, activity (walking/locomotion) and GC levels all changed significantly when elephants were subject to different environmental or social circumstances. Situations which could be assumed to increase stress and therefore decrease welfare levels, such as transportation, novel flooring and being moved into a smaller enclosure were associated with decreased lying rest, increased standing rest, increased stereotypies, increased GC and decreased environmental interactions. Situations which may be

.

associated with improved welfare, such as being moved into pens/paddocks rather than being chained or shackled, and being provided with time-consuming, naturalistic feeding enrichment were associated with reduced stereotypies, increased feeding, increased positive social interactions, increased self-maintenance and increased activity.

Discussion

The aim of this review was to appraise evidence from current peer-reviewed literature on potential welfare indicators for captive elephants, and to synthesise evidence from the literature on the validity and reliability of these potential welfare indicators. An assessment of the peer-reviewed literature identified a selection of potential welfare indicators for which there was evidence of some level of validity. This included construct and criterion validity for the papers which studied behavioural and physiological indicators, and construct and face validity in the papers studying physical condition. The exact methods of recording each of the welfare indicators varied between studies and therefore any future use of welfare indicators should include assessment of the validity and reliability of the indicator in the context in which it is used (examples of validation processes can be found in Whitham & Wielebnowski, 2009 and Wemelsfelder & Mullan, 2014). A full narrative review of the welfare indicators is provided in supplementary material; however, it is worth briefly highlighting some of the strengths and limitations of the main welfare indicators identified.

Behavioural indicators

All of the reviewed studies which assessed behavioural indicators of welfare exhibited some degree of criterion validity by a change of state and a further five also exhibited construct validity through statistical association with other welfare indicators. Quantification of the frequency of observed stereotypical behaviour was the most frequently used measure of welfare in the captive elephant literature. Stereotypies are controversial as a welfare indicator because they may not be indicative of current welfare state. Not all stereotypies are sensitive indicators of current welfare state (Mason & Latham, 2004); the original factors which caused the stereotypy to develop may not

.

313 be present in their current environment, and thus a stereotypy may not be a measure of the current
314 welfare of the individual. However, changes in the level of expression of stereotypic behaviour may
315 still be useful as an indicator of welfare when the motivating reasons underlying the performance
316 of the stereotypy are known and when it is coupled with other measures. It has been suggested that
317 an increase in frequency or intensity of stereotypies may be indicative of a welfare issue, and
318 reduction in stereotypies not caused by direct prevention may be indicative of improved welfare
319 (Mason & Latham, 2004). The use of stereotypies as an indicator of welfare in the reviewed studies
320 suggests that this is true in these reports; there were meaningful correlations between changes in
321 levels of stereotypical behaviours and other welfare measures. For example, an increase in
322 stereotypies, a decrease in lying rest and an increase in faecal GCM, was observed in a bull elephant
323 post-transport (Laws *et al*., 2007). By contrast, a significant decrease in frequency of stereotypical
324 behaviour was observed when elephants were penned rather than chained in a circus (Gruber *et al*.,
325 2000; Schmid, 1995; Friend & Parker, 1999). Formal reliability assessments were reported in seven
326 of the studies and although intra-rater reliability was not assessed, a further two studies used a single
327 observer thereby removing the possibility of inter-rater variation. Used appropriately, i.e. alongside
328 other suitable measures of welfare and in a situation where there is the opportunity for investigation
329 of change over time, stereotypies appear to be an important and well-supported indicator of welfare.
330 Assessment of stereotypies would be particularly useful to assess an elephant's reaction to changes
331 in housing or husbandry practices; which could then be used to inform management decisions for
332 that elephant.

333 Although not yet formally validated as an indicator of welfare in elephants, sleep and rest
334 behaviour were linked to other welfare indicators in the reviewed papers, and changed in a
335 predictable manner in a number of different situations. Reliability assessments were conducted in
336 five of the ten papers which assessed sleep behaviour. Reduction in frequency of sleep was
337 correlated with increased stereotypies and associated with events perceived to be stressful to
338 elephants, such as travel (Laws *et al*., 2007), death of a conspecific (Koyama *et al*., 2012), and

.

339 introduction of novel flooring (Meller *et al*., 2007). Reduced sleep may be indicative of poor welfare
340 in some species, but particularly prolonged periods of time spent asleep may also be indicative of
341 stress (Jones *et al*., 2011, McPhee & Carlstead, 2012). The quality and pattern of sleep may be
342 important to the welfare of zoo housed elephants; however, relatively few studies have investigated
343 the resting behaviour of elephants housed in UK zoos (Williams *et al*., 2015; Holdgate *et al*., 2016b).
344 Elephant keepers and researchers have suggested that elephants lying down to sleep could be
345 interpreted as indicators of positive welfare, and a lack of sleep or not lying down to sleep could be
346 seen as indicators of negative welfare (Chadwick *et al*., 2017). Recent research has shown a
347 relationship between recumbence and substrate, space and social variables in elephants (Holdgate
348 *et al*., 2016b) and between some measures of physical health and recumbence (Yon *et al*.,
349 unpublished) but the complex relationship between rest and recumbence remains unclear. Further
350 research should be undertaken to investigate the factors which affect rest in captive elephants and
351 to investigate the relationship between rest and other welfare indicators, in order to identify if there
352 is an optimal level of lying rest for elephants. However, initial indications suggest that increased
353 lying rest, used in conjunction with other more fully validated measures, could be used as a
354 behavioural measure of welfare in zoo-housed elephants.

355 A relatively small number of authors researched social interactions in elephants, and social
356 interactions did not correlate with any other welfare measures. However, it is possible that because
357 social interactions were not the main focus of these studies, these less frequently performed
358 behaviours were missed, as these studies focused on compiling activity budgets pre- and post-
359 environmental change. Reliability assessments were undertaken in five of the six reviewed papers.
360 Elephants are a highly social species (Poole and Moss, 2008), and reports both in the peer-reviewed
361 literature and by stakeholders suggest that social interactions are an extremely important part of the
362 behavioural repertoire of an elephant. Indeed, in one of the reviewed papers, positive social
363 interactions were greater when elephants were given freedom of choice of social partners and were
364 kept in paddocks rather than being shackled (Schmid, 1995), which provided the possibility of more

.

interaction between elephants. It has been suggested by some elephant experts that persistent or extreme aggression within a captive group may be indicative of an underlying welfare problem for either a particular individual or for the entire group (Chadwick *et al*., 2017). Other aspects of group behaviour which have been studied in species other than elephants, such as behavioural synchrony (Asher & Collins, 2012), or the use of social networks (Asher *et al*., 2009), may also be useful welfare indicators. It is felt by stakeholders that social group size is one of the most important factors affecting elephant welfare (Gurusamy *et al*, 2014). Used in conjunction with other, validated indicators, expression of positive social interactions should be seen as a positive indicator of welfare.

Walking was widely assessed in the reviewed studies and correlations were observed between rest (negative), feeding (negative) and stereotypic pacing (positive). Five of the reviewed papers investigating walking behaviour formally assessed reliability and a further one used a single observer. Distance elephants travel in the wild has been attributed to availability and distribution of resources (Leighty *et al*., 2009); yet to date little is known about how far elephants 'should' walk in order to optimise welfare. I. This study found that elephants housed in larger enclosures and more complex social groups engaged in the greatest amount of walking behaviour (Leighty *et al*., 2009), which may be indicative of naturalistic exploratory behaviours. Distance walked has not been found to be related to health or behavioural outcomes, but distance walked has been found to be greater in groups with unpredictable feed schedules and greater number of elephants in the group (Holdgate *et al*., 2016a). Individual variability between elephants in walking behaviour within the same environment may be important; a lack of motivation to move, or a physical inability to move owing to poor physical health should be considered as a sign of poor welfare. Walking should be used as an indicator of welfare only alongside other, more traditional indicators, and it should also take into consideration the physical health of the individual elephant and the activities the elephants were engaged with whilst walking, rather than just distance travelled.

Environmental interactions did not significantly correlate with any other measure; however, increased environmental interaction was associated with positive social interactions and reduced

.

stereotypies. Four of the six reviewed studies assessing environmental interaction and welfare used formal reliability assessments, however there was disparity between studies in their interpretation and definition of environmental interaction. In order to understand its association with welfare and to increase the validity of this indicator, clear working definitions of activities which constitute environmental interactions must be developed, to enable precision in measuring these behaviours. Environmental interactions could then be used as part of a wider welfare assessment, and if observed in conjunction with other measures such as reduced stereotypies, environmental interactions may be seen as an indicator of an elephant engaging positively with its environment and therefore experiencing positive welfare.

Not all of the indicators identified in this review necessarily have the immediate potential for welfare assessment, but the presence of species-specific behaviour has been suggested as a potential indicator that the needs of the study animal are being met and that it is experiencing good health and well-being (McPhee & Carlstead, 2012). It could therefore be assumed that providing elephants with the opportunity to engage in increased periods of species-typical behaviour are positive for welfare, and that elephants that are engaging in this manner are experiencing good welfare. Species-typical behaviours which require further research before inclusion in welfare assessments for zoo-housed elephants include comfort or maintenance behaviours and feeding. Feeding behaviour correlated negatively with stereotypies and walking in the reviewed studies, however the factors underlying the relationships between these behaviours are not entirely clear. For example, it is not clear whether it is the lack of opportunity to feed that induces stereotypical behaviour in some elephants, nor is it clear whether the manner of food provision is reducing the distance elephants need to walk. Clubb and Mason (2002) suggested that lack of stimulation to engage in foraging activities is one of the main underlying causes of development of stereotypic behaviour. Researchers have suggested that increased food availability is associated with reduced exhibition of stereotypies (Friend & Parker, 1999), and when frequency of foraging is similar to that of wild elephants, relatively little stereotypic behaviour is seen (Koyama *et al*., 2012). Indeed,

.

keepers have also suggested that methods of food presentation which enable elephants to engage in more natural feeding behaviours are important for welfare (Chadwick *et al*., 2017).

Less recorded but nevertheless important behaviour which have been assessed included play and vocalisations. Further research is needed to investigate these indicators before they can be used reliably in welfare assessment. The small number of studies which have recorded play behaviour may represent the infrequency with which it is recorded in generalised activity budget studies (perhaps due to difficulty defining it), especially in adult elephants, whilst vocalisation data is inherently difficult to capture without specialised recording equipment and requires a good knowledge and understanding of the behavioural context for accurate interpretation of the data.

Physiological indicators of welfare

All of the eight reviewed papers on physiological indicators of welfare displayed some degree of either construct or criterion validity; one construct validity only, four criterion validity only and three both construct and criterion validity. Inter-assay reliability assessments were conducted for five of the seven papers. Levels of GC correlated positively with stereotypies and negatively with lying rest. Furthermore, they increased in situations which could be perceived as 'stressful', such as introduction of a new elephant (Dathe *et al*, 1992), the opening of the zoo (Menargues *et al*., 2008) and transport between facilities (Laws *et al*., 2007). Glucocorticoid measurements must be interpreted with caution as an indicator of welfare; GC are produced by the adrenal glands in response to activation of the hypothalamic-pituitary-adrenal (HPA) axis. However, activation of the HPA axis is context dependent and it may be activated during either beneficial or detrimental circumstances (Palme, 2012). Stress responses are an animal's means of coping with their environment (Palme, 2012) although it is widely understood that coping mechanisms differ between individuals, and it is not yet clear if there is an 'optimum' coping strategy (Fanson *et al*., 2013). Glucocorticoids are also affected by the sex, age, physiological stage, and life history of the animal as well as time of day and environmental factors such as temperature (see Mormede *et al*., 2007 for a full review). Assessment of GC should be used with appropriate consideration of these

.

caveats, measured over a suitable time period, with a suitable frequency and where possible and appropriate, at a range of time points throughout the day. They should be investigated in conjunction with a suite of other welfare measures to ensure a complete assessment of welfare.

Physical indicators of welfare

Three papers assessed body condition score of a large number of elephants. These papers only met face validity on our criteria of validity. However, the methods used to assess body condition were designed to increase the accuracy of ratings through thorough assessment, and thus are extremely important when considering the strength of these indicators. A assessment of overall physical condition was achieved through culmination of scores for a number of places on the body, using experienced observers and in the case of Wemmer and colleagues (2006) designing and trialling the questionnaire using multiple observers and providing pictures and descriptions to increase the accuracy of ratings.. Obesity in zoo elephants has been cited as a significant problem, and has been linked to poor foot health, arthritis and reduced reproductive output (Clubb *et al*., 2008; Clubb *et al*., 2009). Assessment of physical welfare using a body condition scoring protocol has the advantage of being relatively easy to learn and quick to conduct (Wemmer *et al*., 2006). Particularly in the captive setting, body condition scoring can be easily incorporated into routine health checks. To investigate the relationship between body condition score and measures of body fat, the method needs to be validated against composition assessments (Wemmer *et al*., 2006). However, as a simple means of reliably assessing the overall physical health of an elephant, body condition scores can be an important welfare indicator.

Only one paper included in this review investigated foot health, and that was studied in relation to stereotypies; the study met the threshold level of construct validity. Elephants with higher levels of stereotypies had poorer foot health, but owing to the high percentage of stereotypies observed in the study elephants the effect could not be identified as casual by the researchers (Haspeslagh, 2013). Nevertheless, assessment of foot health is an important physical health indicator as a stand-alone assessment; assessment of physical health, especially foot health, is being

.

increasingly incorporated into preventative care management approaches to keeping elephants in British and Irish zoos (Walter, 2010).

Physical indicators of welfare are more likely to change only over a longer time scale than behavioural or physiological indicators, making it more challenging to use health measures to assess short term responses to changing conditions. Furthermore, assessment of some physical welfare indicators, such as foot health, may require closer contact with the animal, so assessment would typically need to be undertaken by animal keeping or care staff, working directly with the elephants, rather than by visiting researchers. However, if undertaken by appropriately trained individuals over time, the methods described in the reviewed papers provide a reliable and valid means of assessing physical welfare of elephants.

Evaluation of the reviewed papers

Welfare assessment models, such as that developed by Sharp and Saunders (2011) utilise systematic, comprehensive and transparent processes to enable evidence-based assessments of animal welfare (Baker, Sharp & Macdonald, 2016). The finalised assessment must be developed from knowledge of behavioural and physiological responses to changes in circumstance or intervention (Baker, Sharp & Macdonald, 2016). Such a process was undertaken during this review to identify a suite of potential welfare indicators for use in routine welfare assessment of zoo housed elephants. As is evidenced in this review, there is a paucity of published literature assessing the welfare of captive elephants; however, there were a number of indicators which have been used repeatedly in the literature which could be used to assess welfare in captive elephants. There were some limitations to the reviewed studies such as relatively small sample sizes, number of single institution studies (73% of the reviewed papers were single institution studies) and time period of the studies (80% were conducted in a time period of less than one year). However these limitations are, in fact, common to zoo research and are not limited to the elephant literature; they arise due to the practical difficulties of conducting long-term, multi-institutional research. It is important not to

.

overlook the importance of numerous single-institution, short-term studies when reviewing the literature, especially when there is relatively little published research. The knowledge gained from these smaller studies could be maximised by using similar or standardised methods and surmising findings across studies. There is also likely to be an intrinsic link between animal based welfare indicators and environmental conditions (Gurusamy *et al*., 2014; Meehan *et al*., 2016), so consideration of environmental conditions should be incorporated into future studies with the aim of further validating the identified indicators.

Inclusion of more recently published material

Due to the process required to undertake a systematic review the findings are only current at the time the search was undertaken. Thus, literature published since the review date may be missed. In this instance, after the review was performed, the results of a large scale epidemiological study were released, and so it is prudent to include a short review of that work here, even though it was not a formal part of our review. The authors used eight welfare indicators; three behavioural (recumbence, daily walking distance and stereotypy), three physical (BCS, musculoskeletal health and foot health) and two physiological (ovarian acyclity and prolactin levels) (Meehan *et al*., 2016). Physical health indicators were associated with situations which may lead to poor welfare. In a study of 255 elephants, a link was established between foot and musculoskeletal health (as measured using presence/absence of abnormalities) and period of time spent on hard surfaces (Miller *et al*., 2016). High BCS were prevalent among the studied population of 240 elephants; nearly 75% were considered overweight or obese by the authors (Morfeld *et al*., 2016). There was no link between musculoskeletal and foot health and obesity (Miller *et al*., 2016) but the authors suggest that management practices which lead to reduced obesity may lead to welfare improvements (Morfeld *et al*., 2016). Behavioural measures which were investigated included walking rates and presence of stereotypies. In the 56 elephants studied, distance walked was not related to health or behavioural outcomes. However, walking rates were highest in elephants that had unpredictable feeding schedules and were housed in largest social groups, and were negatively correlated with overnight

.

space; with elephants having access to larger overnight spaces showing lower walking rates (Holdgate *et al*., 2016a). Stereotypical behaviour were the second most prevalent behaviour observed (after feeding) in the study population of 89 elephants (Greco *et al*., 2016). The social environment had a significant association with stereotypic behaviour rates: percent time with juveniles and number of elephant's housed together contributed to reduced risk of stereotypic behaviour, and being housed separately increased stereotypic risk. However, the authors recognised that there are multiple potential causes which contribute to the expression of stereotypic behaviour (Greco *et al*., 2016) and so these effects may not be causal. The final behavioural indicator investigated was recumbence behaviour. Holdgate and colleagues (2016b) studied 72 elephants for on average 4 to 5 days each. Species differences were observed between African and Asian elephants; African elephants were recumbent for on average 1 hour less per day than Asian elephants, and nearly 33% of the studied population were non recumbent for at least one night. An association was observed between substrate type and recumbence for both species, with resting occurring less frequently on hard flooring. In both species, recumbence was highest in elephants that had the greatest outdoor space overnight. Recumbence was also inversely related to age for both African and Asian elephants – with duration of sleep becoming shorter as elephants aged. Lone housed elephants slept longer than group housed elephants, which the authors attributed to a lack of disturbance (Holdgate *et al*., 2016b). These findings are in agreement with research by Yon and colleagues (unpublished), which identified less rest in groups with juveniles (due to them disturbing sleep). Yon and colleagues (unpublished) also identified a positive correlation between poor foot health/gait scores and duration of recumbence, and found that elephants with poorer physical health were recumbent for longer than average. This highlights the importance of lying rest, but also suggests there may be an optimal level of rest. The relationship between recumbence and welfare remains unclear but it is an important area for further investigation. The findings from this study, using a large number of study animals over a long period of time, contribute to our knowledge of

.

546 the impact of a number of husbandry factors on elephant welfare. The studies used indicators which
547 are widely used in the literature and support the findings from this rapid review.

548 ***Evaluation of the critical appraisal tool and methodological limitations***

549 To these authors' knowledge, the critical appraisal tool developed in this paper is the first
550 one developed to review animal welfare measures. This is a significant innovation and one which
551 could be applied to examine welfare measures in other contexts. However, the tool has limitations,
552 some of which relate to adjustments made to the data available. Perhaps the most significant is the
553 use of p-values instead of the more informative effect sizes (Nakagawa & Cuthill, 2007). When
554 attempting to extract effect sizes we found information provided in most papers did not permit
555 calculation of these values. Due to this lack of information, indicators were either classified as
556 either: (a) having a percentage change across different situations or (b) having a significant ($p<0.05$)
557 change across different situations. The questions asked in the critical appraisal tool did not place
558 values on different types of study design or different types of statistics. Other critical appraisal tools
559 consider certain study designs to provide stronger evidence than others, for example in
560 epidemiological studies, randomised control trials are viewed as the most robust (e.g. Sibbald and
561 Roland, 1998; Kaptchuk, 2001; GRADE Working Group, 2004). In future developments of this
562 tool, it would be useful to understand the value of different study designs in support of the validity
563 of welfare measures. There are some widely recognised limitations to research conducted on wild
564 animals at captive facilities (see Hosey *et al*. 2009). Typically in critical appraisal, inclusion criteria
565 would be sufficiently stringent that the review would consider only the best quality research with
566 the most appropriate study designs or sample sizes, but in this context it was not appropriate to be
567 this stringent. If critical appraisal approaches were applied to welfare measures in other species, for
568 which there are fewer limitations on study design and sample sizes, then more selective inclusion
569 criteria should be considered.

570 The approach used to identify relevant literature also had limitations. Systematic reviews are
571 current only at the point in time at which they are conducted, and thus cannot include work which

.

572 is published post search date. This review focused only on peer-reviewed literature which was
573 readily available from Scopus, Web of Knowledge and Ovid. Only papers for which the entire article
574 was available were included in the review. It is likely that more evidence exists in 'grey literature',
575 particularly the wealth of information available from within-zoo studies. Such literature is often too
576 inconsistently reported to allow for application of the critical appraisal tool and so was not in the
577 scope of this review, but a narrative review is provided in Asher and colleagues (2015). The papers
578 reviewed ranged widely from single elephant or single institution studies to multi-elephant or multi-
579 institution studies. They also ranged in terms of the level of validity demonstrated for the measures
580 in each study. Many of the reviewed studies did not claim to be 'assessing welfare'; however, if
581 they assessed behavioural change in situations which may be considered to be 'better for welfare'
582 or 'worse for welfare' they were included in the review. The ability to assess change over time is
583 important for an indicator of welfare. However, although we examined the duration of time over
584 which studies were conducted and the methods used to assess welfare, it was beyond the scope of
585 this review to define over what period of time each welfare indicator must be used in order to reliably
586 assess welfare. In future studies, it would be useful to expand the critical appraisal tool to consider
587 the time period over which welfare indicators were able to detect change. Finally, the biggest
588 constraint when identifying indicators of welfare in captive elephants is that some indicators were
589 more widely used and accepted than others, and these were repeated in the literature. The persistence
590 of the presence of these indicators in the literature doesn't necessarily indicate that they are the best
591 measures of welfare nor does it mean that they are more useful and should be deemed more
592 important in welfare assessment.

593 The indicators identified were largely in agreement with welfare indicators suggested by
594 keepers and elephant experts in focus groups (Chadwick *et al*., 2017), which may be due to the
595 familiarity of many people with the most common welfare measures. However, there were measures
596 which were mentioned in the focus groups which were not identified in this review of peer-reviewed
597 literature, such as assessment of skin, eyes, gait, and muscle tone. Used in combination, reviews of

.

598 existing literature and consultation with stakeholders could help to identify a range of welfare
599 measures to ensure a complete assessment of welfare for a given species. Indeed, Hill and Broom
600 (2009) suggested that a range of measures must be employed to ensure adequate assessment of
601 welfare in elephants.

602 **Conclusion**

603 Based on this rapid review and critical appraisal of peer-reviewed literature and assessment
604 of reliability and validity of the reported welfare measures, we suggest that there is support for the
605 following welfare indicators of improved welfare state: reduced stereotypies, reduced GC and
606 improved body condition scores. Additional measures which are yet to be fully validated but were
607 identified as having strong associations with the listed welfare measures and should therefore be
608 more thoroughly investigated (through inclusion in welfare assessments) are increased lying rest
609 and exhibition of positive social interactions. There is not enough evidence at present to include
610 increased environmental interactions and increased activity (or reduced inactivity) into welfare
611 assessment but they would be worthy of further investigation to establish their future use alongside
612 other, more well established and validated measures. It is important to note that many of these
613 measures represent a cumulative welfare state, rather than the current welfare state. Thus, a suite of
614 these measures should be employed as part of welfare assessment in elephants. Welfare assessments
615 should incorporate both well established and validated measures, and some of those measures
616 detailed in this report which have not yet been fully assessed or as frequently used, because of their
617 potential to capture important aspects of welfare. The use of these measures together would enable
618 the assessment of reliability and validity of the less frequently used measures for their use as future
619 welfare measures. Welfare assessments should be repeated within an individual for monitoring
620 purposes, both for routine monitoring over time as part of an ongoing assessment, and following
621 management or husbandry changes to assess a possible response to those changes. The evidence
622 synthesis and critical appraisal approach applied here to evaluate welfare measures could be usefully
623 applied to other contexts and species. The next stage in accurately identifying indicators of welfare

.

in captive elephants is the systematic assessment of the reliability and repeatability of the indicators detailed in this report across a range of conditions in captive elephants over time. This could be achieved through multi-institutional, longitudinal studies of a large number of elephants in a range of different conditions using a standard assessment criterion.

Animal welfare implications

Hill and Broom (2009) suggested that the most reliable results come from studies which adopt a multidisciplinary approach to assessing the welfare of animals, i.e. measuring a wide range of behavioural, physical or physiological indicators. In order to begin to efficiently assess the welfare of captive elephants, a suite of reliable and valid indicators of welfare must be identified. This paper makes the first steps towards identifying and reviewing welfare indicators used previously in the welfare assessment of zoo housed elephants, synthesising evidence on the reliability and validity of each indicator and identifying from these a selection of behavioural, physical and physiological indicators which could be used in future assessments of captive elephant welfare. This information should be used alongside consultation with zoo staff and other relevant stakeholders, in order to utilise existing knowledge and experience not contained with the scientific literature to identify further possible welfare measures. In this report, a range of different types of welfare indicators have been identified for potential use in assessing the welfare of captive elephants. As previously discussed, the further validation of these welfare indicators would enable the development of a more robust and comprehensive tool for determining captive elephant welfare.

Appendix: supplementary material

Supplementary data to this article can be found online at…

Acknowledgements

This research was conducted as part of a study funded by the Department for Environment, Food and Rural Affairs (DEFRA). We are grateful to members of the Elephant Welfare Group (EWG) Behaviour Sub-group, Dr Samantha Bremner-Harrison, Dr Oliver Burman and Dr Ros Clubb for advice. We are thankful to Douglas Grindlay for his guidance with the review process. We are

.

650 appreciative of the invaluable guidance and other support from the policy and evidence team at
651 Defra, and their external experts, Matt Hartley and Dr John Eddison.

652 **References**

653 **Asher L and Collins L** 2012 Assessing synchrony in groups: Are you measuring what you think
654 you are measuring? *Applied Animal Behaviour Science* 138 (3-4): 162 – 169

655 **Asher L, Collins LM, Ortiz-Pelaez A, Drewe JA, Nicol CJ and Pfeiffer DU** 2009 'Recent
656 advances in the analysis of behavioural organization and interpretation as indicators of animal
657 welfare', *Journal of The Royal Society Interface* 6(41): 1103–1119

658 **Asher L, Williams E and Yon L** 2015 *Developing behavioural indicators, as part of a wider set*
659 *of indicators, to assess the welfare of elephants in UK zoos. Defra WC1081*. Bristol: Defra

660 **Baker SE, Sharp TM, Macdonald DW** 2016 Assessing Animal Welfare Impacts in the
661 Management of European Rabbits (*Oryctolagus cuniculus*), European Moles (*Talpa europaea*)
662 and Carrion Crows (*Corvus corone*). *PloS One* 11(1): e0146298

663 **Belshaw Z, Asher L, Harvey ND and Dean RS** 2015 Quality of life assessment in domestic
664 dogs: An evidence-based rapid review. *The Veterinary Journal* 206(2): 203–212

665 **Born Free** 2015 *Elephants in UK and European Zoos*. Available at:
666 http://www.bornfree.org.uk/campaigns/zoo-check/captive-wildlife-issues/elephants-in-
667 captivity/elephants-in-uk-and-european-zoos/ (Accessed: 14 July 2015)

668 **Chadwick CL, Williams E, Asher L and Yon L** 2017 Stakeholders' opinions on assessing the
669 welfare of elephants (Loxodonta africana and Elephas maximus) housed in UK zoos. *Animal*
670 *Welfare*26: 461 - 472

671 **Clubb R and Mason G** 2002 *A review of the welfare of zoo elephants in Europe*. Horsham, UK:
672 RSPCA

673 **Clubb R, Rowcliffe M, Lee P, Mar KU, Moss C and Mason GJ** 2008 Compromised
674 Survivorship in Zoo Elephants. *Science* 322(5908): 1649–1649.

.

675 **Clubb R, Rowcliffe M, Lee P, Mar KU, Moss C and Mason GJ** 2009 Fecundity and population
676 viability in female zoo elephants: problems and possible solutions. *Animal Welfare* 18: 237 – 247.
677 **Dathe HH, Kuckelkorn B and Minnemann D** 1992 Salivary cortisol assessment for stress detection
678 in the Asian elephant (Elephas maximus): A pilot study. *Zoo Biology* 11(4): 285–289.
679 **Dawkins MS** 2008 What is good welfare and how can we achieve it? In Dawkins MS and Bonney
680 R (eds) *Future of animal farming: renewing the ancient contract*. Malden, MA: Blackwell Pub.
681 **Elzanowski A and Sergiel A** 2006 Stereotypic Behavior of a Female Asiatic Elephant (Elephas
682 maximus) in a Zoo. *Journal of applied animal welfare science* 9(3): 223 – 232.
683 **Fanson KV, Lynch M, Vogelnest L, Miller G and Keeley** T 2013 Response to long-distance
684 relocation in Asian elephants (Elephas maximus): monitoring adrenocortical activity via serum,
685 urine, and feces. *European Journal of Wildlife Research* 59(5): 655 - 664.
686 **Friend TH** 1999 Behavior of picketed circus elephants, *Applied Animal Behaviour Science* 62(1):
687 73 – 88.
688 **Friend TH and Parker ML** 1999 The effect of penning versus picketing on stereotypic behavior
689 of circus elephants, *Applied Animal Behaviour Science* 64(3): 213–225.
690 **Godogama WK, Wemmer C and Ratnasooriya WD** 1998 The body condition of Sri Lankan
691 domesticaded elephants (Elephas maximus maximus). *Ceylon Journal of Science (Biological Sciences)*
692 26: 1 – 6.
693 **Grand AP, Kuhar CW, Leighty KA, Bettinger TL and Laudenslager ML** 2012 Using personality
694 ratings and cortisol to characterize individual differences in African Elephants (Loxodonta africana).
695 *Applied Animal Behaviour Science* 142(1-2): 69–75.
696 **GRADE Working Group** 2004 Grading quality of evidence and strength of
697 recommendations. *BMJ* 328(7454): 1490–0
698 **Greco BJ, Meehan CL, Hogan JN, Leighty KA, Mellen J, Mason GJ and Mench JA** 2016
699 The days and nights of zoo elephants: using epidemiology to better understand stereotypic

.

700 behaviour of African elephants (*Loxodonta africana*) and Asian elephants (*Elephas maximus*) in
701 North American Zoos. *PLoS One* 11(7): e0144276
702 **Gruber TM, Friend TH, Gardner JM, Packard JM., Beaver B and Bushong D** 2000 Variation in
703 stereotypic behavior related to restraint in circus elephants. *Zoo Biology* 19(3): 209–221.
704 **Gurusamy V, Tribe A and Phillips CJC** 2014 Identification of major welfare issues for captive
705 elephant husbandry by stakeholders. *Animal Welfare* 23: 11 – 24
706 **Harker JH and Kleijnen J** 2012 What is a rapid review? A methodological exploration of rapid
707 reviews in Health Technology Assessments. *International Journal of Evidence-Based Healthcare*
708 10(4): 397 - 410
709 **Harris M, Sherwin C and Harris S** 2008 *The welfare, housing and husbandry of elephants in*
710 *UK zoos. Defra WC05007*. Bristol: Defra
711 **Hill SP and Broom DM** 2009 Measuring zoo animal welfare: theory and practice. *Zoo Biology*
712 28(6): 531 – 544.
713 **Haspeslagh M, Stevens J, De Groot E, Dewulf J, Kalmar I and Moons C** 2013 A survey of foot
714 problems, stereotypic behaviour and floor type in Asian elephants (Elephas maximus) in European
715 zoos. *Animal Welfare* 22(4): 437–443.
716 **Hnath PT and Yannessa MN** 2002 Effects of facility modifications on elephant activity levels.
717 *Animal Keepers Forum* 29(10): 421 – 427.
718 **Hosey GR, Melfi V and Pankhurst S** 2009 *Zoo animals: Behaviour, management and welfare*
719 [chapter 14]. New York: Oxford University Press.
720 **Holdgate MR, Meehan CL, Hogan JN, Miller LJ, Soltis J, Andrews J and Shepherdson DJ**
721 2016a Walking behaviour of zoo elephants: associations between GPS measured daily walking
722 distances and environmental factors, social factors and welfare indicators. *PLoS One* 11(7):
723 e0150331
724 **Holdgate MR, Meehan CL, Hogan JN, Miller LJ, Rushen J, de Passille AM, Soltis J,**
725 **Andrews J and Shepherdson DJ** 2016b Recumbence behaviour in zoo elephants: determination

.

of patters and frequency of recumbent rest and associated environmental and social factors. *PLoS One* 11(7): e0153301

Jones MA, Mason GJ and Pillay N 2011 Correlates of birth origin effects on the development of stereotypic behaviour in striped mice, Rhabdomys *Animal Behaviour* 82(1): 149 – 159

Kiiru W 2007 *The sad state of captive elephants in Canada*. Zoo check Canada.

Kaptchuk TJ 2001 The double-blind, randomized, placebo-controlled trial: Gold standard or golden calf? *Journal of Clinical Epidemiology* 54(6): 541–549

Koyama N, Ueno Y, Eguchi Y, Uetake K and Tanaka T 2012 Effects of daily management changes on behavioral patterns of a solitary female African elephant (Loxodonta africana) in a zoo. *Animal Science Journal* 83(7): 562–570.

Laws N, Ganswindt A, Heistermann M, Harris M, Harris S and Sherwin C 2007 A Case Study: Fecal Corticosteroid and Behavior as Indicators of Welfare During Relocation of an Asian Elephant. *Journal of Applied Animal Welfare Science* 10(4): 349–358.

Lewis KD, Shepherdson DJ, Owens TM and Keele M 2010 A survey of elephant husbandry and foot health in North American Zoos. *Zoo Biology* 29(2): 221-236

Leighty KA, Soltis J, Wesolek CM, Savage A, Mellen J and Lehnhardt J 2009 GPS determination of walking rates in captive African elephants (Loxodonta africana). *Zoo Biology* 28(1): 16–28.

Mason GJ 1991 Stereotypies: a critical review. *Animal Behaviour* 41(6):1015 – 1037.

Mason GJ 2010 Species differences in responses to captivity: stress, welfare and the comparative method. *Trends in Ecology & Evolution* 25(12): 713–721.

Mason GJ and Latham NR 2004 Can't stop, won't stop: is stereotypy a reliable animal welfare indicator? *Animal Welfare* 13: S57 – 69.

Mason GJ and Veasey JS 2010 How should the psychological well-being of zoo elephants be objectively investigated? *Zoo Biology* 29(2): 237–255.

.

McPhee ME and Carlstead K 2012 The Importance of Maintaining Natural Behaviors in Captive Mammals, in Kleiman DG, Thompson KV and Kirk Baer C (eds) *Wild Mammals in Captivity: principles and techniques for zoo management, second edition*. Chicago, USA: University of Chicago Press

Meagher RK 2009 Observer ratings: Validity and value as a tool for animal welfare research. *Applied Animal Behaviour Science* 119(1-2): 1–14.

Meehan CL, Mench JA, Carlstead K and Hogan JN 2016 Determining connections between the daily lives of zoo elephants and their welfare: an epidemiological approach. *PLoS One* 11(7): e0158124

Meller CL, Croney C and Shepherdson D 2007 Effects of rubberized flooring on Asian elephant behavior in captivity. *Zoo Biology* 26(1): 51–61.

Menargues A, Urios V and Mauri M 2008 Welfare assessment of captive Asian elephants (Elephas maximus) and Indian rhinoceros (Rhinoceros unicornis) using salivary cortisol measurement. *Animal Welfare* 17: 305 – 312.

Miller MA, Hogan JN and Meehan CL 2016 Housing and demographic risk factors impacting foot and musculoskeletal health in African elephants (Loxodonta africana) and Asian elephants (Elephas maximus) in North American zoos *PLoS One* 11(7): e0155223

Millspaugh JJ, Burke T, Van Dyk G, Slotow R, Washburn BE and Woods RJ 2007 Stress Response of Working African Elephants to Transportation and Safari Adventures. *Journal of Wildlife Management* 71(4): 1257–1260.

Millspaugh JJ, Washburn BE 2004 Use of fecal glucocorticoid metabolite measures in conservation biology research: considerations for application and interpretation. *General and Comparative Endocrinology* 138: 189 – 199

Morfeld KA, Meehan CL, Hogan JN and Brown JL 2016 Assessment of body condition in African (Loxodonta africana) and Asian (*Elephas maximus*) elephants in North American zoos and management practices associated with high body condition scores. *PLoS One* 11(7): e0155146

.

Mormede P, Andanson S, Auperin B, Beerda B, Guemene D, Malmkvist J, Manteca X, Manteuffel G, Prunet P, van Reenen CG, Richard S and Veissier I. 2007 Exploration of the hypothalamic-pituitary-adrenal function as a tool to evaluate animal welfare. *Physiology and behaviour* 92(3): 317 – 339

Nakagawa S and Cuthill IC 2007 Effect size, confidence interval and statistical significance: A practical guide for biologists. *Biological Reviews* 82(4): 591–605

Palme R 2012 Monitoring stress hormone metabolites as a useful, non-invasive tool for welfare assessment in farm animals. *Animal Welfare* 21(3): 331–337

PETA 2015 *Get Elephants Out of Zoos*. Available at: http://www.peta.org/issues/animals-in-entertainment/get-elephants-out-of-zoos.aspx (Accessed: 14 July 2015)

Poole J and Moss C 2008 Elephant sociality and complexity, in Wemmer C and Christen CA (eds) *Elephants and ethics: Toward a morality of coexistence*. Baltimore, MD: Johns Hopkins University Press.

Posta B, Huber R and Moore III DE 2013 The Effects of Housing on Zoo Elephant Behavior: A Quantitative Case Study of Diurnal and Seasonal Variation. *International Journal of Comparative Psychology* 26(1): 37 – 52.

Ralph CR, Tilbrook AJ 2016 INVITED REVIEW: The usefulness of measuring glucocorticoids for assessing animal welfare. *Journal of Animal Science* 94(2): 457-470

Ramanthan A and Mallapur A 2008 A visual health assessment of captive Asian elephants (Elephas maximus) housed in India. *Journal of zoo and wildlife medicine* 39(2): 148 - 154

Rees PA 2004 Low environmental temperature causes an increase in stereotypic behaviour in captive Asian elephants (Elephas maximus) *Journal of Thermal Biology* 29(1): 37 - 43

Rees PA 2009 Activity budgets and the relationship between feeding and stereotypic behaviors in Asian elephants (Elephas maximus) in a Zoo. *Zoo Biology* 28(2): 79–97.

Schmid J 1995 Keeping circus elephants temporarily in paddocks – the effects on their behaviour. *Animal Welfare* 4: 87 – 101.

.

Schmid J, Heistermann M, Ganslober U and Hodges JK 2001 Introduction of foreign female Asian elephants (Elephas maximus) into an existing group: behavioural reactions and changes in cortisol levels. *Animal Welfare* 10: 357 – 372.

Sharp T and Saunders G 2011 A model for assessing the relative humaness of pest animal control methods (Second edition). Australian Government Department of Agriculture, Fisheries and Forestry, Canberra, ACT.

Sibbald B and Roland M 1998 B Understanding controlled trials: Why are randomised controlled trials important? *BMJ* 316(7126): 201

Soltis J 2010 Vocal communication in African elephants (Loxodonta africana). *Zoo Biology* 29(2): 192 - 209

Stoinski TS, Daniel E and Maple TL 2000 A preliminary study of the behavioral effects of feeding enrichment on African elephants. *Zoo Biology* 19(6): 485–493.

Vanitha V, Thiyagesan K and Baskaran N 2011 Social life of captive Asian elephants (Elephas maximus) in Southern India: implications for elephant welfare. *Journal of Applied Animal Welfare Science* 14(1): 42 - 58

Veasey J 2006 Concepts in the care and welfare of captive elephants. *International Zoo Yearbook* 40(1): 63–79.

Walter O 2010 Management guidelines for the welfare of zoo animals: elephants. 3rd edition. UK: BIAZA

Wells DL and Irwin RM 2008 Auditory stimulation as enrichment for zoo-housed Asian elephants (Elephas maximus) *Animal Welfare* 17(4): 335 - 340

Wemelsfelder F and Mullan S 2014 Applying ethological and health indicators to practical animal welfare assessment. *Revue Scientifique et Technique de l'OIE* 33(1):111–120

Wemmer C, Krishnamurthy V, Shrestha S, Hayek LA, Thant M. and Nanjappa KA 2006 Assessment of body condition in Asian elephants (Elephas maximus) *Zoo Biology* 25(3) 187–200

.

Whilde J and Marples N 2011 Effect of a birth on the behavior of a family group of Asian elephants (Elephas maximus) at Dublin Zoo. *Zoo Biology* 31(4): 442–452.

Whitham JC and Wielebnowski N 2009 Animal-based welfare monitoring: Using keeper ratings as an assessment tool. *Zoo Biology* 28(6): 545 – 560

Wilson M, Bloomsmith M and Maple T 2004 Stereotypic swaying and serum cortisol concentrations in three captive African elephants (*Loxodonta africana*). *Animal Welfare* 13：39 - 43

Williams E, Bremner-Harrison S, Harvey N, Evison E and Yon L 2015 An investigation into resting behaviour in Asian elephants in UK zoos. *Journal of Zoo Biology* 34: 406 - 417

Zoos Forum Elephant Working Group 2010 *Elephants in UK Zoos: Zoos Forum review of issues in elephant husbandry in UK zoos in light of the Report by Harris et al (2008)*. Bristol: Wildlife Species Conservation Division, Defra

.

854 **Table 1. Definitions used during assessment of reliability and validity of the study methods**
855 **(based on Meagher, 2009 and Belshaw et al., 2015)**

Type of reliability or validity	Definition
Intra-rater reliability	The consistency of recording within the same rater repeatedly scoring the same animal within a very short timeframe or ideally the same observation of the animal (as recorded by a video camera).
Inter-rater reliability	The consistency of recording between rates scoring the same animal at the same time or using the same observation of the animal (as recorded by a video camera).
Test re-test reliability	The consistency of answers when scored within the same animal expected to be in the same welfare state after a minimum time interval of two days.
Internal reliability	The correlation between items within components of an instrument which are meant to be measuring the same thing. Usually measured with Cronbach's alpha. Allows removal of poor and redundant items during instrument development. Closely related to construct validity.
Content/face validity	Whether the items in an instrument appear to be asking what they should be. Logical explanation as to why measure is representative of an aspect of welfare provided in Introduction or Discussion.
Construct validity	Whether the items in an instrument measure the broad area (construct), which they were designed to measure (e.g. comfort). Assessed by investigating correlations between similar (convergent) and dissimilar (divergent) welfare measures. These may be other behavioural measures or physiological measures or a combination.
Criterion validity (concurrent or predictive)	The results of the instrument are compared to an external, independent criterion measure. The criterion measure is thought to measure the same thing and should ideally be a "gold standard" test, or an alternative established measure. A gold standard measure of welfare could be considered animal choice or strength of motivation either positive or negative; or validated measures of affective state (e.g. cognitive bias); or (depending on welfare definition) comparison with natural or functional behaviour. The criterion measurement is taken from the same animal, and can be at the same time (concurrent to the assessment) or in the future (where the assessment is predictive of the criterion measure). Additional measures of criterion validity would be the ability of the instrument to distinguish between different populations (e.g. attempt to manipulate welfare e.g. provide enriched or impoverished environments and test changes in welfare measures; or compare environments which are believed or previously been shown to have better and worse welfare; or expose to short term welfare intervention).

856

.

857 **Table 2.Behavioural indicators of welfare identified in the peer reviewed literature which have been used in assessment of captive elephant**
858 **welfare**

Type of indicator	Indicator category	Example of indicators used	Mean (range, SD) of the number of elephants studied	Correlation with other measures	Authors which used the indicator*	Significant change	Percent change
Behavioural	Abnormal behaviour	Stereotypies, foot lifting, faeces manipulation, trunk swinging	16 (1 – 140, 32)	Feeding Walking Resting Foot health Cortisol	Koyama et al (2010) Laws et al (2007) Gruber et al (2000) Vanitha et al (2011) Elzanowski & Sergiel (2006) Friend (1999) Hnath & Yannessa (2002) Schmid (1995) Schmid et al (2001) Wells & Irwin (2008) Wilson et al (2004) Rees (2004) Friend & Parker (1999) Meller et al (2007) Stoinski et al (2000) Whilde & Marples (2011) Rees (2009)	9 papers	17 papers
	Sleep/rest	Standing rest, lying rest	7 (1 – 17, 6)	Walking Stereotypies	Koyama et al (2010) Laws et al (2007) Gruber et al (2000) Friend (1999) Hnath & Yannessa (2002) Schmid et al (2001) Posta et al (2013) Friend & Parker (1999) Meller et al (2007) Whilde & Marples (2011)	3 papers	3 papers

.

Feeding	Eating, drinking, ingestion	6 (1 – 17, 5)	Walking Stereotypies	Koyama et al (2010) Gruber et al (2000) Friend (1999) Hnath & Yannessa (2002) Schmid et al (2001) Posta et al (2013) Stoinski et al (2000) Whilde & Marples (2011) Rees (2009) Wells & Irwin (2008)	3 papers	10 papers
Environmental interaction	Enrichment use, investigative/ exploratory behaviour	4 (2 – 7, 2)		Posta et al (2013) Meller et al (2007) Stoinski et al (2000) Whilde & Marples (2011) Hnath & Yannessa (2002) Schmid et al (2001)	3 papers	6 papers
Comfort (self-maintenance)	Dust bathing, mud wallowing, general grooming	9 (1 – 29, 8)		Koyama et al (2010) Gruber et al (2000) Friend (1999) Hnath & Yannessa (2002) Schmid (1995) Schmid et al (2001) Wells & Irwin (2008) Friend & Parker (1999) Stoinski et al (2000) Whilde & Marples (2011) Rees (2009) Posta et al (2013)	4 papers	12 papers
Activity	Walking/locomotion	6 (1 – 14, 4)	Rest Feeding Stereotypies	Posta et al (2013) Koyama et al (2010) Gruber et al (2000) Schmid et al (2001) Wells & Irwin (2008)	5 papers	10 papers

.

					Stoinski et al (2000) Meller et al (2007) Rees (2009) Whilde & Marples (2011) Leighty et al (2009)		
	Inactive		3		Stoinski et al (2000)	1 paper	1 paper
	Social interactions	Positive interactions (affiliation), negative interactions (agression)	10 (2 - 29, 9)		Gruber et al (2000) Schmid (1995) Schmid et al (2001) Wells & Irwin (2008) Posta et al (2013) Stoinski et al (2000)	3 papers	5 papers
	Other	Vocalisations	4 (4 – 4, 0)		Wells & Irwin (2008) Soltis (2010)	1 papers	1 papers
		Play	17 (4 – 29, 13)		Whilde & Marples (2011) Schmid (1995)	1 papers	2 papers
Physical	Assessment of body condition/health (except feet)	Body condition score, assessment of mucous membranes, skin condition, eyesight, oedemas, wounds, abscesses	114 (82 – 140, 24)		Ramanthan & Mallapur (2008) Godogama et al (1998) Wemmer et al (2006)	NA	NA
	Assessment of foot health	Toenail cracks, presence of foot fissures, abscesses	87	Stereotypies	Haspeslagh et al (2013)	NA	NA
Physiological	Cortisol analysis	Salivary cortisol, serum cortisol, faecal glucometabolites, urinary glucometabolites	5 (1 – 8, 2)	Other measures of cortisol Personality Stereotypies	Dathe et al (1992) Fanson et al (2013) Grand et al (2012) Laws et al (2007) Menargues et al (2008) Millspaugh et al (2007) Schmid et al (2001)	8 papers	10 papers

.

859
860 **Table 3. Summary of the 30 articles reviewed**

Authors	Species	Sample size	No. of institutions	Validity	Reliability	Time period	Method of assessing welfare	Study design	Welfare measures
Dathe et al (1992)	EM	2 (0.2)	1	Criterion	Intra and inter assay coefficients of variation	15 - 18 days	Comparison with normal cortisol concentration range	Observational, repeated measures, qualitative	Salivary cortisol[%]
Elzanowski & Sergiel (2006)	EM	1 (0.1)	1	Criterion	None reported	35 days in 1 year	Monitoring behavioural changes following changes to the environment	Experimental, repeated measures, quantitative	Stereotypies[S]
Fanson et al (2013)	EM	8 (1.7)	3	Criterion	None reported	1 year	Compared observed results with expected results, monitored change in cortisol over time	Observational, repeated measures, quantitative and qualitative assessment of personality	Faecal gluco metabolites (FGM)[SC] UGM[SC] Serum Cortisol[SC] Keeper Assessment of Personality[C]
Friend (1999)	EM/LA		1	Criterion	None reported	October 1995 (4 x 8hrs) and	Monitoring behavioural changes prior		Stereotypies[%] Resting[%] Eating/drinking[%]

.

Authors	Species	Sample size	No. of institutions	Validity	Reliability	Time period	Method of assessing welfare	Study design	Welfare measures
		14 EM (0.14), 3 LA (0.3)				April 1996 (3 x 24hrs)	to scheduled events		Dust bathing[%]
Friend & Parker (1999)	EM/LA	14 EM (0.14), 3 LA (0.3)	1	Criterion	None reported	April 1996 and April 1998 (3 x 24hrs during each period)	Monitoring behavioural changes following changes to the environment	Experimental, repeated measures, quantitative	Stereotypies[S] Standing Lying Eating Drinking Dust bathing
Godogama et al (1998)	EMM	140 (72.68)	13 districts		None reported	N/A		Qualitative - BCS, independent	Body condition score
Grand et al (2012)	LA	5 (0.5)	1	Construct	Inter-rater	One month	Correlations predicted between types of cortisol measures and between cortisol and personality characteristics	Observational, repeated measures, qualitative	Keeper assessment of personality[SC] Salivary cortisol[SC] Serum cortisol[SC]
Gruber et al (2000)	EM/LA	EM 11 (1.10) LA 3 (0.3)	1	Criterion	Intra rater and inter-rater	2 months per treatment group	Monitoring behavioural changes between treatment groups	Experimental, repeated measures, quantitative	Stereotypies[S] Aggression Comfort[S] Ingestion[S] Locomotion[S] Resting Social[S]

.

Authors	Species	Sample size	No. of institutions	Validity	Reliability	Time period	Method of assessing welfare	Study design	Welfare measures
Haspeslagh et al (2013)	EM	87 (16.71)	32	Construct	None reported	N/A	correlation predicted between behavioural and physical measures of welfare		Stereotypies[C] Foot health[C]
Hnath & Yannessa (2002)	EM/LA	2 (0.2)	1	Criterion	None reported	3 days per week for 2 weeks, then 4 month break (following environmental change) then 3 days per week for 2 weeks	Monitoring behavioural changes following changes to the environment	Observational, repeated measures, quantitative	Keeper/elephant interaction[%] Feeding[%] Enrichment use[%] Yard investigation[%] Dust bathing[%] Resting[%] Stereotypies[%]
Koyama et al (2012)	LA	1 (0.1)	1	Construct and Criterion	None reported	1 year	Monitoring change in behaviour over time, following presumed stressful event	Prospective, observational, repeated measures, quantitative	Feeding[C%] Comfort[%] Locomotion[C%] Resting[C%] Stereotypies[C%]
Laws et al (2007)	EM	1 (1.0)	2	Construct and Criterion	Intra and inter- assay coefficients of variation	20 days (10 days prior to event and 10	Monitoring change in behaviour and cortisol following	Prospective, observational, repeated measures, quantitative	Stereotypies[%] Sleep[%] Faecal cortisol[S]

.

Authors	Species	Sample size	No. of institutions	Validity	Reliability	Time period	Method of assessing welfare	Study design	Welfare measures
						days post event)	presumed stressful event		
Leighty et al (2009)	LA	7 (0.7)	1	Criterion	None reported	1 year	Monitoring behavioural changes in different scenarios	Experimental, repeated measures, quantitative	Locomotion[S]
Lewis et al (2010)	EM/LA	EM 137 (26.111) , LA 151 (21.130)	78		None reported	N/A			Presence of foot pathologies
Meller et al (2007)	EM	6 (2.4)	1	Criterion	Inter-observer	3 days per observation period (3 periods)	Compared with choices and then monitored overall behavioural change following environmental manipulation	Experimental, repeated measures, quantitative	Locomotion[S] Standing rest[S] Lying rest[S] Foot-lifting Exploratory[S] Stereotypies[S]
Menargues et al (2008)	EM	6 (0.6)	1	Criterion	None reported	4 months	Comparison with normal cortisol concentration range	Observational, repeated measures, quantitative	Salivary cortisol[S]

.

Authors	Species	Sample size	No. of institutions	Validity	Reliability	Time period	Method of assessing welfare	Study design	Welfare measures
Millspaugh et al (2007)	LA	5	1	Criterion	standard assay validation	1 year	Monitoring change over time, comparison with wild	Prospective, Observational, repeated measures, quantitative	FGM[S]
Posta et al (2013)	LA	2 (1.1)	1	Criterion	Inter-observer	2 years	Comparison with wild	Experimental, repeated measures, quantitative	Feed[%] Nurse[%] Stand[%] Lie[%] Walk[%] Enrichment Use[%] Self-directed[%] Investigation[%] Affiliation[%] Aggression[%]
Ramanthan & Mallapur (2008)	EM	82 (33.49)	10		None reported – measures taken to increase reliability	N/A		Qualitative - BCS, independent	Mucous membrane Body condition score Skin condition Foot fissures Toenail cracks Edema Eyesight Wounds Abscess
Rees (2004)	EM	8 (2.6)	1	Criterion	None reported – measures taken to	35 days	Monitoring how stereotypies change over time	Observational, repeated measures, quantitative	Stereotypies[S]

.

Authors	Species	Sample size	No. of institutions	Validity	Reliability	Time period	Method of assessing welfare	Study design	Welfare measures
					increase reliability				
Rees (2009)	EM	8 (2.6)	1	construct and Criterion	None reported – measures taken to increase reliability	35 days	Monitoring how activity budgets change over time	Repeated measures, observational, quantitative	Dust bathing[S] Feeding[C] Locomotion Standing Stereotypies[SC]
Schmid (1995)	EM/LA	EM 19 (0.19), LA 10 (0.10)	4	Criterion	None reported	4 to 11 days	comparison of species typical behaviours between keeping systems	Experimental, repeated measures, quantitative	Social – attractive Social – cohesive[S] Social – repulsive Comfort[S] Object play[S] Stereotypies[S]
Schmid et al (2001)	EM	7 (0.7)	1	construct and Criterion	intra and inter assay coefficients	7 months	Correlation between behavioural and physiological indicators of welfare, looking at changes following presumably stressful event	Experimental, repeated measures, quantitative	Stereotypies Social Comfort Feeding Locomotion Resting Manipulation/exploration Cortisol
Soltis (2010)	LA	4 (4.0)	1	Criterion	None reported	14.5 months	Documenting elephant rumbles in	Observational, repeated measures	Vocalisation[S]

.

Authors	Species	Sample size	No. of institutions	Validity	Reliability	Time period	Method of assessing welfare	Study design	Welfare measures
							different social situations		
Stoinski et al (2000)	LA	3 (0.3)	1	Criterion	Inter-rater	1 month	Monitoring change over time	Experimental, repeated measures, quantitative	Feed[S] Drink[S] Object exam Faeces manipulation Locomotion Social Sterotypies Mud wallowing Self-directed Inactive[S] Contact[S]
Vanitha et al. (2011)	EM	140	80	Criterion	None reported	2 years		Retrospective, observational, independent measures, objective questionnaire	Stereotpyies[S]
Wells and Irwin (2008)	EM	4 (0.4)	1	Criterion	Test re-test	Initial study over 21 days, study repeated 4 months later for 3 days	Environmental manipulation	Experimental, repeated measures, quantitative	Stand Move Socialise Aggression Dust bathe Object interaction Eat Drink Vocalise Abnormal behav[S]

.

Authors	Species	Sample size	No. of institutions	Validity	Reliability	Time period	Method of assessing welfare	Study design	Welfare measures
Wemmer et al. (2006)	EM	119 (58.61)	7		None reported – measures taken to increase reliability	Single point		Qualitative - BCS, independent	Body Condition Scoring
Whilde and Marples (2011)	EM	4 (0.4)	1	Criterion	None reported – measures taken to increase reliability	10 days prior to event, 2 months post event	Monitoring behavioural changes following an event	Experimental, repeated measures, quantitative	Walk[S] Feed Stand[S] Stereotypy Trunk swing Maintenance Manipulation of non-food items[S] Play Lie Associations[S]
Wilson et al. (2004)	LA	3 (0.3)	1	construct and Criterion	Inter-rater	10 weeks during 2001	Comparison of behavioural changes at two points in time (1992, 1994, 2001)	Repeated measures, observational, quantitative	Blood cortisol[%C] Stereotypies[%C]

861 [S] Measure identified as being statistically significant in the study, [C] Measure correlated with another welfare measure in the study, [%] Percentage
862 change in the study
863
864

865

.

Section 1. INFORMATION ON STUDY POPULATION

1. Species [African (Loxodonta Africana); Asian (Elephas maximus)]
2. Sub-species [L. africana africana; L. africana cyclotis; E. maximus maximus; E. maximus indicus; E. maximus sumatranus]
3. Sample size (males: females)
4. Age range
5. Type of facility or facilities studied [Zoo; Safari park; Circus; Timber Camp; Other]
6. Number of establishments involved in the study
7. Number of enclosures or groups
8. Approximate size(s) of enclosure

Section 2. STUDY DESIGN

9. Study design [*as many as applicable from:* Observational- qualitative ; Observational-quantitative; Retrospective; Prospective; Experimental ; Repeated-measures design; Independent-measures design]
10. Number of repeated measures of same animal
11. Control group used [Yes, No]
12. Study manipulations [Yes, No]
13. Rater blind to study manipulations? [Yes, No]

Section 3. WELFARE INDICATOR METHODS

14. Welfare indicators used *list then complete the remainder of relevant questions for each welfare indicator*
15. Media for data collection [Live observations; Video observations; Proxy assessor (e.g keeper questionnaire); Records]
16. Sampling method [Scan, Focal, Instantaneous, Conspicuous behaviour]
17. Recording method [Continuous; Instantaneous, One-zero]
18. Hours of observations
19. Study time period
20. Time of day of samples (how representative of the time period are the samples) [e.g. Consistent time, spread throughout day, spread throughout night, spread throughout 24 hours, etc]

Section 4**. RELIABILITY AND VALIDITY**

21. Types of reliability and validity which have been assessed [*list all that apply and whether they reached criteria for acceptance* Intra-rater reliability; Inter-rater reliability; Test re-test reliability; Internal reliability; Content/face validity; Construct validity; Criterion validity (concurrent or predictive)]
22. Which method was used to assess Criterion Validity of welfare measure? [Presence or absence of motivated items (welfare measured when has and doesn't have items it is motivated to access, approach or avoid); Correlation with behavioural measures of welfare; Correlation with physiological indicators of welfare; Correlation with affective measures of welfare (e.g. cognitive bias); Short term manipulation of welfare state (hours); Long term manipulation of environment (days); Comparison with natural or functional behaviour]
23. What is evidence that criterion validity has been demonstrated? [*For* Comparison with choices and manipulations of welfare state *list Effect size (Mean difference between groups/ standard deviation across groups)*; *For* Correlational designs, *list Correlation coefficients*; *For* Comparison with wild, *list percentage difference in time captive vs wild/ mean time spent in activity in wild*.]
24. Statistics used in the paper [Non-parametric (e.g. Spearman's, correlation, Kruskal Wallis, Wilcoxon); Parametric with no random effects (e.g. Pearson's, t-test, ANOVA, GLM); Modelling or other control for random effects (e.g. mixed models, multi-level)]

.

Figure 1. Critical Appraisal form for extracting information for critical appraisal of welfare measures used in published studies. *Instructions are listed in italics.* [Unless otherwise stated mutually exclusive options are listed in square brackets.]

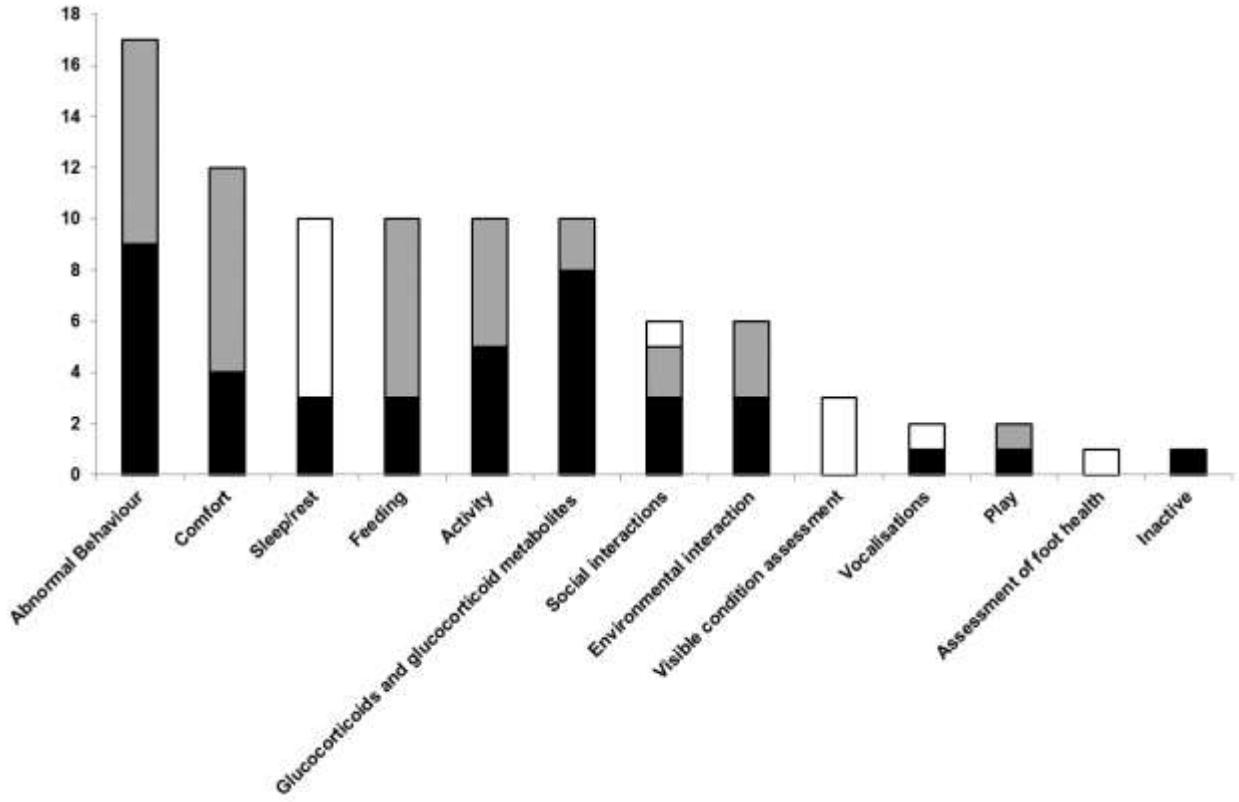

Figure 2. The total number of papers on captive elephants which conducted studies on each welfare indicator (height of bars), along with the number of papers which have shown a percentage change between treatments presumed to influence welfare (height of grey bars), and the number of these which have demonstrated a significant difference (height of black bars).