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	14 EM (0.14), 3 LA (0.3)	1	Criterion	None reported	October 1995 (4 x 8hrs) and April 1996 (3 x 24hrs)	Monitoring behavioural changes prior to scheduled events		Stereotypies[%] Resting[%] Eating/drinking[%] Dust bathing[%]
	EM/LA	14 EM (0.14), 3	1	Criterion		April 1996 and April	Monitoring behavioural	Experimental, repeated	Stereotypies[S] Standing

Authors	Species	Sample size	No. of institutions	Validity	Reliability	Time period	Method of assessing welfare	Study design	Welfare measures
Friend & Parker (1999)		LA (0.3)			None reported	1998 (3 x 24hrs during each period)	changes following changes to the environment	measures, quantitative	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]
Haspeslagh et al (2013)	EM	87 (16.71)	32	Construct	None reported	N/A	correlation predicted between behavioural		Stereotypies[C] Foot health[C]

Authors	Species	Sample size	No. of institutions	Validity	Reliability	Time period	Method of assessing welfare	Study design	Welfare measures
							and physical measures of welfare		
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	Monitoring change in behaviour and cortisol following	Prospective, observational, repeated measures, quantitative	Stereotypies[%] Sleep[%] FGM[S]

Authors	Species	Sample size	No. of institutions	Validity	Reliability	Time period	Method of assessing welfare	Study design	Welfare measures
						10 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	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
							concentration range		
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

Authors	Species	Sample size	No. of institutions	Validity	Reliability	Time period	Method of assessing welfare	Study design	Welfare measures
Rees (2004)	EM	8 (2.6)	1	Criterion	None reported – measures taken to increase reliability	35 days	Monitoring how stereotypies change over time	Observational, repeated measures, quantitative	Stereotypies[S]
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	Experimental, repeated measures, quantitative	Stereotypies Social Comfort Feeding Locomotion Resting Manipulation/exploration Glucocorticoids

Authors	Species	Sample size	No. of institutions	Validity	Reliability	Time period	Method of assessing welfare	Study design	Welfare measures
							presumably stressful event		
Soltis (2010)	LA	4 (4.0)	1	Criterion	None reported	14.5 months	Documenting elephant rumbles in different social situations	Observational, repeated measures	Vocalisation[S]
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	Environmental manipulation	Experimental, repeated	Stand Move Socialise Aggression

Authors	Species	Sample size	No. of institutions	Validity	Reliability	Time period	Method of assessing welfare	Study design	Welfare measures
						repeated 4 months later for 3 days		measures, quantitative	Dust bathe Object interaction Eat Drink Vocalise Abnormal behav[S]
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	Repeated measures, observational, quantitative	Blood cortisol[%C] Stereotypies[%C]

Authors	Species	Sample size	No. of institutions	Validity	Reliability	Time period	Method of assessing welfare	Study design	Welfare measures
							(1992, 1994, 2001)		