Zoo Biology 24:499–518 (2005)

Research Article

Scientific Approaches to Enrichment and Stereotypies in Zoo Animals: What's Been Done and Where Should We Go Next?

Ronald R. Swaisgood[1]* and David J. Shepherdson[2]

[1]*Center for Reproduction of Endangered Species, Zoological Society of San Diego, San Diego, California*
[2]*Oregon Zoo, Portland, Oregon*

The zoo scientific community was among the first to focus attention on captivity-induced stereotypic behaviors, their causes, and methods of eradication. Environmental enrichment has emerged recently as the main husbandry tool for tackling this problem. An increasing number of research publications have attempted to evaluate the effectiveness of enrichment in reducing stereotypic behavior and to develop further concepts to explain how effective enrichment works. A review and meta-analysis of this literature indicates that enrichment is a successful technique for reducing stereotypic behavior in zoo animals. Enrichment was associated with significant reduction in stereotypy performance about 53% of the time. Published enrichment and stereotypy research is lacking for most zoo species, with most studies on large, charismatic, and often endangered species, but it is unclear whether stereotypies are more prevalent in these species. In addition, problems with scientific methods and data presentation, quantitatively detailed in this work, severely limit the conclusions drawn from zoo research. Further understanding of what kinds of enrichment works and what doesn't will require greater attention to experimental design, sample size, statistical analysis, and better descriptions of enrichment properties and the form of stereotypy. We recommend that future studies focus on increasing sample size (e.g., through multi-institutional studies), appropriate repeated measures design (e.g., with multiple baseline and experimental phases), providing full statistical

*Correspondence to: Ronald R. Swaisgood, Conservation and Research for Endangered Species, Zoological Society of San Diego, P.O. Box 120551, San Diego, CA 92112.
E-mail: rswaisgood@sandiegozoo.org

Received 09 August 2004; Accepted 28 April 2005

DOI 10.1002/zoo.20066
Published online 22 July 2005 in Wiley InterScience (www.interscience.wiley.com).

information about the behavioral changes observed (including standard error), and ultimately the development of a predictive science for enrichment, stereotypies, and wellbeing. Zoo Biol 24:499–518, 2005.

Key words: enrichment; stereotypy; wellbeing; literature review; scientific method

INTRODUCTION

The zoo community was among the first to raise concerns over abnormal and stereotypic behaviors in captive animals and to begin to develop environmental enrichment strategies to deal with the perceived problem. Stereotypies are relatively invariant, repetitive behaviors that seem to have no immediate function [Mason, 1991]. Environmental enrichment can be defined loosely as an animal husbandry principle that seeks to enhance the quality of captive care by identifying and providing environmental stimuli necessary for optimal psychological and physiological wellbeing [Shepherdson, 1998].

The relationship between the performance of stereotypic behavior and wellbeing is complicated, however, and studies strongly suggest that there is not a one-to-one relationship [Mason and Mendl, 1993; Mason and Latham, 2004]. Performance of stereotypies may not correspond to current wellbeing because stereotypies may be a "scar" from previous suboptimal environments. In addition, because stereotyping may be a means of coping with an aversive environment, individual animals that perform stereotypies in suboptimal environments may well have better welfare than those that do not perform stereotypies in the same environment. Stereotypies are on the whole more than just loosely correlated with wellbeing, however: Mason and Latham's [2004] survey showed that 68% of environments that cause stereotypies are associated with diminished welfare. When discussing the meaning of stereotypies in a zoo environment, we thus follow their advice that "stereotypies should always be taken seriously as a warning sign of potential suffering, but never used as the sole index of welfare." In the absence of corroborating data on wellbeing, we consider stereotyping animals at high risk for suboptimal wellbeing.

In the zoo community, environmental enrichment has become almost a catchall term for husbandry activities with the specific aim of improving wellbeing and as such is the method of choice for reducing stereotypic behavior. Environmental enrichment involves changing the environment of the zoo animal to provide opportunities or choices not available before. Often a heavy emphasis is placed on the importance of providing enrichment that is appropriate to the specific biology (to the extent to which it is known) of the species under consideration [Mellen and MacPhee, 2001]. Biologically appropriate complexity can be increased in many ways, for example, by adding substrates such as dirt, litter, mulch, vegetation, or trees. These substrates increase "information content" of the environment and elicit foraging and investigatory behavior by concealing food, smells, naturally occurring insects, or other wildlife, etc. Barriers and landscaping can provide privacy, promote territorial behavior, provide escape routes, and thus improve social interactions. Toys and novel objects elicit exploration and creative play. Climbing structures allow more efficient use of space and provide shade and temperature gradients for choice of microclimate. They can also provide hiding places from conspecifics, the public, and keepers. Cognitive challenges, such as mechanical apparatuses, puzzle feeders, or computer interaction with visitors, put captive animals in a position in which they can learn to actively control and explore some aspect of their environment. Various feeding devices and practices such as carcass and whole-fruit feeding allow animals to acquire and process food in diverse, versatile, and more natural ways. More recently, the potential of training, not just as a management tool but also as cognitive enrichment for captive animals, has begun to be realized [Laule and Desmond, 1998].

What Lessons Have We Learned from Zoo Enrichment Studies?

With more than three decades of research and development of the enrichment ethos in the zoo community behind us, a timely assessment of what we have learned from these experiences seems in order. In an earlier analysis we sought to do just that [Swaisgood and Shepherdson, in press]. We reviewed the literature on enrichment studies designed to reduce stereotypic behavior in zoo animals since 1990 and attempted a meta-analysis to try to discern what kinds of enrichments work and what doesn't. We discovered a great deal of creativity and ingenuity in the way that zoo researchers go about developing enrichment programs. Unfortunately, we were unable to find much evidence that some kinds of enrichment work better than others and gained little insight into the motivational basis of stereotypies and enrichment effects. We concluded that one reason we failed to find differences in the efficacy of enrichment was that investigators probably tailored the enrichment program to the particular needs of the study animals and their situation. A thoughtful approach such as this, if used by most investigators, would obscure any universal trends explaining why some enrichment strategies work better than others. That is, one enrichment strategy may work best with species A in situation X, whereas another may work better with other species in other situations. The fact that many types of enrichment reduce stereotypies is also consistent with the argument that stereotypies do not arise from a single underlying motivational state [Rushen et al., 1993].

Our failure to discover what makes enrichment work cannot solely be attributed to these factors. There were also limitations in the literature that presented obstacles to our analysis. Most prominently, the literature contained many fewer peer-reviewed publications on the topic than expected. Only 23 studies provided the needed information and met our criteria for inclusion in the analysis. Because these studies covered a diverse array of species and included many variations in the captive environment and enrichment strategy tested, we had few representatives for each category of enrichment and many confounding variables obfuscating the analysis.

This analysis led us to explore systematically another aspect of zoo enrichment research, namely, how enrichment practitioners go about studying enrichment in a zoo setting. Here we present quantitative data on the design and methods used by zoo researchers. The analysis is based on a systematic review of the literature. We have expanded the number of included studies since the first analysis to 41 by relaxing one of our criteria from our previous analysis: that authors reported the level of stereotypic behavior before and after enrichment (see Methods). With this more robust sample, we examine such issues as the taxonomic units selected for study, sample size, experimental design, statistics, dependent variables measured, as well as further measures of enrichment efficacy.

We in the zoo community deserve much of the credit for the development of the concept and application of enrichment. The purpose of this review is to assess precisely what has been done and to provide timely feedback to the zoo community. The successes are many, but what are the shortcomings of this body of work? How can we refocus our enrichment research efforts in the future to get better answers to our questions and ultimately develop a more predictive science for structuring enrichment programs?

MATERIALS AND METHODS

Inclusion Criteria for Literature Review and Analysis

We reviewed all studies published from 1990 to 2003 in three peer-reviewed journals: *Animal Welfare*, *Applied Animal Behavior Sciences*, and *Zoo Biology*. Other peer-reviewed journals rarely publish zoo research on the topic. Conference proceedings and journals that are not peer-reviewed are also a rich source of information, but we chose not to include them because one of our objectives was to evaluate the level of scientific and statistical sophistication of zoo-based research. To find articles, we searched for the following keywords: stereotypy, stereotypic, abnormal behavio(u)r, enrichment, wellbeing, and welfare (except in *Animal Welfare*, where welfare is always a keyword). From this sample, we included only publications of empirical studies meeting the following criteria: (1) the animals were studied in two different situations that varied in terms of enrichment quality (i.e., control vs. enriched); (2) the effects of enrichment on stereotypies were evaluated with inferential statistics; and (3) the study was conducted at a zoological park, aquarium, or conservation breeding center (e.g., studies at biomedical research facilities were not included). We found 25 publications meeting these criteria and completed a four-page questionnaire for each. We read each articles and answered a series of specific questions relating to the scientific process, enrichment strategy, and results.

Experimental Design and Statistics Used by Zoo Researchers

We categorized several aspects of the experimental design and statistics used in the studies included in our analysis. We recorded the sample size, as determined by the number of individuals used in the statistical analysis of the effects of enrichment on stereotypy performance, which was sometimes smaller than was the sample size for the article as a whole. Experimental design was classified as following [see also Saudargas and Drummer, 1996]: (1) before/after design, also referred to as baseline (A)/experimental (B) or AB, same individuals studied before and after a permanent change in enrichment conditions (repeated-measures design); (2) repeated treatment design (ABA, ABAB, etc.), as above, but baseline and experimental phases repeated more than once (used when enrichment can be given and withdrawn repeatedly or animals can be moved into and out of different enclosures); and (3) between-subjects study, some subjects given Treatment A (or baseline) and some subjects given Treatment B (different subjects are exposed to different environments that vary in degree or type of enrichment). We also determined whether only one specific enrichment variable was changed or whether a suite of variables was changed simultaneously.

Statistics were categorized as follows: (1) individual-animal statistics: statistical tests reported for individual animals across multiple trials; (2) analysis of pooled data: data from multiple animals analyzed together, but one animal contributes more than one observation to the analysis without the use of specific repeated measures in the statistical model, e.g., before/after, time of day, season, etc. (day of observation was not considered a legitimate repeated measure because there is no reason to believe that day has a systematic effect; where the effects of potential habituation are of interest, trial number is considered legitimate); and (3) "legitimate" between-subjects or within-subjects statistics, such as analysis of variance (ANOVA) and *t*-tests and their nonparametric counterparts, where appropriate degrees of freedom were used (see below).

We also quantified several aspects of study effort: duration of observation periods, total observer time spent collecting data (including controls), and duration of the study.

Species Characteristics

For each species in our sample, we recorded its taxonomic family, its foraging strategy (herbivore, carnivore, and omnivore), its sociality (social/asocial), and whether it is listed by the IUCN as a threatened or endangered species.

Categorization of Enrichment

In a previous analysis, we categorized enrichment according to several variables that might influence its success [Swaisgood and Shepherdson, in press]. Here we present descriptive information on exactly what enrichments were used. Nonfeeding enrichment included the following. (1) Objects: nonpermanent addition of (novel) objects for manipulation/play. (2) Olfactory: the addition of scents or scented material to the enclosure. (3) Training: using operant conditioning to get the animal to do something that it has not done before for the express purpose of challenging the animal's cognitive skills. (4) Enclosure rotation: moving the animal repeatedly between two or more enclosures (similarly enriched) to retain some novelty value. (5) Major exhibit changes: enclosure undergoes major renovation or animal is moved to a new enclosure. The following changes were noted: (a) live vegetation; (b) shelter, hiding place; (c) climbing structure, elevated perches; (d) water source; (e) loose substrate (for digging, resting); (e) other permanent moveable furnishings; (f) other unmovable permanent furnishings; and (g) increased enclosure size.

Feeding enrichment was classified according the following nonexclusive criteria: (1) designed to increase search time (e.g., scatter or hide); (2) designed to increase capture time (e.g., live prey); (3) designed to increase extraction time (e.g., puzzle feeder); (4) designed to increase processing time related to handling and mastication (e.g., vegetation/browse, bones, ice blocks with food, whole food, carcasses); (5) designed to increase temporal variability of feeding times (change from feeding at set times); and (6) designed to increase number of feeding times/day.

Dependent Variables and Data Analysis

For each behavior measured by the researchers, we recorded whether or not enrichment had a statistically significant effect. We used this categorical variable in our analyses across studies. This type of literature meta-analysis, called "vote counting,'' is not the preferred method. The recommended method of meta-analysis is to calculate an effect size for each study, for example, the magnitude of change from baseline to experimental conditions [Lipsey and Wilson, 2001]. This is the method we used in our previous study [Swaisgood and Shepherdson, in press], but because authors did not always report stereotypy levels, our sample size was much smaller (23 instead of 41). Another problem we encountered was that authors sometimes do not report measures of variability, such as standard error, which are needed to calculate a correction factor for meta-analysis. Nonetheless, the most significant criticism of the vote-counting method is that statistical significance for individual studies included in the meta-analysis is strongly affected by sample size. Sample sizes in our literature sample were consistently small, however, and varied little.

When reported by the authors, we categorized the stereotypy according to the following stereotypic forms: oral (e.g., self-bite/suck, tongue flick, object-bite/suck, regurgitation), locomotor (pacing and repetitive swimming patterns), repetitive movement (head, body, limb movements, etc.), and other. Authors often lumped several forms of stereotypy together for analysis, which we labeled ''composite stereotypy.''

Authors often reported other variables affected by enrichment, and we categorized these as natural/normal behavior (e.g., a behavior normally seen in the wild, such as feed, forage, locomote, explore), behavioral diversity (increased number of behavior patterns displayed), active (nonstereotypic activity levels), and corticoids.

Our sample consists of 25 publications (see below). Some statistics reported are based on this unit of analysis, which we refer to as ''publication.'' Other reported statistics are based on what refer to as a ''study'' ($n = 41$). A publication contained more than one study when one of three criteria was met: (1) the authors included two completely separate experiments testing different enrichment strategies; (2) the authors included two or more types of enrichment as independent variables in the analysis; or (3) the authors provided separate analyses for different species. We consider these statistical replicates to be relatively independent because in each a successful or unsuccessful finding could be obtained. For example, if one type of enrichment did not work with a study group that does not mean that another type of enrichment will also fail. Similarly, if enrichment reduces stereotypies in one species, it does not follow that the same enrichment will work for other species.

Many results we present are simply descriptive statistics. Most reported inferential statistics are based on the χ^2 statistic. When assumptions of this test were not met (i.e., $>20\%$ of cells with expected values less than 5), we used the Fisher's exact test. For trends across years we used Spearman's rank correlation. We report results from more than 10 statistical tests but do not use any correction factors for familywise errors (e.g., Bonferroni adjustment). We chose not to do this for two reasons. First, it is only necessary to correct for multiple testing among related hypotheses, i.e., those hypotheses attempting to address the same general research question [Quinn and Keough, 2002], but our statistical tests are conducted on a series of unique hypotheses with unique predictions. Second, our data set yields limited statistical power, thus increasing statistical conservatism would stack the odds too strongly in favor of negative findings (see recent discussion by Nakagawa [2004]).

RESULTS AND DISCUSSION

Zoo-Based Studies on Enrichment and Stereotypies

The keyword search of the literature turned up 101 publications using one or more of the key words. Of these publications, 31 were conducted at biomedical or university facilities and 14 occurred at agricultural facilities. Most of the study subjects in these publications were not species typically held in zoos. The remaining 66 publications were based on work conducted at zoos and breeding centers for rare wildlife, accounting for well over half the published studies in this area. The journals searched were chosen because they were likely to contain zoo-based research, however, whereas much of the biomedical and agricultural research is published elsewhere. Of 66 zoo-based publications, 48 included a test of the effectiveness of an enrichment treatment, of which 25 measured stereotypies. If we assume that most enrichment studies include a measurement of stereotypies if they existed, this latter statistic suggests that many enrichment studies are conducted with animals that do not perform stereotypies (or stereotypies occur at levels too low to merit analysis). Zoo researchers should be commended for their efforts to enrich the lives of captive animals even when this most common index of compromised wellbeing is lacking.

Experimental Design Used by Zoo Researchers

Sample size is one of the most important aspects of scientific investigation because larger numbers of individuals increase external validity. Without a sufficient number of subjects, valid statistical procedures are not possible and any conclusions cannot be generalized to the larger population. For example, a sample size of six subjects tested in two experimental conditions is required before the Wilcoxon matched-pairs test can reach significance. Here the performance in zoo research needs improvement. In our survey, the median sample size was 4, with a range of 1 to 17, excluding an outlier of 257 from a questionnaire survey. Unsurprisingly, sample sizes for social species are larger (median = 9) presumably because several can be observed at once at the same institution (bringing into question their statistical independence). These small sample sizes mean that many of the conclusions drawn from these studies may not hold up to future investigation. To discover whether there have been improvements in recent years, we ran a Spearman's rank correlation between year and sample size on a per publication basis, and found a weak but non-significant trend toward increasing sample sizes ($\rho = 0.28$, $Z = 1.4$, $P = 0.16$; Fig. 1). Figure 1 also reveals that over the same period of time average observation period length (in hours) decreased nonsignificantly ($\rho = 0.36$, $Z = 1.7$, $P = 0.08$). Other measures of study effort, total hours of observation time and study duration (in months), seem flat between 1990 and 2003.

Taken together, these graphs suggest that zoo researchers have maintained a relatively constant effort over the past 13 years, but are beginning to observe more animals for a shorter period of time. We hope that this trend continues because it is the number of subjects studied, not the number of samples collected, that increases statistical power and improves generalizability [Machlis et al., 1985]. There are times when intensive observation on few animals is useful, but as a rule, research time is better spent observing more animals, even if this means collaborating with multiple institutions.

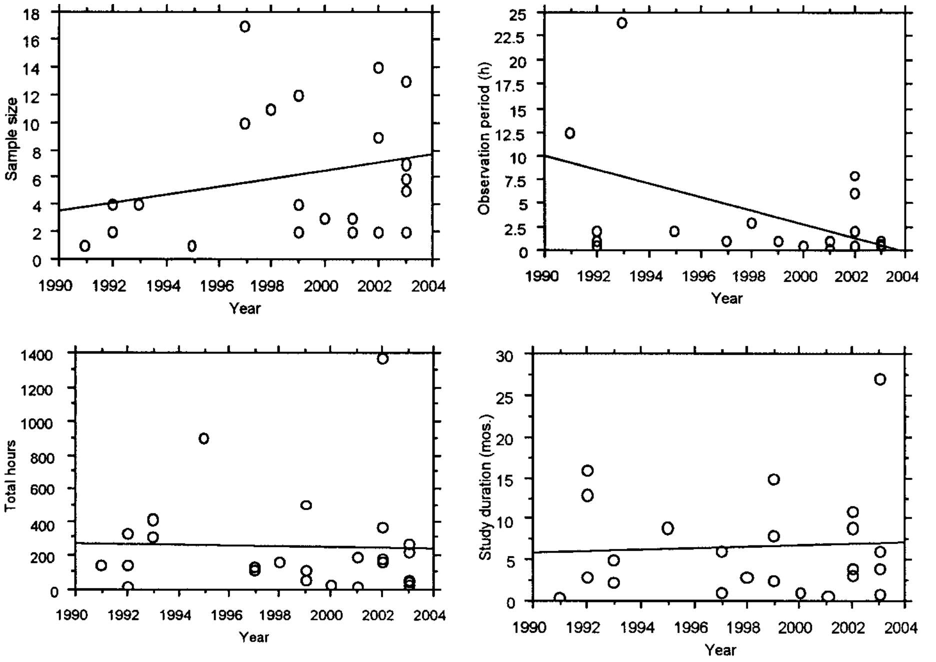

Fig. 1. Relationship between year and several variables related to research effort. Spearman's rank correlation is not significant. Regression line included for descriptive purposes only. Outlier (N = 257) from questionnaire survey excluded.

We also evaluated the type of experimental design used by zoo researchers. The before/after design was used in 11 of the publications. This design was most often used when an animal's enclosure was renovated or the animal was moved to a new enclosure. It is also often used when a new enrichment program is first implemented. In general, this design is used for large-scale changes that are relatively difficult to reverse. Although it has its place in experimental design, it suffers more from the potential for confounding variables, such as changing season or reproductive condition.

The repeated-treatment design was used in 10 publications. This design improves upon the first by increasing internal validity. With increasing repetition of treatment and baseline, the possibility of confounding variables is gradually diminished. In one type of this design, multiple observations are made for baseline condition and examined for stability across time. If no trend is evident, the investigator can proceed with the enrichment treatment relatively secure that a trend change can be attributed to the enrichment [Saudargas and Drummer, 1996]. This method is repeated with each AB cycle. With increasing numbers of AB cycles, however, this trend analysis becomes less important because alternative explanations can be ruled out. For example, if a subject is studied on 20 days without enrichment (A) and 20 days with enrichment (B), it is highly unlikely that potential confounding variables such as noise or crowd size will occur coincidentally with either A or B. Although these sources of variability will remain present, they become "statistical noise" rather than potential systematic bias.

In between-subjects studies, seen in three of the publications, the investigator measures the behavior of some individuals in enrichment condition A and a different set of individuals in B. This design reduces confounding variables only if the subjects within each treatment are independent of one another, i.e., not belonging to the same social group or all housed at the same location [Hurlbert, 1984]. With small sample sizes, individual histories and temperament become important confounding variables, and results should be interpreted with caution. The internal validity of this design thus depends on the circumstances. A final way to measure the effects of enrichment on stereotypy is the questionnaire survey, observed once in our sample. In this example, the investigators [Bashaw et al., 2001] obtained responses from 49 institutions on the occurrence of stereotypies in 257 giraffe and okapi. Although these data are not fully independent, this impressive sample size allowed these researchers to correlate several enclosure and husbandry variables with stereotypies. More of this sort of study, with follow-up observational and experimental work, is clearly needed. Ideally, experimental or at least observational work is needed to validate the reliability of questionnaire surveys [Shepherdson, 2001].

In general, the most enlightening studies are multi-institutional (or at least based on reasonable sample sizes) and are of the multiple-cycle ABAB-type design. We encourage the increased use of these kinds of studies in the zoo community.

Another aspect of experimental design that can slow progress toward understanding what sorts of enrichment really work is the use of many different forms of enrichment simultaneously. This everything-but-the-kitchen-sink approach can produce marked results, but subsequent practitioners cannot determine the enrichments that were most important in producing the result [Swaisgood and Shepherdson, in press]. Sixteen of the publications relied on this method. In 3 of these 16 cases, however, the investigators studied multiple enrichments but with only one present at a time, and hence could analyze the effects of each separately. In the remaining 9 (of 25) publications only one aspect of enrichment was changed, which although potentially less efficacious on the whole yields clear results for a specific enrichment. We sympathize with the kitchen-sink approach because zoo personnel are often justly motivated to bring their charges to optimal wellbeing as quickly as possible. We need to acknowledge, however, that this is done at the expense of gaining a better understanding of enrichment strategies that will ultimately enhance wellbeing across multiple institutions. Moreover, in our sample the multiple-enrichment strategy was no more effective at reducing stereotypy than was the single-change strategy ($\chi^2 = 1.7$, $P = 0.44$), suggesting that the kitchen-sink approach may not be an improvement over thoughtful selection of single enrichments.

In a related issue, zoo researches seem to choose their enrichments thoughtfully, only testing enrichments that they predict will be effective for the circumstances of their subjects. They can hardly be faulted for this, but again this limits our ability to determine what works and what doesn't or when it works. We do not advocate that investigators purposely compromise animal wellbeing by providing inappropriate enrichment. If they design studies that test the effects of several forms of enrichment, however, including those they believe are unrelated to the motivational basis for stereotypies in their animals, perhaps then we will find that all enrichments don't work equally well and, more importantly, why. Another factor compromising our meta-analysis is the near certainty that the published literature is highly biased toward positive findings, largely because of unfavorable peer review.

A second way of advancing understanding for why some enrichment doesn't work, then, is to include negative results (perhaps from earlier studies) in publications of results with positive findings.

Statistical Analyses Used by Zoo Researchers

In most publications in our sample, authors used what we view to be legitimate statistical analysis according to our criteria. In only one publication did authors analyze their data with a legitimate between-subjects analysis. The scarcity of this analysis is testament to the small sample sizes that prevail in zoo studies. By contrast, authors frequently incorporated repeated measures (within-subjects design) into the analysis ($n = 13$), thereby increasing the degrees of freedom and statistical power above the threshold necessary to attain statistical significance. Repeated measures were often varying enrichment conditions, but other blocking variables such as season or time of day were also used. Not only does this make legitimate statistical analysis feasible, but it also reduces the effects of extraneous variables that add statistical noise to the analysis.

A simple example will make this point clear. The investigator wishes to evaluate the effects of enrichment on circulating corticoids, a measure of stress. From previous studies she knows that both time of day (Morning/Afternoon) and season (Spring/Summer/Fall/Winter) affect corticoid levels, and therefore decides to use them as blocking variables in the analysis as a statistical control. Four subjects are exposed to two different forms of enrichment during each season and time period. The resulting analysis contains 64 data points (2 time periods $\times$ 4 seasons $\times$ 4 subjects $\times$ 2 treatments). The residual error for the denominator for the F-statistic in the ANOVA model contains 52 degrees of freedom (*df*) if no interactions are included in the model. The caveat for this kind of approach is that researchers cannot just add blocking variables without regard to biological relevance. For example, although separating the data set into two time periods may be legitimate, separating it into 24 1-hr periods to increase the degrees of freedom 24-fold would clearly be inappropriate. There is no agreed upon criterion for how few individuals may be used in a repeated-measures analysis. If enough factors are added to the statistical model, it is indeed possible to attain statistical significance for a single individual. Even so, it is clearly unreasonable to generalize to the population based on $n = 1$. Somewhat arbitrarily we recommend a goal of at least six subjects, with a minimum of four, but with cautious interpretation (see below).

Six of the publications in our sample suffer from data pooling errors. When n subjects are observed k times, and the data are pooled to create a sample size of $n \times k$, the data are not independent and degrees of freedom are artificially inflated, resulting in inflated Type I error rate [Machlis et al., 1985]. Simply adding subject as a factor to the statistical model (usually ANOVA) does not guard against data pooling, unless legitimate repeated measures account for the multiple data points (see above). Statistically, large samples of behavior should not be equated with large numbers of individuals. Clearly, obtaining 100 samples from three individuals is not as generalizable to the population level as are 100 samples from 100 individuals.

A common alternative to pooling, found in four publications in our sample, is to compensate for small sample size by calculating statistics on data from individual animals separately. Although less than ideal, we consider this statistical method a relatively legitimate approach to dealing with small samples sizes [see also Saudargas and Drummer, 1996]. Authors should always acknowledge the limitation of individual-animal statistics, however, by stating clearly in the abstract and elsewhere that they cannot generalize to the larger population and can only make claims to the effects for the specific subjects in the study [e.g., Owen et al., 2004]. No authors in our sample did this. There are, however, cases where individual-animal statistics are advantageous, and therefore generally preferred to pooling. Many aspects of captive management, such as enrichment plans and stress mitigation, require understanding the needs of individual animals, which can be highly variable, so that individualized plans of action can be crafted. For example, one animal may display signs of stress in response to vehicular traffic noise whereas another may not. In these cases, statistics on the individual are more enlightening than are population trends [Swaisgood, Owen et al., in press].

When sufficient individuals are not available and research is conducted nonetheless, we recommend authors resort to descriptive statistics or individual-animal statistics. We believe, however, that such studies are warranted only when the authors are addressing a somewhat novel research question, one where even a small amount of new information will be useful. Alternatively, this approach may be used when the information is needed to fine-tune management for individual animals, but this may be of limited interest to the scientific community at large.

These cases represent the most extreme of numerous forms of pseudoreplication (meaning that replicates are not statistically independent), a slippery concept that has spawned much debate. In their review of pseudoreplication, Bart et al. [1998] recognize that ''reasonable people may disagree'' and conclude that the ''best general advice is simply that investigators should clearly specify the population to which their statistical inferences apply'' (p. 181). We concur and advocate this approach in reporting zoo-based research findings, which are often marked by a greater degree of pseudoreplication than is biological research in general. We also point out that the best way out of this dilemma is the increased use of multi-institutional studies that minimize pseudoreplication and maximize external validity.

Species Characteristics

The species characteristics of animals included in stereotypy research may illuminate the traits that promote stereotypy, assuming investigators are more likely to study animals that perform stereotypy. Choice of animals may also reveal the researchers' biases. Table 1 displays the species and families covered in our review. The felids were the most studied species, followed by ursids and hominids. These families also contain some of the most notoriously stereotypic zoo animals. Figure 2 presents data on the percentage time that animals from these families spend performing stereotypies before enrichment. These data are instructive and provide an idea of the extent of stereotypies in zoo animals, but these values should not be used to make comparisons among families because the small sample sizes and variable observation times and effort mean that they may not be representative. Arguably, many of the species represented in Table 1 could be classified as ''charismatic megafauna.'' These large-bodied animals tend to have large home ranges, however, a factor known to correlate with stereotypy performance in carnivores [Clubb and Mason, 2003]. The ursids and felids also tend to be solitary in nature and may fare worse in captivity than social species because they do not have access to or benefit from interaction with conspecifics. Indeed, asocial species (28)

TABLE 1. Taxonomic classification of animals studied in the publications used in the analysis (some publications include more than one species or family)

Family	Species	No. of articles
Cercopithecidae	Mandrill	1
Elephantidae	Elephant	1
Felidae	Cheetah, jaguar, jungle cat, leopard, leopard cat, lion, lynx, tiger	9
Giraffidae	Giraffe, okapi	2
Hominidae	Chimpanzee, gorilla	5
Phocidae	Harbor seal, gray seal	2
Pongidae	Orangutan	2
Psittacidae	Conure	1
Suidae	Babirusa	1
Tapiridae	Tapir	1
Ursidae	American black bear, Asiatic black bear, brown bear, giant panda, polar bear, sloth bear, spectacled bear	5

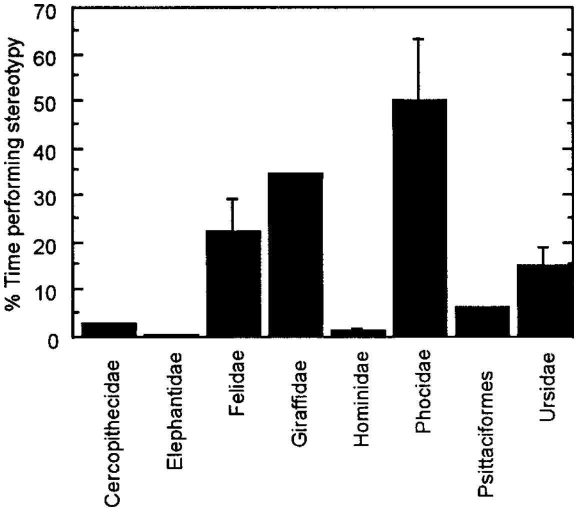

Fig. 2. Stereotypy performance as a function of taxonomic family. Analysis based on data reported for control or pre-enrichment baseline conditions (N = 21).

also predominated over social species (16) in our sample. Carnivory was the most frequently represented foraging strategy in our sample, with 24 studies on animals classified as carnivorous. By contrast, only 11 herbivores and 9 omnivores were included. Carnivores also tend to be solitary and have large home ranges, so this factor may again offer a partial explanation for their overrepresentation.

Our data set included 18 studies of threatened or endangered species and 22 that had no conservation status. Species with larger home range size are also more vulnerable to extinction [Woodroffe and Ginsberg, 2000], which may mean that many species that are endangered are also more vulnerable to stereotypy development. Nonetheless, because the majority (85%) of animals in zoos are not endangered [Magin et al., 1994], this suggests that managers and researchers are targeting endangered species with enrichment programs to improve wellbeing. Endangered species get more than their fair share of attention in many regards, but the link between enrichment, wellbeing, and reproduction [Carlstead and Shepherdson, 1994] and behavioral competence for reintroduction to the wild [Shepherdson, 1994] must play an important role in these efforts. For example, enrichment programs are playing a major role in conservation breeding efforts for the highly endangered giant panda [Swaisgood, Zhang et al., in press; Swaisgood et al., 2003].

In a previous report we were unable to discern any effects of taxonomic position on the efficacy of enrichment in reducing stereotypy [Swaisgood and Shepherdson, in press], although final conclusions require further evaluation. There were clearly no differences between the carnivore, primate, and other families, however, with all groups showing about a 50% reduction in stereotypy performance after enrichment. Given the wide variety of ecological strategies within these groups, this result is perhaps unsurprising. In the current analysis, we examined whether foraging strategy determined whether enrichment significantly reduced stereotypy levels. We found a marginally nonsignificant effect ($\chi^2 = 5.5$, $P = 0.06$): it seemed most effective for carnivores (9/13), moderately effective for herbivores (11/21), and least effective for omnivores (1/6). We await more data with fewer confounding variables (e.g., omnivores may have received some less effective forms of enrichment in this sample), however, before drawing any conclusions on the effects of foraging strategy. By contrast, sociality was more obviously unrelated to enrichment efficacy: enrichment was associated with a significant reduction in stereotypies in 11/21 tests for social species and in 9/16 of tests on asocial species ($P = 0.81$).

Excluding tests on groups of mixed species with varying conservation status, enrichment with endangered species (4/14) was significantly less likely to reduce stereotypies than was enrichment with nonendangered species (15/22; $\chi^2 = 5.5$, $P = 0.02$). Because the median sample size was five for endangered species, sample size is a little better than for the survey as a whole (four), and cannot explain lack of statistical effect. A better explanation lies in the fact that endangered species were displaying less stereotypic behavior before the enrichment was tested (Mann-Whitney *U*-test: n = 6, 10; Z = 2.6, $P = 0.009$; Fig. 3). This low level of stereotypy performance in baseline conditions makes it difficult to significantly reduce this level

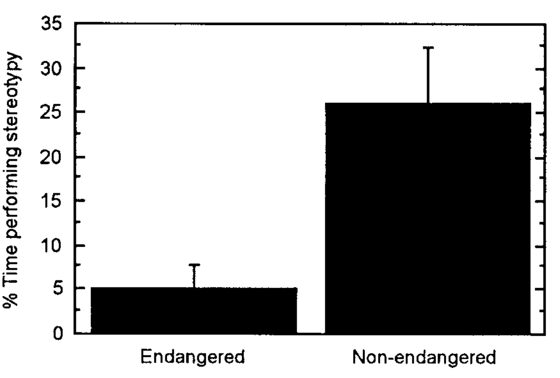

Fig. 3. Stereotypy performance as a function of species conservation status. Analysis based on data reported for control or pre-enrichment baseline conditions (N = 16).

further. The most plausible interpretation is that endangered species are already living under more enriched circumstance in baseline conditions. This is good news for endangered species, which deserve the extra attention to promote reproduction, but it may make them less suitable for understanding the relationship between enrichment and stereotypies.

Studies on smaller, less charismatic species are notably lacking. We advocate more research on these overlooked species, which often will help meet two objectives: (1) larger *n* because of greater numbers of animals held at individual institutions; and (2) more diversity in terms of phylogeny, life-history strategies, foraging strategies, and so forth.

Measuring the Effects of Enrichment on Stereotypy

Table 2 displays the types of enrichment we found in the literature sample. This list gives a rough idea of what enrichments were provided and how frequently they were used; however, in many cases the enrichments were so poorly described that we were unable to classify each change made. This was especially true for major exhibit changes. Here we can see that feeding enrichments were common, especially those that increased search, extraction, or processing effort. Scheduling changes to increase the number of feedings or reduce the predictability of feeding times were also relatively common. Of the nonfeeding enrichments, major exhibit changes were most common, followed by rotating animals through multiple equally enriched enclosures.

Although our data are limited for several forms of stereotypy, it is important to look at the various forms that stereotypies take because they can provide insight into

TABLE 2. Frequency of different types of enrichment used in studies included in the analysis (description of details of enrichment varied, so frequencies are approximations based on inference)

Enrichment type	Frequency
Novel objects	3
Olfactory	4
Training	0
Major exhibit change	7
Live vegetation	4
Shelter	1
Climbing structure	2
Water source	1
Substrate	3
Moveable furnishings	4
Unmovable furnishings	2
Increased size	5
Enclosure rotation	5
Other non-feeding	3
Feeding	
Search time	7
Capture time	3
Extraction time	10
Processing time	10
Temporal variability	5
No. feeding times	5
Other feeding	0

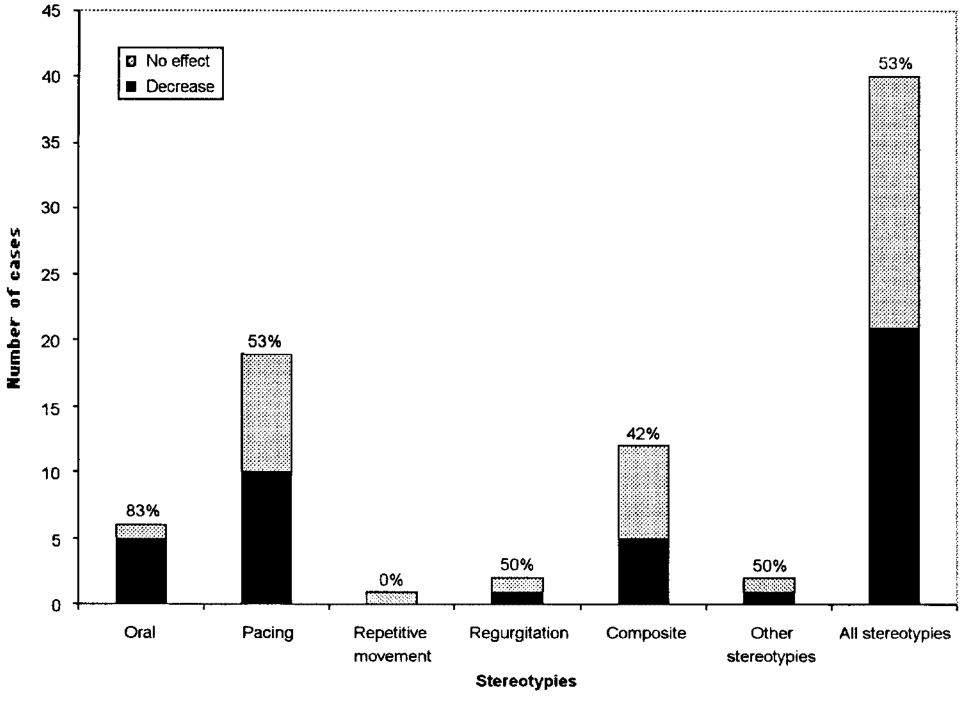

Fig. 4. The effects of enrichment on stereotypy performance. “Number of cases” refers to the number of times the stereotypic behavior was observed in the data set. “Composite stereotypies” was used when the authors lumped several different kinds of stereotypies together for analysis. “All stereotypies” refers to the effects of enrichment on any form of stereotypy in the other categories. “No effect” indicates that enrichment had no statistical effect on stereotypic behavior, whereas “Decrease” indicates that enrichment significantly decreased stereotypic behavior. The numbers above the columns refer to the percent of cases in which enrichment significantly decreased stereotypy performance.

motivation and causation [Mason, 1993]. Different forms of stereotypy may have different etiologies and arise from different motivational states. Locomotor stereotypies (pacing) were by far the most common form in our survey, and 53% of studies found a significant decrease after implementation of enrichment (Fig. 4). Analysis of more than one form of stereotypy together was also common, and enrichment was effective in 42% of these cases. Oral stereotypies were measured in six studies, and enrichment yielded a significant reduction in 83% of these studies. Repetitive movement, regurgitation, and other forms occurred rarely. When all 41 studies are considered together, enrichment significantly reduced stereotypies about half the time (53%).

With this limited data set, we cannot say much about the persistence of different forms of stereotypy when countered with enrichment. Is one form of stereotypy more difficult than another is to reduce or abolish? Again, one obstacle to such insights was the failure of some authors to report the occurrence of individual forms of stereotypy. “Composite” stereotypies are often reported because the frequency of individual forms is not sufficient for analysis, an understandable constraint; indeed one of us is guilty of this oversight [Swaisgood et al., 2001]. In the future, we encourage authors to present at least descriptive statistics for individual stereotypy forms to facilitate meta-analyses across many publications.

Another problem encountered in many of these publications is the lack of specific, detailed definitions for stereotypic behavior. In some cases the definitions were exceptionally vague. For example, some authors used the term "pacing" without further description. In our experience, one person's perception of what constitutes pacing can be quite different from another's. How invariant and how repetitive does the locomotion have to be to merit the designation of pacing? Does the footfall pattern have to be nearly identical or is it sufficient to wander around in the same general area for a length of time? The same sort of ambiguity was found in the other forms of stereotypy defined. Indeed, only one author provided a quantitative definition of what constituted stereotypy. Pacing was defined as "more than three traverses of a definite path, such as the back wall of the exhibit, or the area around the edge of the pool. Score as move until three traverses completed" [Forthman et al., 1992]. Authors need to present more of this sort of detail so we can be sure that we are comparing similar stereotypic behaviors when evaluating results across studies. In addition to improved definitions, it will also be useful for authors to report more detailed quantification in the results. Mason [1993] advocates that the morphology be measured in terms of body parts used, bout length, and rapidity of repetition. Studies on "loose stereotypies" (predictable subroutines embedded in more variable patterns, often evident over longer time scales) are also important. With stereotypies, and their plasticity in the face of enrichment, the devil (i.e., the underlying motivation) may well be in the detail.

In addition to reporting the details of enrichment and stereotypy, understanding can be deepened by the inclusion of other measures of wellbeing, both behavioral and hormonal [for a review see Young, 2003]. Corticoids are the most common hormonal measure of wellbeing, but disappointingly the pioneering work in this area by Carlstead et al. [1993], who found a reduction in corticoids after enrichment, has not been seized upon by the zoo community. Of 41 studies in our sample, this study is the only one that measured corticoids. Authors presented analyses of natural behaviors on 16 occasions and found a significant increase in response to enrichment in 13. In 14 of 19 reported analyses, activity levels were found to be higher post-enrichment. Enrichment was associated with a significant increase in behavioral diversity in all three studies that presented these data. Clearly, enrichment has diverse effects on behavioral and physiological profiles well beyond just stereotypies, and when measured, they offer additional insights into the relationship between enrichment and animal welfare. Although the validity of many of these behavioral indices for welfare is hotly debated [review in Veasy et al., 1996; Young, 2003], documenting the interrelationship among these variables will be instrumental to a better understanding of stereotypies and is an important goal of future work in this area.

General Discussion

Based on the information presented here, we can make a number of suggestions for guiding future research on stereotypic behavior and enrichment in zoos that we believe will help to shed more light on some of these questions in the future. First, we stress that although we conducted a meta-analysis looking for types of enrichment that work better than others do, we do not advocate a one-size-fits-all approach to enrichment. We all have experiences with the idiosyncrasies of animals even within a species, where enrichment works well for some individuals but not others. We should not, however, let this natural variation prevent us from searching for larger general principles and global strategies that are more likely to work more of the time than alternatives, then looking further for factors influencing variation in enrichment efficacy (e.g., species, age, and gender).

Perhaps the most urgent need is to increase the sample size of animals included in these studies. Most often this will mean more multi-institutional studies. Of 41 studies in our sample, only three were multi-institutional studies and one of these was a questionnaire survey to 49 zoos. Questionnaire surveys are a valuable way of collecting information on large sample sizes at multiple institutions, but they cannot substitute for direct observation. Correlative studies [e.g., Carlstead, 1998] that measure the relationship between observed behaviors and environmental variables are an effective way of learning about the variables that influence stereotypic behavior and wellbeing. A variant of the correlational study is to capitalize on "accidental" experiments, e.g., studying animals that are housed in different environments. Of course the gold standard is the controlled experimental approach. This is challenging in the zoo context but not impossible, as Wielebnowski et al. [2002] demonstrated in an exemplary study; hopefully, others will now follow.

Another problem that our review revealed was the relatively poor level of description of enrichment and stereotypy. Terms such as "scattered food," "novel objects," and "naturalistic environment" could cover a wide variety of different objects and activities. If we are to gain a deeper understanding of which kinds of enrichment work best, we need to know exactly what was tried. Mench [1998] and Sambrook and Buchanan-Smith [1997] have advocated strongly the need for careful description of enrichment items and how different properties influence efficacy. In an empirical attempt to follow this advice, Swaisgood, White et al. [in press] quantified 13 properties of five different enrichments and tested how these mapped onto the behavioral profiles associated with enrichments and their effects on wellbeing. The need for details also applies to descriptions of stereotypic behavior, which are rarely described in much detail (for a study that does attempt this see Lewis [2000]), a plea that Mason made back in 1993 [Mason, 1993]. Different kinds of stereotypic behavior may respond to treatments differently, but if stereotypies are not described quantitatively this information will be hard to gain.

In the course of our review we also noted that many zoo studies lacked the strong theoretical framework and hypothesis-testing approach that is more prevalent in the farm and lab animal research. Various competing theories were often alluded to, but not specifically tested. Moreover, conceptual overlap between theories meant that clear distinctions could not be made. To aid in the advancement of a theoretical framework to guide enrichment and to enable the evaluation of competing theories, it is clear that varying theories need to be articulated in a way that allows mutually exclusive alternative predictions to be generated. A systematic evaluation of all the theories advanced to explain stereotypy development and enrichment efficacy will do much to revitalize future research and understanding in this field. This will require the amalgamation of some overlapping theories and the establishment of sharper conceptual boundaries between those remaining. Once done, zoo researchers need to take up the challenge to test these models explicitly. Only such tests will allow us to understand the motivations underlying the performance of stereotypies and why some enrichments are better at addressing these motivations than others.

If these goals can be realized, we foresee a future where, in 10–15 years, we can conduct a literature meta-analysis and determine not only what theories best explain observed patterns of stereotypy performance and enrichment success, but have much to say about what kinds of enrichment strategies do and don't work, and under what circumstances. It has been said that the goal of science is prediction, and we believe this should be the goal of enrichment research in zoos, namely, to:

1. Predict when stereotypies are most likely to develop (i.e., what species in what kinds of enclosures, subject to what kind of husbandry practices?)
2. Predict what forms of enrichment are required to reduce and abolish stereotypies across various taxonomic units and captive circumstances.

Undoubtedly, this goal will never be realized fully, but the closer we come, the better will be our arsenal for improving wellbeing for captive animals. We hope that enrichment will continue to be practiced at ever increasing levels of enthusiasm and sophistication and that more studies will be conducted that take into account some of the problems revealed by our meta-analysis.

Conclusions

1. Zoo researchers tackle stereotypies with diverse and creative enrichment strategies, resulting in a significant reduction of stereotypy performance in 53% of the studies included in our literature meta-analysis.
2. The value of zoo research on enrichment and stereotypies can be increased by improved experimental design, statistics, and descriptive details.
3. To understand underlying motivations for stereotypy performance and the effects of enrichment, zoo researchers need to test hypotheses stemming from predictions of specific theoretical models.
4. Although this literature analysis could not draw definitive conclusions on the effects of taxonomic position on enrichment and stereotypies, zoo researchers can make a major contribution to the field by expanding the phylogenetic range of research subjects.
5. The ultimate goal of enrichment-stereotypy research should be to predict when stereotypies will develop and which enrichments will abolish them.

ACKNOWLEDGMENTS

We thank L. Brudney, M. Owen, and S. Wong for their assistance with the literature review and M. Bashaw, G. Mason, M. Owen, and J. Rushen for their instructive reviews of an earlier version of this article.

REFERENCES

Bart J, Fligner MA, Notz WI. 1998. Sampling and statistical methods for behavioral ecologists. Cambridge, UK: Cambridge University Press.

Bashaw MJ, Tarou LR, Maki TS, Maple TL. 2001. A survey assessment of variables related to stereotypy in captive giraffe and okapi. Appl Anim Behav Sci 73:235–47.

Carlstead K, Shepherdson DJ. 1994. Effects of environmental enrichment on reproduction. Zoo Biol 13:447–58.

Carlstead K. 1998. Determining the causes of stereotypic behaviors in zoo carnivores: toward appropriate enrichment strategies. In: Shepherdson DJ, Mellen JD, Hutchins M, editors. Second nature: environmental enrichment for captive animals. Washington: Smithsonian Institution Press. p 172–83.

Clubb R, Mason G. 2003. Captivity effects on wide-ranging carnivores. Nature 425:473–74.

Forthman DL, Elder SD, Bakeman R, Kurkowski TW, Noble CC, Winslow SW. 1992. Effects of feeding enrichment on behavior of three species of captive bears. Zoo Biol 11:187–95.

Hurlbert SH. 1984. Pseudoreplication and the design of ecological field experiments. Ecol Monogr 54:187–211.

Laule G, Desmond T. 1998. Positive reinforcement training as an enrichment strategy. In: Shepherdson DJ, Mellen JD, Hutchins M, editors. Second nature: environmental enrichment for captive animals. Washington, DC: Smithsonian Institution Press. p 302–13.

Lewis K. 2000. Individual differences and environmental factors influencing stereotypic behavior in captive polar bears [MSc]: Portland State University. 117 p.

Lipsey MW, Wilson DB. 2001. Practical meta-analysis. Thousand Oaks, California: Sage Publications. 247 p.

Machlis L, Dodd PW, Fentress JC. 1985. The pooling fallacy: problems arising when individuals contribute more than one observation to the data set. Anim Behav 68:201–14.

Magin CD, Johnson TH, Groombridge B, Jenkins M, Smith H. 1994. Species extinctions, endangerment and captive breeding. In: Olney PJS, Mace GM, Feistner ATC, editors. Creative conservation: interactive management of wild and captive animals. London: Chapman & Hall. p 4–31.

Mason G, Mendl M. 1993. Why is there no simple way of measuring animal welfare? Anim Welfare 2:301–19.

Mason GJ. 1991. Sterotypies: a critical review. Anim Behav 41:1015–37.

Mason GJ. 1993. Forms of stereotypic behaviour. In: Lawrence A, Rushen J, editors. Stereotypic animal behaviour. Wallingford, UK: CAB International. p 7–40.

Mason GJ, Latham N. 2004. Can't stop, won't stop: is stereotypy a reliable animal welfare indicator. Anim Welfare 13(Suppl):57–69.

Mellen J, MacPhee MS. 2001. Philosophy of environmental enrichment: past, present, and future. Zoo Biol 20:211–26.

Mench JA. 1998. Environmental enrichment and the importance of exploratory behavior. In: Shepherdson DJ, Mellen JD, Hutchins M, editors. Second nature: environmental enrichment for captive animals. Washington: Smithsonian Institution Press. p 30–46.

Nakagawa S. 2004. A farewell to Bonferroni: the problems of low statistical power and publication bias. Behav Ecol 15:1044–5.

Owen M, Swaisgood RR, Czekala NM, Steinman K, Lindburg DG. 2004. Monitoring stress in captive giant pandas (*Ailuropoda melanoleuca*): behavioral and hormonal responses to ambient noise. Zoo Biol 23:147–64.

Quinn GP, Keough MJ. 2002. Experimental design and data analysis for biologists. Cambridge: Cambridge University Press.

Rushen J, Lawrence AB, Terlouw EMC. 1993. The motivational basis of stereotypies. In: Lawrence A, Rushen J, editors. Stereotypic animal behaviour. Wallingford, UK: CAB International. p 41–64.

Sambrook TD, Buchanan-Smith HM. 1997. Control and complexity in novel object enrichment. Anim Welfare 6:207–16.

Saudargas RA, Drummer LC. 1996. Single subject (small N) research designs and zoo research. Zoo Biol 15:173–81.

Shepherdson D. 1998. Tracing the path of environmental enrichment in zoos. In: Shepherdson D, Mellen J, Hutchins M, editors. Second nature: environmental enrichment for captive animals. Washington: Smithsonian Institution Press. p 1–12.

Shepherdson D. 2001. Environmental enrichment. In: Bell CE, editor. Encyclopedia of the world's zoos. Chicago: Fitzroy Dearborn Publishers. p 421–4.

Shepherdson DJ. 1994. The role of environmental enrichment in the captive breeding and reintroduction of endangered species. In: Mace G, Olney PJS, Feistner A, editors. Creative conservation: interactive management of wild and captive animals. London: Chapman & Hall. p 167–77.

Swaisgood RR, Zhou X, Zhang G, Lindburg DG, Zhang H. 2003. Application of behavioral knowledge to giant panda conservation. Int J Comp Psychol 16:65–84.

Swaisgood RR, Owen MA, Czekala NM, Hawk K, Kinoshita R, Tang J. In press. Evaluating stress and well being in giant pandas: a system for monitoring. In: Wildt DE, Zhang AJ, Zhang H, Janssen D, Ellis S, editors. Giant pandas: biology, veterinary medicine and management. Cambridge: Cambridge University Press.

Swaisgood RR, Shepherdson DJ. In press. Environmental enrichment as a strategy for mitigating stereotypies in zoo animals: a literature review and meta-analysis. In: Mason GJ, Rushen J, editors. Stereotypic animal behaviour: fundamentals and applications to welfare. 2nd ed. Wallingford, UK: CAB International.

Swaisgood RR, White AM, Zhou X, Zhang G, Lindburg DG. In press. How do giant pandas respond to varying properties of enrichments? A comparison of behavioral profiles among five enrichment items. J Comp Psychol.

Swaisgood RR, White AM, Zhou X, Zhang H, Zhang G, Wei R, Hare VJ, Tepper EM, Lindburg DG. 2001. A quantitative assessment of the efficacy of an environmental enrichment programme for giant pandas. Animal Behaviour 61: 447–57.

Swaisgood RR, Zhang G, Zhou X, Zhang H. In press. The science of behavioral management: creating biologically relevant living environments in captivity. In: Wildt DE, Zhang AJ, Zhang H, Janssen D, Ellis S, editors. Giant pandas: biology, veterinary medicine and management. Cambridge: Cambridge University Press.

Veasy JS, Waran NK, Young RJ. 1996. On comparing the behaviour of zoo housed animals with wild conspecifics as a welfare indicator. Anim Welfare 5:13–24.

Wielebnowski N, Ziegler K, Wildt DE, Lukas J, Brown JL. 2002. Impact of social management on reproductive, adrenal and behavioural activity in the cheetah (*Acinonyx jubatus*). Anim Conserv 5:291–301.

Woodroffe R, Ginsberg JR. 2000. Ranging behaviour and vulnerability to extinction in carnivores. In: Gosling LM, Sutherland WJ, editors. Behaviour and conservation. Cambridge: Cambridge University Press. p 125–40.

Young RJ. 2003. Environmental enrichment for captive animals. Oxford, UK: Blackwell Science. 228 p.