Human visual identification of individual Andean bears Tremarctos ornatus

Authors: Horn, Russell C. Van, Zug, Becky, LaCombe, Corrin, Velez-Liendo, Ximena, and Paisley, Susanna

Source: Wildlife Biology, 20(5) : 291-299

Published By: Nordic Board for Wildlife Research

URL: https://doi.org/10.2981/wlb.00023

Wildlife Biology 20: 291–299, 2014
doi: 10.2981/wlb.00023

Subject Editor: Al Glen. Editor-in-Chief: Ilse Storch. Accepted 4 July 2014

Human visual identification of individual Andean bears *Tremarctos ornatus*

Russell C. Van Horn, Becky Zug, Corrin LaCombe, Ximena Velez-Liendo and Susanna Paisley

R. C. Van Horn (rvanhorn@sandiegozoo.org) (orcid.org/0000-0002-5789-8822) and C. LaCombe, Inst. for Conservation Research, San Diego Zoo Global, PO Box 120551, San Diego, CA 92112-0551, USA. – B. Zug, Nelson Institute for Environmental Studies, Univ. of Wisconsin – Madison, WI 53706, USA. – X. Velez-Liendo, Centro de Biodiversidad y Genética, Univ. Mayor de San Simon, Cochabamba, Bolivia. – S. Paisley, Durrell Inst. of Conservation and Ecology, Univ. of Kent, Canterbury, Kent, CT2 7NR, UK

It is often challenging to use invasive methods of individual animal identification for population estimation, demographic analyses, and other ecological and behavioral analyses focused on individual-level processes. Recent improvements in camera traps make it possible to collect many photographic samples yet most investigators either leap from photographic sampling to assignment of individual identity without considering identification errors, or else to avoid those errors they develop computerized methods that produce accurate data with the unintended cost of excluding participation by local citizens. To assess human ability to visually identify Andean bears *Tremarctos ornatus* from their pelage markings we used surveys and experimental testing of 381 observers viewing photographs of 70 Andean bears of known identity. Neither observer experience nor confidence predicted their initial success rate at identifying individuals. However, after gaining experience observers were able to achieve an average success at identifying adult bears of 73.2%, and brief simple training further improved the ability of observers such that 24.8% of them achieved 100% success. Interestingly, observers who were initially more likely to falsely identify two photos of the same bear as two different bears than vice versa were likely to continue making errors and their bias became stronger, not weaker. Such biases would lead to inaccurate population estimates, invalid assessments of the bears involved in conflict situations, and underestimates of bear movements. We thus illustrate that in some systems accurate data on individual identity can be generated without the use of computerized algorithms, allowing for community engagement and citizen science. In addition, we show that when using observers to collect data on animal identity it is important to consider not only the overall frequency of observer error, but also observer biases and error types, which are rarely reported in field studies.

Visual identification of individuals, either from direct sightings or from imagery such as camera trap photos, benefits many lines of research, from studies of development (Swanson et al. 2013) and behavioral ecology (Charpentier et al. 2008) to population estimation (Ngoprasert et al. 2012) and other analyses with direct conservation implications (reviewed by McGregor and Peake 1998). Thus, for decades natural markings have been evaluated for the noninvasive identification of individuals of various species (Pennycuick 1978, Jarman et al. 1989). However, although natural markings never produce perfect individual identification (Pennycuick 1978, Jarman et al. 1989), identification errors are often unreported (but see Stevick et al. 2001, Frasier et al. 2009). There are two types of identification errors: false matches (i.e. incorrectly identifying images from multiple individuals as images from one), and false mismatches (i.e. incorrectly identifying multiple images of the same individual as images from multiple individuals). These two error types may skew subsequent analyses and conclusions differently. For example, false matches may lead to underestimates of population size, while false mismatches can lead to overestimates of population size (Stevick et al. 2001, Hastings et al. 2008, Goswami et al. 2012). Thus, to reach valid conclusions it is critical that researchers quantify and characterize identification errors (Yoshizaki et al. 2009). Error rates may differ with training and experience (Stander et al. 1997, Diefenbach et al. 2003, Schofield et al. 2008), but experience does not automatically produce low error rates (Diefenbach et al. 2003, Patton and Jones 2008, Evans et al. 2009) even if the observer is confident in their own ability to identify the target species (De Angelo et al. 2010). Thus, regardless of observer experience or confidence, observer ability must be quantified. Actual errors of identification cannot be determined solely from non-invasive photos of wild individuals (Ríos-Uzeda et al. 2007, Bashir et al. 2013), but accuracy and precision can be assessed by testing identification assigned blindly to known (i.e. captive) individuals (Higashide et al. 2012).

Variation in natural markings has been used to identify individuals of some bear species (Noyce et al. 2001, Higashide et al. 2012, Ngoprasert et al. 2012). Because the markings on the face, throat and neck of Andean bears

Tremarctos ornatus have been thought to differ among individuals (Thomas 1902, Hornaday 1911), some researchers have begun using them to assign individual identity to bears in photos from camera traps (Ríos-Uzeda et al. 2007, Zug 2009, Jones 2010). However, these methods have not been thoroughly tested. Although Roth (1964) and Eck (1969) described variation in markings among captives, neither examined many bears ($n = 19$, $n = 5$ respectively). In addition, although the markings are present from birth (Saporiti 1949, Roth 1964, Dathe 1967, Eck 1969), their permanence is untested and it is unknown if any changes in the markings would affect individual identification. We suspect that the highest rate of apparent disappearance of 'known' wild Andean bears will occur during subadulthood due to increased mortality as cubs become independent of their mothers and due to primary dispersal; neither of these processes has been studied in this species. If the markings of cubs change enough during maturation to confound individual identification then this will further increase the apparent disappearance of known individuals, inflating estimates of mortality and dispersal. So, although comparing images of cubs and adults may be a less common task for researchers than comparing images of adults, the former task warrants special consideration.

Although some researchers have developed project-specific protocols for identifying individual Andean bears (Zug 2009, Jones 2010), methods are not standardized across studies (Garshelis 2011) and there are typically no estimates of error (Goldstein and Márquez 2004, Garshelis 2011), making it pointless to compare results across studies. For example, it is possible that two studies might produce similar estimates of bear density even though one study was conducted in an area with a lower density of bears, simply because there was a higher and unmeasured rate of false mismatches in that data set.

There are numerous methods for computer-assisted or automated identification of individuals of several mammal species (Kelly 2001, Karlsson et al. 2005, Hiby et al. 2009, Goswami et al. 2012). These methods allow for rapid identification, which can otherwise be labor-intensive with many individuals or photos, and they may achieve better accuracy than manual identification when identification is challenging (Kelly 2001). Those two advantages are likely not relevant for Andean bears, whose markings appear different, and which are thought to live at low densities (but see Garshelis 2011). In addition, manual identification harnesses the ability of humans to correct for image variation due to occlusions and shadows, which remains challenging for imaging processing software (Allen et al. 2011). If manual identification of individuals achieves high accuracy and good precision, this would avoid three key disadvantages of computerized identification in research and conservation of Andean bears: the development of such methods requires expertise and funds not often available to field programs, they preclude field identification of individual bears during direct observations, and they require technicians to be computer literate, excluding most local residents.

Conservation science is concerned not just with knowledge, but also with conservation impact, which may be enhanced through local participation (Danielsen et al. 2007). Local attitudes have important effects on conservation of Andean bears (Velez-Liendo 2005), especially because the bears come into conflict with humans (Treves et al. 2006), local communities are active inside many protected areas (Naughton-Treves et al. 2006), and most tropical forests lie outside of protected areas (Chazdon et al. 2009). Engaging local people in research can capitalize on their knowledge and skills (Stander et al. 1997, Zuercher et al. 2003, Sharma et al. 2005) and lead to better communication and better conservation outcomes (Peyton 1989, Byers 1999). We therefore assess the permanence of Andean bear markings and whether observer characteristics, experience, and training affect their performance, to explore whether the use of minimal technology may produce high quality data and enhance their potential conservation impact.

Material and methods

We collected portraits of captive Andean bears of known identity and age from zoo personnel and field researchers in North America, Europe and South America. We discarded images with poor resolution, lighting or clarity, but did not reject images with extreme camera angles. The images we retained illustrated a wide range of facial markings, ranging from no facial markings to broad full circles around both eyes and depigmentation or 'grizzling' across much of the rest of the face. To assess the permanence of facial markings we visually compared across time the markings of the 24 bears for which we had photos as both cubs and adults. We also looked for evidence of grizzling in the photos of all 64 known-age bears in our sample.

To assess humans' ability to identify individual Andean bears we first created online surveys in English and Spanish, the common language across the species' range, using 65 different photographs of 39 known bears. To evaluate participants with a variety of personal and professional backgrounds we solicited participation in the survey by emails to colleagues, peers, and personal contacts, as well as through an announcement in the International Bear News (Paisley et al. 2010). Participants reviewed 21 pairs of images: six pairs of images of adults spanning up to 13 years from the same bear, and 15 pairs that included one photo of a cub and one photo of an adult ('cub–adult' pairs) spanning up to 23 years from the same bear. For each pair, participants responded to the question "Are the photos above of the same bear?" with one of three responses: "yes", "no", and "unable to determine". This task mimics some of the identification tasks faced during field research, such as when a bear under observation in a corn field must be compared to a photo taken earlier during a similar event, or when a recently retrieved camera trap photo must be compared to a camera trap photo taken at another location. To examine whether success was affected by participants' background or personal characteristics we collected information on participant sex, age and experience working with bears, Andean bears, or with visual identification of individuals of any wildlife species. To remove potential biases caused by poorly motivated participants, or by missing data, we only analyzed data from participants who answered at least 15 of the 21 questions. We measured participant performance as the proportion of responses that were correct. The average age of the

120 online participants (50 men, 70 women) who answered at least 15 of the 21 questions was 36.4 ± 12.1 years. Nineteen participants (16%) had experience working with bears but not *Tremarctos*, 10 participants (8.3%) had experience working with *Tremarctos*, and 68 participants (56.7%) had experience with visual identification of individual wild animals.

We assessed whether participant performance differed from random and if it differed between adult pairs and cub–adult pairs, and if it differed depending on whether markings changed during maturation, using t-tests and paired t-tests. We then used an information theoretic approach (Burnham and Anderson 2002) to compare all possible models for online participant performance, built with data on participant characteristics (i.e. sex, age, experience with bears, experience with *Tremarctos*, and experience with visual identification of individuals). We did not include interaction terms in potential models, and we used AIC_c as a key criterion for model comparison (Burnham and Anderson 2002).

We also assessed the effects of experience and simple training through experimental testing of staff, volunteers and visitors at San Diego Zoo's Inst. for Conservation Research, using 94 photographs of 55 known bears. We implemented experimental sessions to groups in a pre-post test study design with 'experience' groups (E groups) and 'experience and training' groups (E–T groups, Oppenheim 1992). All sessions were less than 30 min long and began with a 5-min overview of Andean bear ecology and conservation, conservation research, the purpose of the session and instructions on how to enter their responses into our Classroom Performance System (ver. 1.50.063), an interactive system that allows participants to use remote controls to record their answers. A portrait of an Andean bear was shown while it was explained that individual Andean bears might be recognized by their unique markings, including muzzle freckles. We told participants that they would review pairs of images and that for each pair they would have 20 s to respond to the question "Are these the same bears?" with one of three responses: "Yes, these are the same bears", "No, these are not the same bears", and "I am not sure if these are the same bears". We stressed that each possible answer, including uncertainty, was viable. We asked participants to compare as many features of the markings as possible, excluding nose color, and using caution when interpreting photos with extremes of lighting or orientation.

Within each session we displayed 60 pairs of images sequentially, in the same order in both treatments. In both treatments the first 15 comparisons were the pre-test, while the last 15 comparisons were the post-test. Thus, the treatments differed only in the presentation of the middle 30 comparisons. In the E treatment the transition between the three sections was seamless with a 3-min break at question 34. Thus, any change in E participant performance, as measured by a comparison between the first 15 comparisons ('initial' performance) and the last 15 comparisons ('final' performance), would be due to viewing additional images of Andean bears. In the E–T treatment, while viewing the middle 30 pairs of images the participants received simple training. After viewing each pair for 20 s a group discussion was held in which participants shared their answers and reasoning aloud. The instructor then revealed the correct answer and highlighted meaningful comparisons between images. Three points of comparison were illustrated for pairs of images showing the same bear (matches) and 2–3 points of comparison were highlighted for pairs of images showing different bears (mismatches). If it was not possible to determine if a pair of images displayed the same or different bears, the instructor illustrated this (e.g. differences in image angle). We conducted sessions as competitions in which the highest score on pre- and post-tests earned a small non-monetary prize.

We used the same sequences of the same images in both treatments. Although the pairs of images differed between the pre- and post-tests, each test contained six matches, six mismatches and three comparisons that were impossible to identify as a match or mismatch. Both the pre- and post-tests were composed of six pairs of adult images and nine cub–adult pairs. Among the middle 30 comparisons (14 cub–adult pairs and 16 adult pairs) there were 14 matches, 14 mismatches and two comparisons that were impossible to identify as matches or mismatches.

We collected data on four participant characteristics: sex, age (four categories), self-perceived ability to identify Andean bears five categories), and the frequency with which the participant observed wildlife (five categories). We ensured confidentiality and anonymity by collecting no personal identifying information. Because we had no prior knowledge of how well participants might succeed, and which participant characteristics might affect initial performance, we compared participant performance to random and then used an information theoretic approach to compare all possible models built with data on participant characteristics. We used a similar approach to investigate changes in measurements of interest (e.g. success identifying individuals), comparing models built on 'treatment' with those also containing 'group' nested within 'treatment'; we did not include interaction terms in potential models and we used AIC_c as a key criterion for model comparison. To assess whether participant success with cub–adult pairs might respond differently than participant success with adult images, we conducted statistical analyses of identification separately for cub–adult pairs, and for adult pairs.

We could not predict whether matches or mismatches would be more common among observers' errors. For some species false matches have been found less common than false mismatches (Stevick et al. 2001) while in other systems the reverse was true (Patton and Jones 2008) or the bias was small (Frasier et al. 2009). We therefore assumed that participant errors would be equally divided among false matches and false mismatches, yielding a 1:1 ratio of false matches:mismatches. We analyzed ratios of error types as we did participant success, to investigate participants' initial ratio of false matches:mismatches, and the change in error type ratio across participants and treatments. Three hundred twenty people participated in experimental tests. Technical difficulties with the first two groups of participants (n = 55) and incomplete data from a few participants (n = 4) led us to discard those data, leaving data from 136 E participants (seven groups) and 125 E–T participants (seven groups).

Unless otherwise noted all quantities are expressed as $\bar{x} \pm$ SD; statistical significance refers to two-tailed $p = 0.05$. Statistical analyses were conducted in JMP ver. 9.0.3

Results

The facial markings did not change during maturation in most Andean bears for which we have photos as both cubs and adults (66.7%, n = 24). However, the markings of some cubs became thinner or less obvious between approximately the first and second year of life. This appears to occur symmetrically on both the left and right sides of the face and if a cub's marking was thin or faint it may not be apparent when the bear is an adult (Fig. 1). If an Andean bear survives long enough its appearance may change again through grizzling; based on photos of 25 grizzled Andean bears it appears that grizzling first appears around the eyes and pre-existing markings and can eventually spread across the entire face (Fig. 2). The youngest bear for which we have evidence of grizzling was eight years old; most bears photographed when over 10 years old showed some grizzling (80%, n = 30).

Figure 1. The markings of most individuals do not change, but some become less pronounced during the first 2–3 years of life. If a marking was thin or faint on a cub, it may not be apparent when that individual is an adult; this occurs symmetrically. (a) 'Joaquin' when less than a year old. (b) 'Joaquin' when 10 years old. (c) 'Diamond' when less than a year old. (d) 'Diamond' when 19 years old. (e) 'BJ' when less than a year old. (f) 'BJ' when 3 years old.

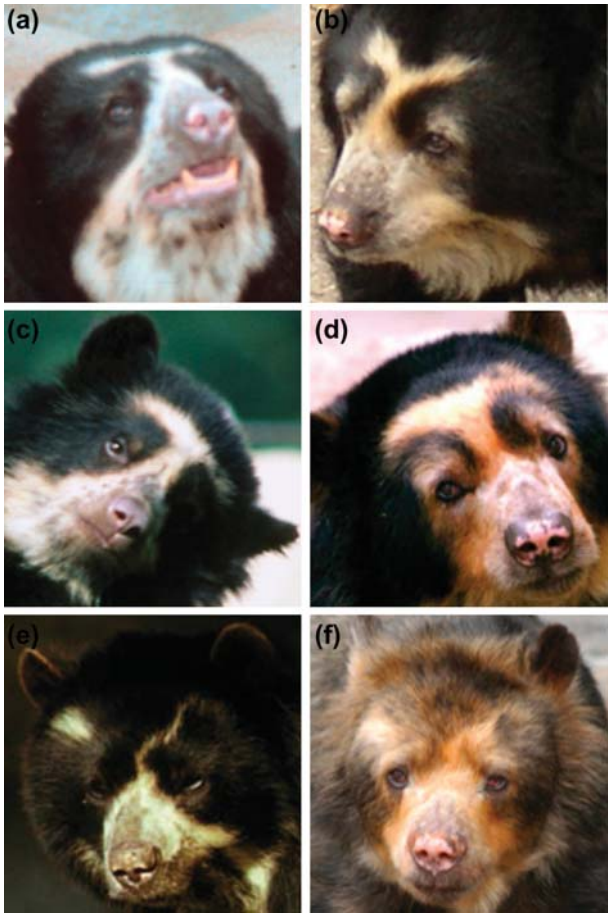

Figure 2. Grizzling of the face begins around the eyes and pre-existing markings and spreads variably. (a) 'Roxanne' when 11 years old. (b) 'Roxanne' when 24 years old. (c) 'Willie' when 6 months old. (d) 'Willie' when 24 years old. (e) 'Chris' when 13 years old. (f) 'Chris' when 30 years old.

The average success of online participants at identifying cub–adult pairs differed depending on whether the cub's markings thinned during maturation. Participants correctly answered on average 58.6% ($\pm$ 15.7%) of questions in which cub's markings did not change (n = 10), which was not as expected at random (i.e. 50%; Fig. 3, $t = 6.019$, DF = 119, $p < 0.001$). They correctly answered on average 30.1% ($\pm$ 21.1%) of questions with cub–adult pairs in which the markings did change (n = 5), which was also not random ($t = -10.32$, DF = 119, $p < 0.001$) and different from their average success with cub–adult pairs in which the markings did not change (t-ratio = –11.808, DF = 119, $p < 0.001$). The best model for participant average success with cub–adult pairs in which markings did not change (i.e. 'age') was not well supported by the data ($R^2 = 0.005$, DF = 119, F-ratio = 1.124, $p = 0.206$). Similarly, the best model for average online success with cub–adult pairs in which markings did change (i.e. 'age' and 'experience with bears') was also not well supported by the data ($R^2 = 0.055$, DF = 119, F-ratio = 2.244, $p = 0.087$). Thus, no participant characteristic had a meaningful impact on how well participants in the online survey could identify cub–adult pairs, whether or not cubs' markings changed.

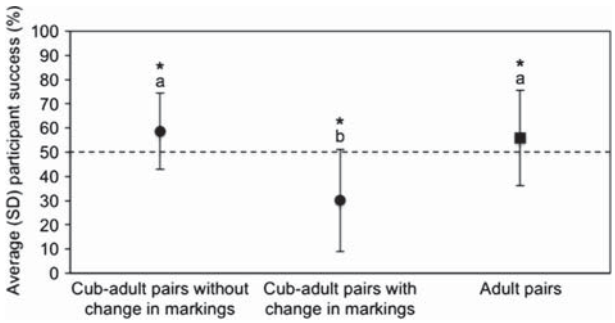

Figure 3. Changes in markings affected average online identification success. The average online success of 120 participants was highest for adult pairs and cub–adult pairs in which the markings did not change, and better than expected at random in both cases. However, the average online success of participants was lower for cub–adult pairs in which the markings did change; participants performed worse than expected at random. Asterisks indicate significant differences from random and letters indicate statistically significant groupings of averages.

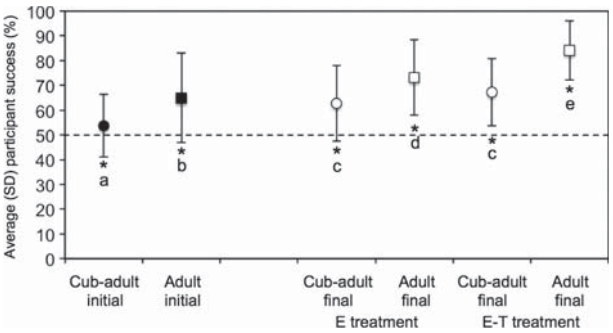

Figure 4. During experimental testing the initial average participant success at identifying individuals was better than expected at random for both cub–adult pairs and for adult pairs. This average success improved regardless of treatment, indicating that simple exposure to more images improved success. Simple training did provide marginally better average performance than experience alone. Asterisks indicate significant differences from random and letters indicate statistically significant groupings of averages.

The average success of online participants at identifying adult pairs of images (55.9 ± 19.7%) was better than expected at random (Fig. 3, $t = 2.852$, $DF = 119$, $p = 0.005$). Although it was not statistically different than these participants' average success at identifying cub–adult images of cubs whose markings did not change (t-ratio = −1.769, DF = 119, $p = 0.08$), average success identifying adult pairs was significantly different than participant success with cub–adult images of cubs whose markings did change (t-ratio = –9.431, DF = 119, $p < 0.001$). The best model for average online success with adult images (i.e. 'age') was not well supported by the data ($R^2 = 0.016$, DF = 119, F-ratio = 1.94, $p = 0.166$). Thus, average participant success online differed between adult pairs and cub–adult pairs only if the markings changed during maturation, and average online success with adult pairs was not explained by any measured participant characteristic. However, averages do not reveal the success of the best participants, which is relevant for assessing potential to produce high-quality data. Although the average online success rate with adult images was far below 100%, five online participants (4.2%) did successfully rate all adult pairs.

In experimental testing, participant initial success in identifying cub–adult image pairs was slightly better than expected at random (Fig. 4, 53.7 ± 12.6%, $n = 261$, $t = 4.734$, $p < 0.001$); none of the illustrated cubs' markings changed during maturation. The best model for initial success in identifying cub–adult image pairs (i.e. 'participant sex') was not well supported by the data ($R^2 = 0.002$, DF = 260, F-ratio = 0.617, $p = 0.433$). Thus, no participant characteristic had a meaningful impact on their initial success in identifying cub–adult image pairs. After treatment, during which six of eight cub–adult pairs illustrated changes in markings during maturation, participants were better at identifying cub–adult image pairs (i.e. the % change in performance was not 0; Fig. 4, 11.2 ± 17.8%, $n = 261$, $t = 10.14$, $p < 0.001$); the cub–adult pairs shown after treatment did not illustrate changes in markings. The best model for the improvement in identification of cub–adult image pairs after treatment (i.e. 'treatment') explained almost none of the variation in the data ($R^2 = 0.012$, DF = 260, F-ratio = 3.19, $p = 0.075$). Thus, training did not obviously improve the ability of participants to identify cub–adult image pairs more than did simple experience; in the end the E participants could identify 62.7% (± 15.3%) of cub–adult image pairs and the E–T participants could identify 67.2% (± 13.5%) of cub–adult image pairs. Motivated exposure to additional images of bears improved participant success at identifying cub–adult pairs.

The initial performance of experimental participants at identifying adult image pairs was better than expected at random (Fig. 4, 64.9 ± 18.1%, $n = 261$, $t = 13.316$, $p < 0.001$) and it was better than these participants' initial performance at identifying cub–adult pairs (t-ratio = –8.673, DF = 260, $p < 0.001$). The best model for the initial success of experimental participants in identifying adult image pairs (i.e. 'treatment') was not well supported by the data ($R^2 = 0.014$, DF = 260, F-ratio = 3.559, $p = 0.06$). Thus, there was no indication that any participant characteristic had a meaningful impact on the initial success in identifying adult image pairs. After treatment, participants were better at identifying adult image pairs (i.e. the % change in performance was not 0; Fig. 4, 13.5 ± 21.1%, $n = 261$, $t = 10.35$, $p < 0.001$). The best model for the improvement in identification of adult image adult pairs (i.e. 'treatment') explained little of the variation in the data ($R^2 = 0.026$, DF = 260, F-ratio = 6.843, $p < 0.001$). Thus, training did not obviously improve the average ability to identify adult image pairs more than did simple experience; in the end, on average E participants could identify 73.2% (± 15.2) of adult image pairs and E–T participants could identify 84.1% (± 11.9) of adult image pairs. In other words, motivated exposure to additional images of bears, but not simple training, improved average participant success at identifying adult pairs. Participants in the E treatment and in the E–T treatment were both able to identify on average more adult image pairs than cub–adult pairs (t-ratio = −6.803, DF = 135, $p < 0.001$ and t-ratio = −10.276, DF = 124, $p < 0.001$, respectively). Before treatment, 3.7% of E participants (5 of 136) successfully

identified all pairs of adult images; after treatment, 7.4% (10 of 136) of E participants did so, illustrating that their experience identifying bears did not significantly increase the proportion of participants that were entirely successful (e.g. $\chi^2 = 1.76$, DF = 1, $p = 0.184$). Before treatment, 4.8% of E–T participants (6 of 125) successfully identified all pairs of adult images; after treatment, 24.8% (31 of 125) of the E–T participants did so, revealing that simple training made it more likely that some participants correctly identified all pairs of adult images ($\chi^2 = 19.83$, DF = 1, $p < 0.001$). In fact, the proportion of participants responding perfectly was greater after simple training than after motivated exposure to images ($\chi^2 = 14.97$, DF = 1, $p < 0.001$).

Were errors in individual identification of adult image pairs equally divided among false matches and false mismatches? Experimental participants initially had a strong bias for false mismatches: the initial ratio of false matches to false mismatches was 0.44 ± 0.64 (n = 154, DF = 153, $t = -11.03$, $p < 0.001$). The best model for this bias (i.e. 'treatment') was not well supported by the data ($R^2 = 0.028$, DF = 153, F-ratio = 4.31, $p = 0.04$). Thus, there was no indication that any characteristic of these participants had a meaningful impact on the ratio of their error types when initially assessing adult images. Among participants who continued to make errors identifying adult image pairs, the final ratio of false matches to false mismatches was 0.22 ± 0.47, still not 1:1 (n = 173, DF = 172, $t = -21.65$, $p < 0.001$). This was a change of -0.17 ± 0.81 from the initial ratio, different from the change expected by chance (i.e. 0; n = 112, $t = -2.278$, $p = 0.025$): among those who continue to make errors the bias for false mismatches was stronger after treatment than before. The best model for the change in ratio of false matches to false mismatches included only 'treatment' and it explained virtually none of the variation in the data ($R^2 < 0.001$, DF = 111, F-ratio = 0.006, $p = 0.94$). Thus, neither treatment nor group nested within treatment had a meaningful impact on the change in the ratio of false matches to false mismatches, although the bias for false mismatches across pairs of adult images was stronger after either treatment.

Why did the bias for false mismatches increase after experience, or experience and training? Was it because the participants who continue to make errors were a biased subset of the overall pool of participants? Perhaps they had greater difficulty with the task from the beginning, in which case their initial success rate would be lower than the people who made no errors after treatment, or perhaps they had a stronger bias for false mismatches from the beginning. An AIC analysis of the full model set (n = 3) including as variables the participant's initial success rate across adult images, and the participant's initial ratio of false matches to false mismatches across adult images, revealed that the second explanation is better supported by the data. The best model for the ratio of false matches to false mismatches across adult images included as a predictor only the participant's original ratio of false matches to false mismatches across adult images ($R^2 = 0.6387$, DF = 111, F-ratio = 194.5, $p < 0.001$, $\varpi = 0.741$), and it was better supported than the other two models ($\Delta AIC_c > 2.0$ for the other two models). In other words, those participants who continued to misidentify adult images finished with a stronger bias for false mismatches than the overall pool of participants after beginning with a stronger bias for false mismatches.

Discussion

A bear's age affects individual identification. The oldest Andean bears may be quickly identified by grizzling, if wild bears live so long, and the markings of some cubs become thinner during maturation, which made it more difficult for our online participants to identify the cubs as adults. However, changes in the markings may not be the only characteristic that makes it difficult to identify cubs as adults. Although trained participants in the experimental tests were shown how the markings might change during maturation, they then still found it challenging to identify cubs as adults even when the markings had not changed. This suggests that great caution is needed when comparing images of cubs and adults: only observers with proven ability should make such comparisons. This is especially important because this identification challenge arises during a poorly understood life stage when various processes might cause an individual to appear to vanish from a population (e.g. dispersal).

We believe there are three reasons why participants in the online survey were unsuccessful at identifying individual bears, regardless of their prior experience with bears, or with *Tremarctos*. First, there were only 10 participants with any experience working with *Tremarctos*, and the type and duration of their experience varied. Second, because each participant with experience of *Tremarctos* was probably exposed only to bears in their own work, it was unlikely that any participant had seen as many different bears, and as much variation in markings, as in the images we presented. Third, although experience sometimes confers improved ability (Stander et al. 1997, Diefenbach et al. 2003, Schofield et al. 2008), this is not always the case (Diefenbach et al. 2003, Patton and Jones 2008, Evans et al. 2009); experience may have conferred overconfidence. Initial success in the experimental assessment was not only low, it was unrelated to participant confidence. This disconnect between self-perceived and true ability is not novel. Competent individuals may have an accurate perception of their performance, but individuals who are not competent at a task are sometimes unable to evaluate their own performance (Kruger and Dunning 1999, Dunning et al. 2003, De Angelo et al. 2010). Thus, experience and self-perceived ability to identify individuals do not guarantee data of any particular quality. Some people's performance in this somewhat subjective task improves with experience, as illustrated by our experimental data, but other people's performance does not.

Although simple training did not improve people's ability on average to identify individual Andean bears more than did viewing images of Andean bears, simple training yielded more trainees who were able to successfully identify all pairs of images. Many trainees could not, but we believe that additional training of motivated observers would magnify the beneficial effect we observed and allow for collection of accurate data. We suspect there is no one living alongside Andean bears who is an expert at identifying more than a few familiar individual bears, but training followed by assessment could improve the research value of some local residents who

already have additional research skills. Others have also seen that minimal training can sometimes circumvent the need for complex technological approaches (Patton and Jones 2008, Schofield et al. 2008, Jones 2010), although the use of image manipulation tools might improve identification success in some circumstances.

Due to their original bias, those participants who continued to make errors were twice as likely to make a false mismatch than a false match. False mismatches, which can lead to overestimates of population size (Stevick et al. 2001, Hastings et al. 2008, Goswami et al. 2012), are particularly troublesome given that the Andean bear is vulnerable to extinction (Goldstein et al. 2010). This suggests that some people will require additional training to achieve acceptable levels of success, or that they should not identify individual bears. To further safeguard against errors, we recommend that whenever possible, the individual identity of any bear be independently assigned by at least three observers of demonstrated ability, whether they are research staff or citizen scientists. We suggest that, as in Mendoza et al. (2011), each observer independently assign individual identity before reviewing the assignments made by the other observers, and finally reaching a consensus assignment of individual identity. The use of multiple observers should also help mitigate differences that may exist among observers in their initial willingness to assign individual identity rather than cautiously withholding judgment as to which animal has been detected (Mendoza et al. 2011). That level of caution may vary not only between observers but may also vary with sample size, such that as the sample size decreases the observer's level of caution may also decrease in a perhaps unconscious attempt to ensure sufficient analytical power. Although it is not obvious to us that variation in caution across observers or sample sizes will create a bias towards matches or mismatches, the impact of such errors will be larger when sample sizes are small, suggesting that evaluation of data quality becomes even more important as data become scarce.

With training, many observers are capable of success over 95%, which we believe should be the minimum level of identification success across studies. As has been suggested by others (Foster and Harmsen 2012), we urge investigators to quantify and report the rates at which they make errors, the types of errors they make, and the potential effects of those errors on their results. In addition, for at least some identification tasks, there is intra-observer variation over time (Bindemann et al. 2012). Thus, assessment of identification success and quantification of error rates should be an ongoing process and not just an evaluation of the efficacy of training. We suggest that images of 'known' identity be inserted amongst the photographs of 'unknown' identity at intervals that cannot be anticipated by the people being assessed; to assess observer performance with sufficient precision at least 20 images of 'known' identity would need to be included in such an evaluation (i.e. 19/20 = 95%). Images of 'known' identity may either be images of captive bears, as we have used, or they may be high quality images of wild bears that can be repeatedly and consistently assigned identity. As data are collected from the field it will likely be impossible to evaluate the true accuracy of individual identification since there may be no independent source of bear identity. However, even when independent measures of accuracy are not possible, researchers should estimate intra-observer consistency and inter-observer agreement as indicators of data reliability (Higashide et al. 2012, Ngoprasert et al. 2012). Some inter-observer agreement will arise from chance so we recommend the use of the kappa statistic instead of percent agreement (Forcada and Aguilar 2000, Watkins and Pacheco 2000, Viera and Garrett 2005).

We have illustrated that with proper training and assessment it should be possible to engage residents of local communities in the identification of individual wild Andean bears. We believe this will also be true for other species of non-social mammals living at low densities, allowing accurate data on individual identity to be generated through community engagement and citizen science if training and evaluation are sufficiently rigorous and multiple observers are involved. In addition, we have shown that investigators should consider not only the overall frequency of observer error after training, but also observer biases, which are rarely reported in field studies. The benefits, disadvantages, and socio–political dynamics of bear research and conservation vary widely across contexts so the decision to use local people to collect data on individual bear identity will need to be context-specific. Our hope is that by considering observer biases, and engaging local citizens in data collection whenever possible, researchers will not only generate reliable data and replicable results, they will also achieve better conservation outcomes.

Acknowledgements – We thank the volunteer participants in the online survey and in the experimental testing at San Diego Zoo's Inst. for Conservation Research, where we also thank Maggie Reinbold, Robin Keith and Samantha Young. We thank the following individuals for contributing photographs and data from captive Andean bears: Valerie Abbot, Mark Brayshaw (Durrell Wildlife Conservation Trust), Sander Hofman (Antwerp Zoo), Dr. Friederike von Houwald (Basel Zoo), Dr. Lydia Kolter (Köln Zoo), and Dr. Florian Sicks (Dortmund Zoo). We also thank the following zoos and zoological parks and gardens for contributing photos of their captive bears: Brookfield (IL), Cheyenne Mountain (CO), Cincinnati (OH), Cleveland Metroparks (OH), Connecticut's Beardsley (CT), Gladys Porter (TX), Houston (TX), Minnesota (MN), Oglebay's Good (WV), Racine (WI), Reid Park (AZ), Rolling Hills Wildlife Adventure (KS), Salisbury (MD), San Antonio (TX), San Diego Global (CA), Smithsonian National Zoo (DC), and Smoky Mountain (TN). RVH and CLC were supported by San Diego Zoo Global, while BZ was supported by the US Dept of Education. The online survey was made possible by Yuri Nataniel Daza (Centro de Biodiversidad y Genética).

References

Allen, W. L. et al. 2011. Why the leopard got its spots: relating pattern development to ecology in felids. – Proc. R. Soc. B 278: 1373–1380.

Bashir, T. et al. 2013. Estimating leopard cat *Prionailurus bengalensis* densities using photographic captures and recaptures. – Wildl. Biol. 19: 462–472.

Bindemann, M. et al. 2012. Who can recognize unfamiliar faces? Individual differences and observer consistency in person identification. – J. Exp. Psychol. Appl. 18: 277–291.

Burnham, K. P. and Anderson, D. R. 2002. Model selection and multimodel inference: a practical information-theoretic approach, 2nd edn. – Springer.

Byers, T. 1999. Perspectives of aboriginal people on wildlife research. – Wildl. Soc. Bull. 27: 671–675.
Charpentier, M. J. E. et al. 2008. Paternal effects on offspring fitness in a multimale primate society. – Proc. Natl Acad. Sci. USA 105: 1988–1992.
Chazdon, R. L. et al. 2009. Beyond reserves: a research agenda for conserving biodiversity in human-modified tropical landscapes. – Biotropica 41: 142–153.
Danielsen, F. et al. 2007. Increasing conservation management action by involving local people in natural resource monitoring. – Ambio 36: 566–570.
Dathe, H. 1967. Bemerkungen zur Aufzucht von Brillenbären, *Tremarctos ornatus* (Cuv.), im Tierpark Berlin. – Zool. Gart. 34: 105–133.
De Angelo, C. et al. 2010. Traditional versus multivariate methods for identifying jaguar, puma and large canid tracks. – J. Wildl. Manage. 75: 1141–1153.
Diefenbach, D. R. et al. 2003. Variability in grassland bird counts related to observer differences and species detection rates. – Auk 120: 1168–1179.
Dunning, D. et al. 2003. Why people fail to recognize their own incompetence. – Curr. Direct. Psychol. Sci. 12: 83–87.
Eck, S. 1969. Über das Verhalten eines im Dresdener Zoologischen Garten aufgezogenen Brillenbären (*Tremarctos ornatus* [Cuv.]). – Zool. Gart. 37: 81–92.
Evans, J. W. et al. 2009. Determining observer reliability in counts of river otter tracks. – J. Wildl. Manage. 73: 426–432.
Forcada, J. and Aguilar, A. 2000. Use of photographic identification in capture-recapture studies of Mediterranean monk seals. – Mar. Mamm. Sci. 16: 767–793.
Foster, R. J. and Harmsen, B. J. 2012. A critique of density estimation from camera-trap data. – J. Wildl. Manage. 76: 224–236.
Frasier, T. R. et al. 2009. Sources and rates of errors in methods of individual identification for North Atlantic right whales. – J. Mammal. 90: 1246–1255.
Garshelis, D. L. 2011. Andean bear density and abundance estimates – how reliable and useful are they? – Ursus 22: 47–64.
Goldstein, I. and Márquez, R. 2004. Monitoring Andean bear activity and movement along natural trails using non-invasive techniques in Venezuela. – Int. Bear News 13: 23.
Goldstein, I. et al. 2010. *Tremarctos ornatus* Andean bear. IUCN 2010. 2010 IUCN Red List of Endangered Species. Ver. 2013.2 <www.iucnredlist.org>. Accessed 24 March 2014.
Goswami, V. R. et al. 2012. Optimizing individual identification and survey effort for photographic capture–recapture sampling of species with temporally variable morphological traits. – Anim. Conserv. 15: 174–183.
Hastings, K. K. et al. 2008. Evaluation of a computer-assisted photograph-matching system to monitor naturally marked harbor seals at Tugidak Island, Alaska. – J. Mammal. 89: 1201–1211.
Hiby, L. et al. 2009. A tiger cannot change its stripes: using a three-dimensional model to match images of living tigers and tiger skins. – Biol. Lett. 5: 383–386.
Higashide, D. et al. 2012. Are chest marks unique to Asiatic black bear individuals? – J. Zool. 288: 199–206.
Hornaday, W. T. 1911. The spectacled bear. – Bull. N. Y. Zool. Soc. 45: 747–748.
Jarman, P. J. et al. 1989. Macropod studies at Wallaby Creek VIII. Individual recognition of kangaroos and wallabies. – Aust. Wildl. Res. 16: 179–185.
Jones, T. 2010. Detection probability and individual identification of the Andean bear (*Tremarctos ornatus*) using camera trapping methods. – MS thesis, Univ. of Wisconsin at Madison.
Karlsson, O. et al. 2005. Photo-identification, site fidelity, and movement of female gray seals (*Halichoerus grypus*) between haul-outs in the Baltic Sea. – Ambio 34: 628–634.
Kelly, M. J. 2001. Computer-aided photograph matching in studies using individual identification: an example from Serengeti cheetahs. – J. Mammal. 82: 440–449.
Kruger, J. and Dunning, D 1999. Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. – J. Pers. Soc. Psychol. 77: 1121–1134.
McGregor, P. and Peake, T. 1998. The role of individual identification in conservation biology. – In: Caro, T. M. (ed.), Behavioral ecology and conservation biology. Oxford Univ. Press, pp. 31–55.
Mendoza, E. et al. 2011. A novel method to improve individual animal identification based on camera-trapping data. – J. Wildl. Manage. 75: 973–979.
Naughton-Treves, L. et al. 2006. Expanding protected areas and incorporating human resource use: a study of 15 forest parks in Ecuador and Peru. – Sustainability Sci. Practice Policy 2: 32–44.
Ngoprasert, D. et al. 2012. Density estimation of Asian bears using photographic capture–recapture sampling based on chest marks. – Ursus 23: 117–133.
Noyce, K. V. et al. 2001. Differential vulnerability of black bears to trap and camera sampling and resulting biases in mark–recapture estimates. – Ursus 12: 211–226.
Oppenheim, A. N. 1992. Questionnaire design, interviewing and attitude measurement. – Continuum.
Paisley, S. et al. 2010. Facial markings of Andean bears – can you tell one from another? – Int. Bear News 19: 26.
Patton, F. and Jones, M. 2008. Errors that occur when using photo-identification to identify individual black rhinos. – Pachyderm 44: 35–44.
Pennycuick, C. J. 1978. Identification using natural markings. – In: Stonehouse, B. (ed.), Animal marking: recognition markings of animals in research. Macmillan Press, pp. 147–159.
Peyton, B. 1989. The ecology of conservation: a case for an ecosystem approach. – In: Rosenthal, M. A. (ed.), Proc. 1st Int. Symp. Spectacled Bear. Lincoln Park Zool. Soc., pp. 74–91.
Ríos-Uzeda, B. et al. 2007. A preliminary density estimate for Andean bear using camera-trapping methods. – Ursus 18: 124–128.
Roth, H. H. 1964. Ein Beitrag zur Kenntis von *Tremarctos ornatus* (Cuvier). – Zool. Gart. 29: 107–129.
Saporiti, E. J. 1949. Contribución al conocimiento de la biología del oso de lentes. – Anal. Soc. Científ. Argentina 147: 3–12.
Schofield, G. et al. 2008. Investigating the viability of photo-identification as an objective tool to study endangered sea turtle populations. – J. Exp. Mar. Biol. Ecol. 360: 103–108.
Sharma, S. et al. 2005. Identification of individual tigers (*Panthera tigris*) from their pugmarks. – J. Zool. 267: 9–18.
Stander, P. E. et al. 1997. Tracking and the interpretation of spoor: a scientifically sound method in ecology. – J. Zool. 242: 329–341.
Stevick, P. T. et al. 2001. Errors in identification using natural markings: rates, sources, and effects on capture–recapture estimates of abundance. – Can. J. Fish. Aquat. Sci. 58: 1861–1870.
Swanson, E. M. et al. 2013. Ontogeny of sexual size dimorphism in the spotted hyena (*Crocuta crocuta*). – J. Mammal. 94: 1298–1310.
Thomas, O. 1902. On the bear of Ecuador. – Ann. Mag. Nat. Hist. 9 (Series 7): 215–217.
Treves, A. et al. 2006. Co-managing human–wildlife conflicts: a review. – Hum. Dim. Wildl. 11: 383–396.
Velez-Liendo, X. 2005. Bolivia: reason to kill a bear, "I did it because it was there". – Int. Bear News 14: 23.
Viera, A. J. and Garrett, M. 2005. Understanding interobserver agreement: the kappa statistic. – Fam. Med. 37: 360–363.

Watkins, M. W. and Pacheco, M. 2000. Interobserver agreement in behavioral research: importance and calculation. – J. Behav. Ed. 10: 205–212.
Yoshizaki, J. et al. 2009. Modeling misidentification errors in capture–recapture studies using photographic identification of evolving marks. – Ecology 90: 3–9.
Zuercher, G. L. et al. 2003. Identification of carnivore feces by local peoples and molecular analyses. – Wildl. Soc. Bull. 31: 961–970.
Zug, B. 2009. Individual identification and habitat use of Andean bears on private lands in the Ecuadorian Andes. – Univ. of Wisconsin-Madison.