2012

An Investigation of Human-Error Rates in Wildlife Photographic Identification; Implications for the Use of Citizen Scientists

Megan Chesser

University of Massachusetts Amherst

Follow this and additional works at: https://scholarworks.umass.edu/theses

Part of the Bioinformatics Commons, Ecology and Evolutionary Biology Commons, and the Science and Technology Studies Commons

Retrieved from https://scholarworks.umass.edu/theses/768

This thesis is brought to you for free and open access by ScholarWorks@UMass Amherst. It has been accepted for inclusion in Masters Theses 1911 - February 2014 by an authorized administrator of ScholarWorks@UMass Amherst. For more information, please contact scholarworks@library.umass.edu.
AN INVESTIGATION OF HUMAN-ERROR RATES IN WILDLIFE PHOTOGRAPHIC IDENTIFICATION; IMPLICATIONS FOR THE USE OF CITIZEN SCIENTISTS

A Thesis Presented

by

MEGAN E. CHESSER

Submitted to the Graduate School of the University of Massachusetts Amherst in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

February 2012

Wildlife and Fisheries Conservation
ACKNOWLEDGMENTS

The few dozen pages that follow are incomplete reflections of the wealth of knowledge and enthusiasm that have been shown to me by so many people during my time here at UMass. It is only with their unfailing support in the form of encouraging words, heart-felt questions, smiles, hugs, offers of chocolate, and friendship that I have been able to finish my degree. I am deeply indebted to Jack Finn, whose infinite patience, positive energy, and selflessness taught me much about what it means to be a mentor. To Paige and Charlie, I wish I could give you extra hours in your days- so valuable are each and every one of your contributions not only to my own work and personal development, but also every student or colleague that walks through your doors; You are our department’s strong and steady backbone, offering students unprecedented research and learning opportunities, serving the community, and envisioning growth and potential for positive change in the world. To Kevin, I thank you for setting the bar high for each of your students, for modeling unwavering dedication to superb teaching and for demonstrating successful integration of real-world problems into course and lab-work. I thank you for, and encourage your continued commitment to, long-term research and for granting me the opportunity to be part of such a unique project that allowed me to interact with so many different people groups. Many thanks to John Kim for his patience, attention to detail, and willingness to collaborate with me in this new application for mandermatcher.

To my labmates L. Gamble, S. Haire, J. Seavey, B. Timm, E. Plunkett, T. Portante, and W. Sytsma, I am forever grateful for inspiration, solidarity, assistance, and many, many fond memories. I owe special thanks to J. Connolly, S. Spencer, M. Notestine, A. Soper, L. and J. Parfrey, A. and M. Swedlund, J. and G. Harvey (as well as everyone at Wesley UMC), Laurie Sanders and Fred Morrison, E. Bell, C. Penton, and A. Morgan for being such bright rays of hope and happiness in my life. Deeper friendships with you are one of the most valuable products of
my time spend working towards my degree. I thank my family for their unwavering love and support, even when they didn’t understand why I would want to spend many cold rainy nights outside hunting for salamanders.

Finally, I am thankful for the tireless efforts of our office staff – especially Carolyn, Linda, Roxann, and Lori – and for the remarkable Dan Pepin (jack of all trades) whose assistance saved me countless hours on numerous occasions. And last but certainly not least, I express my gratitude for the exceptional dedication and enthusiasm of our numerous volunteers and technicians, especially C. Jordan, N. Hahn, L. Moriarty, P. Reidy, P. Noonan, N. Damon, and C. Wise. Without you, none of this work would have been possible. I also offer many thanks to the National Science Foundation for providing the funding that allowed us to conduct this work.
ABSTRACT

AN INVESTIGATION OF HUMAN-ERROR RATES IN WILDLIFE PHOTOGRAPHIC IDENTIFICATION; IMPLICATIONS FOR THE USE OF CITIZEN SCIENTISTS

FEBRUARY 2012

MEGAN E. CHESSER, B.S., NORTH CAROLINA STATE UNIVERSITY

M.S., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Professor Kevin McGarigal

Rapid technological advancements in digital cameras and widespread public access to the internet have inspired many researchers to consider alternative methods for collecting, analyzing, and distributing scientific data. Two emerging fields of study that have capitalized on these developments are “citizen science” and photo-id in wildlife capture-mark-recapture (CMR) studies. Both approaches offer unprecedented flexibility and potential for acquiring previously inconceivable datasets, yet both remain dependent on data collection by human observers. The absence of rigorous assessment of observer error rates causes many scientists to resist citizen science altogether or to fail to incorporate citizen-collected data into ecological analyses. This same need for consistent measurement and documentation of the type and frequency of errors resulting from different observers is mirrored in numerous ecological studies employing photographic identification. The driving question of interest behind this thesis rests at the intersection of these two fields: can citizen scientists provide an effective alternative to commonly utilized computer-assisted programs used with large photo-id databases from wildlife studies?

To address this question we reviewed the history of wildlife photo-id in order to gain a better understanding of knowledge gaps caused by a failure to consistently report human error rates (Chapter 1). We then piloted a crowdsourcing approach to distributed photographic
analysis by soliciting responses to image comparisons from a large number of untrained observers (Chapter 2).

We found that observers correctly assessed 99.6% of all comparisons, but that the predictor variables for the two types of error (false positive and false negative) differed. Building upon a deeper understanding of the history, limitations, key issues, and recommendations for researchers considering using photo-id, we recommend the expanded use of citizen science methods as an effective alternative to computer-assisted approaches with large image libraries. Error rate improvements should allow scientists to more readily accept data collected by untrained observers as valid, and will also contribute to improved accuracy of ecological estimates of population size, vital rates, and overall conservation management of threatened or endangered species. Additionally, the general public will benefit from expanded opportunities to engage with and learn about the scientific process.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACKNOWLEDGMENTS</td>
<td>iv</td>
</tr>
<tr>
<td>ABSTRACT</td>
<td>vi</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>x</td>
</tr>
<tr>
<td>CHAPTER</td>
<td></td>
</tr>
<tr>
<td>1. AN OVERVIEW OF WILDLIFE PHOTOGRAPHIC IDENTIFICATION</td>
<td>1</td>
</tr>
<tr>
<td>1.1 Abstract</td>
<td>1</td>
</tr>
<tr>
<td>1.2 Introduction</td>
<td>2</td>
</tr>
<tr>
<td>1.3 Type of natural mark</td>
<td>5</td>
</tr>
<tr>
<td>1.3.1 Morphology</td>
<td>5</td>
</tr>
<tr>
<td>1.3.2 Pigmentation</td>
<td>5</td>
</tr>
<tr>
<td>1.4 Image comparison method</td>
<td>6</td>
</tr>
<tr>
<td>1.4.1 Manual</td>
<td>6</td>
</tr>
<tr>
<td>1.4.2 Alpha-numeric encoding</td>
<td>7</td>
</tr>
<tr>
<td>1.4.3 Computer-assisted</td>
<td>9</td>
</tr>
<tr>
<td>1.5 Key issues in photographic identification</td>
<td>13</td>
</tr>
<tr>
<td>1.5.1 Tag-ability</td>
<td>13</td>
</tr>
<tr>
<td>1.5.2 Distinctiveness</td>
<td>14</td>
</tr>
<tr>
<td>1.5.3 Stability</td>
<td>15</td>
</tr>
<tr>
<td>1.5.4 Photographic quality</td>
<td>16</td>
</tr>
<tr>
<td>1.5.5 Measurability</td>
<td>19</td>
</tr>
<tr>
<td>1.5.6 Error rate estimation</td>
<td>21</td>
</tr>
<tr>
<td>1.6 Conclusions and final recommendations</td>
<td>25</td>
</tr>
<tr>
<td>2. ANALYSIS OF HUMAN ERROR RATES RELATED TO PHOTOGRAPHIC IDENTIFICATION IN ECOLOGICAL DATABASES: IMPLICATIONS OF INCORPORATING CITIZEN SCIENTISTS</td>
<td>28</td>
</tr>
<tr>
<td>2.1 Abstract</td>
<td>28</td>
</tr>
<tr>
<td>2.2 Introduction</td>
<td>29</td>
</tr>
<tr>
<td>2.3 Materials and methods</td>
<td>33</td>
</tr>
<tr>
<td>2.3.1 Collection of image library</td>
<td>33</td>
</tr>
<tr>
<td>2.3.2 Study design</td>
<td>34</td>
</tr>
</tbody>
</table>
LIST OF FIGURES

Figure | Page
--- | ---
2.1. Screenshot from “real matching” on mandermatcher.com. With each comparison an observer determined if the salamanders were the same individual on the basis of their pattern information. The reference image on the left was constant throughout an entire trial. The comparison image on the right changed every time the user clicked the “Next” button at the bottom of the page. .......................... 62

2.2. Raw observer response data subdivided by the known, true relationship between the reference and comparison images. For 143841 (141815 + 2026) out of a possible 144373 comparisons, observers correctly identified the comparison image (99.63%); in 533 (127 + 406) comparisons, observers made an error (0.37%). In order to display all results on the same figure, please note that we have inserted a gap in the y-axis. .......................... 63

2.3. The total number of errors of each type observed per unique observer (all trials combined). False negatives are much more common and occur more frequently than false positives, which occur rarely except in higher frequencies for observers #34, and #13. .......................... 64

2.4. Hypothetical scenarios for observers from the general public committing false positive (FP) errors while participating in ecological photographic identification. We examined differences in the probability of making a FP error as a result of changing observer current experience (# of previous images viewed) for both novice and experienced participants. The thick blue line (very close to zero) represents the population average, and thinner multi-colored lines represent individual observers. .......................... 65

2.5. Hypothetical scenarios for observers from the general public committing false positive (FP) errors while participating in ecological photographic identification. We examined differences in the probability of making a FP error as a result of changing the number of matches in the catalog for a hypothetical novice observer with no current experience and having already viewed an average number of images. The probability of making a false positive after viewing the maximum number of images is essentially invariant from the average number of images. The thick blue line (near zero) represents the population average, and thinner multi-colored lines represent individual observers. Additionally (not shown) these same hypothetical scenarios with an experienced observer show similar trends over a much smaller range of probabilities. .......................... 66
2.6. Hypothetical best and worst case scenarios for observers from the general public committing false negative errors while participating in ecological photo-id. Under a worst-case scenario (on the left), variables not plotted on the x-axis were chosen to reflect poor match-finding conditions: largest catalog size (300 images), smallest number of matches (1), and longest time interval between matching images. Under a best-case scenario (on the right), variables not plotted on the x-axis were chosen to reflect optimum match-finding conditions: smallest catalog size (75), largest number of matches (8), and shortest time interval between matching images. In order to plot both these best and worst-case scenarios we used average age and average current observer experience values. We examined differences in the probability of making a false negative error as a result of changing A) time interval between matching images, B) number of matches in the catalog, C) size of the catalog viewed in a single sitting, and D) observer experience (# of previous images viewed). The thick black line in each plot indicates the population average; thinner multi-colored lines represent individual observers.
CHAPTER 1
AN OVERVIEW OF WILDLIFE PHOTOGRAPHIC IDENTIFICATION

1.1 Abstract

The use of photographic identification in capture-mark-recapture studies has expanded rapidly since the 1960’s as scientists look for less-invasive, permanent, and cheaper marking techniques for large numbers of individual animals. Finding and comparing previous methods, results, outstanding challenges, and suggested solutions across dozens of different scientific journals presents a difficult task for a researcher newly interested in such a widely applied method. To address the need for a comprehensive guide, we present a broad overview of the types of natural marks (e.g. morphological and pigmented) that have previously been used to successfully identify individuals and we outline the three typical methods used to compare images: manual, alpha-numeric encoding, and computer-assisted approaches. We highlight important issues related to image collection and analysis (e.g.; photographic quality, pattern stability, and error rate), and offer recommendations (e.g.; double-marking to address pattern stability) for future studies incorporating photographic identification as a method of CMR.

Synthesis and applications. Researchers interested in photo-id should consider, among other factors, the size of their population and image catalog, the recapture rate and duration of their study, and their budget before deciding if, and what type of visual recognition approach is appropriate for their study. Recognizing and addressing key issues will serve to improve experimental design and analysis, facilitate comparisons between studies, and efficiently advance the methodology and technology associated with photographic identification in the future.
1.2 Introduction

Identifying and tracking individual animals across space and time remains essential to the study of population dynamics, life history, movement ecology, and connectivity (Wursig & Jefferson 1990). To this end, capture-mark-recapture (CMR) studies are defined by repeated sampling of individuals. Upon its first capture, an individual is given a unique mark or tag and is then released back into the population so that it has some chance of being recaptured on subsequent sampling occasions (Crosbie & Manly 1985). Though generally field-efficient and adaptable to the specific study system, applied tags like numbered or colored leg bands or passive integrated transponders (PIT-tags) can be costly and prone to lack of permanence over time (e.g. Nietfield, Barrett, & Silvy 1994; Arntzen et al. 2004). Most importantly, all applied tags are invasive to varying degrees and can affect behavior or probability of recapture as well as introduce potential health risks to the individual (e.g. Jackson & Wilson 2002; McCarthy & Parris 2004). Amid growing concerns associated with applied tags, many researchers are increasingly interested in utilizing less invasive marking techniques.

Naturally unique morphology and pigmentation patterns inherent to different individuals in a population offer a plausible alternative to applied tags. Examples of “natural tags” include but are not limited to: the shape and/or pigmentation of fins and flukes, scar patterns, and fur, skin, or scale patterns (Yoshizaki 2007). In most cases, these natural tags are static and permanent through time, eliminating the risk of tag-loss associated with applied tags (Hammond 1986; Blackmer, Anderson, & Weinrich 2000). Natural markings also tend to be universal within a species, enabling researchers to tag and potentially recapture all individuals (Arzoumanian, Holmberg, & Norman 2005). Furthermore, documenting these visual differences between individuals often does not require their capture or harm, and thus is far less invasive than required for applied tags (Moon, Ivanyi, & Johnson 2004). In particular, minimizing
handling of animals reduces the physical stress associated with applied tags that might lead to behavior changes or health risks. This makes natural tags an especially advantageous choice for researchers studying threatened or endangered species (Forcada & Aguilar 2000; Van Tienhoven et al. 2007).

Photographic identification\(^1\) is, by far, the preferred method of natural tagging in wildlife studies. Photographic approaches can be divided into two broad categories, active and passive, depending on the level of interaction (and thus potential to impart stress) required with the animals. Aerial photography (common with whales, e.g. Sears et al. 1990; Hillman et al. 2008) and the use of remote or infrared camera traps (common with tigers, e.g. Karanth 1995; bobcats, e.g. Heilbrun et al. 2003; or fishing cats, e.g. Cutter 2009 ‘unpublished data’) are considered passive since they do not involve direct human interaction with the animal being photographed. Alternatively, moving closer to an animal (e.g., by boat) or capturing it in order to obtain a photograph (e.g. Wursig & Jefferson 1990; Gilkinson et al. 2007) is considered an active approach since it involves direct human interference with the animal. Whether active or passive, and particularly since the development of the digital camera, photography remains one of the fastest and cheapest means of "marking" large numbers of individuals in a population (Harting, Baker, & Becker 2004). Photographs are usually stored in a library to be visually examined later for matches and to develop individual capture history files (Speed, Meekan, & Bradshaw 2007). Though it is possible to manage small numbers of photographs manually, the task becomes

\(^1\) The exact term for this technique varies depending on a variety of factors that might include taxon, author preference, precedent, or publication source. For the purposes of this paper, we will use the term “photographic identification”, abbreviated photo id. It is important to recognize that literature searches should contain all possible combinations of terms to maximize returns and ensure a comprehensive review. Similar terms include but are not limited to: photo-identification, photographic individual identification, pattern recognition, pattern identification, pattern mapping, visual biometrics, biometric identification, image identification, natural marking identification, individual marking variation, non-metric identification, and individual numeric(al) encoding.
increasingly inefficient and error-prone when library sizes are large (Arntzen et al. 2004). As a result, many photographic identification studies now employ one or more of the following method modifications to limit the extent of manual matching: 1) coding of distinct features into a searchable database (e.g. Petersen 1972), 2) computer-assisted extraction of morphometric information (e.g. Araabi et al. 2000), or 3) semi-automated matching with pattern recognition algorithms (e.g. Beekmans et al. 2005; Gamble, Ravela, & McGarigal 2008).

Researchers entering this rapidly expanding field should consider several major factors before deciding if, and what type of visual recognition method is appropriate for their study context. A basic understanding of the history, limitations, and key issues associated with photographic identification will serve to improve experimental design and analysis, facilitate comparisons between studies, and efficiently advance the methodology and technology associated with this field. Herein we describe a framework for classifying the vast array of previous visual-based wildlife mark-recapture studies. Our main objectives are to: 1) provide a comprehensive review of current wildlife photographic identification literature and methodology, and 2) identify key issues and offer recommendations for future researchers selecting a photographic identification method. Due to the complexity of approaches employed in previous studies, we found it most useful to organize our review into four sections: 1) type of natural mark/pattern on the species of interest (e.g., morphology or pigmentation), 2) image comparison method (most basic to most advanced), 3) key issues in photo-id (e.g. accounting for differences in photographic quality), and 4) recommendations.
1.3 Type of natural mark

1.3.1 Morphology

Visual differences in the physical structure and form of individual animals provide one type of natural mark for populations. The degree of success and the nature of the morphological mark used to distinguish individuals vary widely based on species. For large terrestrial mammals like elephants or rhinoceros, ear outlines with notches and nicks (Ardovini, Cinque, & Sangineto 2008) or snout wrinkles, and horn shape (Goddard 1966; Patton & Jones 2008) have been used. For numerous cetacean species, dorsal fin and/or fluke outlines have long been the mark of choice (Hammond 1986; Markowitz, Harlin, & Wursig 2003; Mazzoil et al. 2004). For some reptile species with a lack of pigmentation pattern, the size and/or shape of scales or pineal spots has been found to be unique among individuals (Buonantony 2008; Reisser et al. 2008). If animals are exposed to wound infliction regularly, whether through intra/inter-species competition (as is the case with sea otter noses during the breeding season; Gilkinson et al. 2007), abiotic environmental interactions (e.g. scraping of the body along sharp ice surfaces; Sears et al. 1990), or anthropogenic injury (e.g. propellers on the dorsal surfaces of manatees; Langtimm et al. 2004), the size and shape of scar tissue can provide a type of natural mark suitable for individual identification.

1.3.2 Pigmentation

Patterns of pigmentation provide a second type of natural mark for populations. Like morphological marks, the degree of success and nature of the pigmentation vary widely based on species. We find it helpful to recall the alternate term of “pattern mapping” in order to subdivide kinds of pigmentation into groups based on points (spots), lines (stripes), and polygons (non-circular patches and mottling). Perhaps the most ubiquitous pigmentation
pattern- spots- have been used for photographic identification from salamanders (Hagstrom 1973; Milanovich et al. 2006), to big cats (lions- Pennycuick & Rudnai 1970; leopards- Miththapala et al. 1989; cheetahs- Kelly 2001), to marine animals (harbor seals- Hastings, Small, & Hiby 2001; penguins- Burghardt et al. 2004; whale sharks- Arzoumanian, Holmberg, & Norman 2005; and ragged tooth sharks- Van Tienhoven et al. 2007). Stripes are less commonly found in the photo-id literature, but have been successfully used with zebras (Petersen 1972) and tigers (Karanth 1995; Karanth & Nichols 1998). Non-circular patches of pigment and mottling have been used to identify individual amphibians (Doody 1995; Church et al. 2007; Gamble, Ravela, & McGarigal 2008), reptiles (Sheldon & Bradley 1989; Moon, Ivanyi, & Johnson 2004; Perera & Mellado 2004; Nowak 2005), mammals (Hiby & Lovell 1990; Mizroch & Harkness 2003), and even crustaceans (Frisch & Hobbs 2007).

However, not all studies fit neatly into either the morphology or pigmentation natural mark categories. Many studies use a combination of natural marks to reliably identify the same individual across time (Wursig & Jefferson 1990; Auger-Méthé & Whitehead 2007). Additionally, doubling tagging with some form of applied tag or genetic samples can provide a reliable method of assessing both “stability” issues with natural marks and sources of human or computer processing methodological errors.

1.4 Image comparison method

1.4.1 Manual

By far the most utilized method of photographic identification, “manual” implies a simple visual comparison of every image to every other image in the library, unassisted by computer pattern recognition algorithms (although computers may be used to view the images). The advantages to this method are that it can be implemented easily at little to no cost (except
to purchase the camera and film, if necessary), and that it requires no technical skills. One disadvantage to this method is that in some cases where the natural mark is complex and/or its interpretation is difficult, training of participants conducting the comparisons might be required to minimize error rates (Agler 1992; Friday et al. 2000). More important is recognizing that factors like the number of samples and duration of the study impact image library size; as library size increases, the number of pair-wise comparisons, and thus processing time, grows exponentially (Sacchi et al. 2007).

From a sampling context, manual comparisons are ideal for small (< 200, Huele et al. 2000) to medium (<850, Hammond, Mizroch, & Donovan 1990) sized databases. As with any marking technique, but particularly with photo-id where multiple photographs of the same individual can be linked, higher recapture rates can reduce the length of time spent searching for matches in the database and improve the “matching success” (Van Tienhoven et al. 2007; Hastings, Hiby, & Small 2008). Using this method with relatively simple mark types and not, for instance, complicated mottling or very small spots (which make it more difficult to distinguish differences rapidly by eye) can also reduce both the amount of time spent per comparison as well as the human error rate (Kelly 2001). Regardless of these precautions, at larger database sizes and with complex patterns, manual photographic identification is likely to be the least efficient in terms of per-image processing time (Araabi et al. 2000).

1.4.2 Alpha-numeric encoding²

Going a step further than the manual comparison method, some researchers assign a number, letter, or combination alpha-numeric code or category to each image/individual during

---

² Often times the marine mammal literature refers to encoding as “landmarking”, or the designation by an observer of a particularly noticeable “landmark” on the individual that can be used for identification purposes.
its initial processing. These codes usually represent key features unique to the individual (e.g. fluke “color type” for humpbacks, Friday et al. 2000; arrangement of spots on leopards, Miththapala et al. 1989; arrangement of black ventral scales on wall lizards, Sacchi et al. 2007; or the dorsal ratio in dolphins, Defran, Shultz, & Weller 1990). Whenever a new image is compared to the database, its code(s) serve(s) as a search term to easily narrow down the list of potential matches. Beyond the initial encoding and filtering of the catalogue, this method remains the same as the basic manual method, with the observer comparing the new (reference) image to every other potential match with the same code(s) or category. Like the manual method, alpha-numeric encoding is relatively easy and straightforward in that it requires little or no technical skills. On the other hand, the additional encoding step can lengthen initial image handling time, and create a source of potential human bias or error. For example, differences in personal interpretation of identifying features between observers or even within observers across sampling occasions can lead to improper encoding of individuals (Huele et al. 2000). This difficulty assigning codes affects the effectiveness of this method.

Alpha-numeric encoding can be applied to images in all catalogue sizes but it is most often used with relatively medium to large numbers of images (used with approximately 2000 photographs (Huele et al. 2000), and approximately 3000 photographs (Nace, Richards, & Hazen 1973; Friday et al. 2008)). Easily quantified or classified patterns such as the number (and shape) of spots in each quadrant of a leopard frog (Nace, Richards, & Hazen 1973), or the shape of zebra stripe intersections (Petersen 1972) are ideal mark types for this method. Other simple or even moderately complex patterns might be adapted to this method if the encoding process is modified to select only certain components of the pattern (e.g., the pigmentation level of a whale fluke as light, medium, or dark; Friday et al. 2000). Over time and with larger image libraries, alpha-numeric encoding is ideally faster per comparison than the manual method
because it allows for an immediate reduction in the percentage of the total library that must be viewed to check for matches (Kreho et al. 1999).

1.4.3 Computer-assisted

All studies employing “computer-assisted” methods fall along a continuum of image enhancement and pre-processing steps. Evolving from tedious sorting, measuring, and calculating by hand, computer-assisted photo-id today includes a range of applications from efficient, computerized extraction of morphometric information to fully-automatic sorting of matches by algorithms. Hammond, Mizroch, and Donovan offer several great examples of the earliest forms of computer assistance from the marine mammal literature (1990). Despite impressive diversification and technological advances since the initial efforts, almost all computer-assisted photo-id of wildlife is still unable to identify an individual without human assistance. The one published exception to this statement is the limited scale, fully-automated penguin matching achieved by Sherley et al. (2010). Even here, researchers caution that manual verification of millions of matches would be necessary to maintain high system performance at larger database sizes (Sherley et al. 2010). Thus, ultimately, the final matching decisions, as with the manual method, remain up to the researcher (Whitehead 1990; Kelly 2001). Though the detailed nature of the process may vary across taxa and among studies, the computer-assisted photo-id method has three main phases.

First, distinctive visual information is input into the computer. If whole images are used, this might simply mean downloading digital photographs (or video frames), digitizing a printed image, or scanning a negative. For photographs where only part of the animal/pattern is used for identification purposes, particular components, such as an arrangement of spots (treated as points; Van Tienhoven et al. 2007), trailing fluke edges (treated as lines, Hillman et al. 2003), or
pigmentation patches (treated as polygons, Mizroch, Beard, & Lynde 1990) might be individually digitized or encoded. Human-mediated input is necessary because, as of yet, completely automated feature detection remains unfeasible\(^3\) or unreliable due to issues related to the poor image quality common in many wildlife studies (Ardovini, Cinque, & Sangineto 2008). In some cases, this digital information is then used to extract additional morphological values such as the dorsal ratio for dolphins (Wilkin, Debure, & Roberts, 1998). This digital information can also be used to derive mathematical descriptors or metrics (e.g., proportional distance of each marked point along a ridgeline for narwhales, Auger-Méthé 2008; chest width and shape of penguins, Burghardt et al. 2004; or reference triangles between triplets of coordinating points for whale sharks, Arzoumanian, Holmberg, & Norman 2005). Comparing images based on these metrics has the advantage of being less sensitive to potential differences in human interpretation of images (Huele et al. 2000), but can create problems if sources of input/process error (e.g., mistaking a glare for a spot in the pattern) related to photographic quality are not taken into account (Arzoumanian, Holmberg, & Norman 2005). Still other applications store portions of the digital pattern information for individuals as a matrix of numbers representing pixel values derived from visual properties (Caiafa et al. 2005; Gamble, Ravela, & McGarigal 2008).

Second, images on the computer are compared and ranked from most to least similar using some form of a scoring system. Typically, a single image at a time serves as a reference while a pattern comparison algorithm compares it to all other images in the catalog, simultaneously ranking from the most to least similar. In many cases a similarity coefficient is calculated for each comparison by a matching algorithm, though it is important to recognize than not all matching algorithms utilize the same coefficient so the computed values are

\(^3\) Sherley et al. 2010 have achieved fully automated feature detection for a subset of 1000 “detections” of 114 individual penguins, however they acknowledge the difficulty in scaling their method up due to image quality and pattern detection capabilities.
generally not comparable across studies. Most long-term research studies have developed species-specific software, which blends the input and calculation phases. For example, Caiafa et al. (2005) developed a highly technical use of the ‘eigenfaces’ method (Turk & Pentland 1991) with elephant seals, which deconstructs images to a baseline of characteristic features and then compares these to identify new individuals. Whitehead (1990) developed and later expanded (Beekmans et al. 2005) upon a ‘highlight’ algorithm, which compares a series of identifying points for distinctive features such as nicks on sperm whale flukes. The Mid-Atlantic Bottlenose Dolphin Catalogue (Urian 2011) is one of many examples of data collections that process images utilizing a variation of a curve matching algorithm within a program called Finscan©, which was originally developed by Hillman et al. (2003). Additional curve-based algorithms are used in the Europhlukes database for numerous European cetaceans (Huele et al. 2000; Evans 2003) and in calculating dissimilarity values between different elephant ear shapes (Ardovini, Cinque, & Sangineto 2008). A slightly different, string-based matching algorithm has been used with dolphins by Araabi et al. (2000). Rather than points or lines, some algorithms compare elements of pixel-based vectors (e.g., marbled salamanders, Gamble, Ravela, & McGarigal 2008, penguins, Burghardt et al. 2004) or matrices\(^4\) directly (e.g., with gray seals, Hiby & Lovell 1990; Karlsson et al. 2005; harbor seals, Hastings, Small, & Hiby 2001; and cheetahs, Kelly 2001). Because there are differences in thresholds related to the definition of an “acceptable error” in the positioning of the initial pixel array, researchers should exercise caution in comparing resulting similarity values and/or or ranked weights from matching algorithms across studies.

Third, for the given reference image, researchers manually check the list of suggested matches to a specific depth or similarity score value and make final decisions about the recapture status of individuals. Ideally, the computer has facilitated their task significantly by

\(^4\) Also described as “identifying arrays (IA)” or “measurement regions”
reducing the number of images from the library they need to examine for potential matches. In comparing among published computer-assisted methods, it is important to be aware of differences in reporting of results. For datasets with known matches (i.e., experimentally constructed data sets for the purpose of assessing accuracy of the method), efficiency might be reported as: 1) length of time to find a match (Araabi et al. 2000; Auger-Méthé 2008), 2) total length of time to find all matches (Auger-Méthé 2008), 3) whether or not the correct match was found within a fixed depth of the library (e.g., 70 or 130 images evaluated) (Mizroch & Harkness 2003; Ardovini, Cinque, & Sangineto 2008), 4) whether or not a match was found (or error rates reported) above a certain critical value for the similarity threshold (Hastings, Small, & Hiby 2001), or 5) whether or not the match was found within a fixed proportion of the library based on rank similarity (e.g., top 0.3%, 0.5%, and 1% of an ordered comparison lists), (Hastings, Hiby, & Small 2008). Accuracy, if reported, might be quantified as: 1) percentage of matches found (Sherley et al. 2010), 2) number of overall commission (false positive) and omission (false negative) errors (Hastings, Hiby, & Small 2008) or within a designated depth or percentage of the database (Beekmans et al. 2005), or 3) the depth/percentage of the database that had to be searched before the match was found (Kreho et al. 1999; Araabi et al. 2000; Mizroch & Harkness 2003).

Computer-assisted photo-id is increasingly advantageous as image library size increases in that it offers the greatest reduction in time spent to find matches (Auger-Méthé 2008). Computing power/speed, sophisticated algorithms, and ranking systems combine to shorten the time it takes to extract valuable information and match large numbers of images. The ability of computer-assisted approaches to semi-automatically extract morphological or pigmentation information enables researchers to photographically mark and recapture animals with more complex patterns and handle increasingly large image catalogs (Hillman et al. 2003; Andersen et
al. 2010; Sherley et al. 2010). It has also been suggested that computer-assisted approaches offer a more standardized feature extraction, encoding, or matching process compared to manual methods that are susceptible to human bias. (Kniest, Burns, & Harrison 2010). Ultimately though, it is important to realize that this method still relies on the manual method either to build test sets of images or to make the final comparison of images (Whitehead 1990; Kelly 2001). Implied in the definition, this method also requires extensive technical (computer-science/programming) skills or collaboration. This may mean contracting work (and thus high startup costs for wildlife researchers on tight budgets), but it also provides unique opportunities for rich interdisciplinary partnerships. Digitalized data, databases, and algorithmic code can be easily transported and shared, facilitating collaboration and comparisons between studies (Mizroch, Beard, & Lynde 1990). Lengthy computer/algorithm design and preparation periods, as well as money and effort spent training observers to use a graphical input interface, can pay off with long-term, large databases of images.

1.5 Key issues in photographic identification

1.5.1 Tag-ability

Though CMR studies presume all animals have a constant and equal probability of capture (and thus marking) for each trapping occasion, in reality this tag-ability assumption is not always met. Because photo-id is often passive and does not require handling the animal to acquire an image, there can be distinct differences between “sighting” an animal, and actually being able to “mark” or photographically capture it (Hammond 1986). Researchers should consider, as well as report in subsequent publications, how they handle deviations from this assumption. To have truly equal probabilities of capture, all individuals must have the same

---

5 Also described as markability or universality (Burghardt 2008)
probability of being sighted AND photographically marked (Hammond 1986). Regardless of mark type or method of matching, differences in age, sex, body condition, and stochastic behavior or location between individuals and sampling occasions can lead to unequal photographic representation in the library. For example, sexual dimorphism in certain species can mean that only paler-colored, female grey seals (Hiby & Lovell 1990) or male lions with manes (Kays & Patterson 2002) are able to be photographically marked. Behavior related to breeding condition (e.g., swimming shallow with a calf), territoriality (e.g., a particular family group avoiding a sampling location because it overlaps territories with a rival), or movement style (e.g., angle of diving such that a full/partial dorsal fin or fluke is displayed) can also dramatically affect tag-ability in some species (Hammond 1986).

1.5.2 Distinctiveness

In photo-id studies the probability of recognizing, and thus recapturing, a marked individual is dependent on three major factors: 1) individual distinctiveness, 2) stability of the pattern, and 3) quality of the photograph (Hammond, Mizroch, & Donovan 1990; Friday et al.2000). The first of these factors, distinctiveness, is based on another primary assumption of CMR studies -- that each individual in a population has a distinct mark by which they can be identified. “Distinctiveness” has also been defined as “recognisability” (Hammond, Mizroch, & Donovan 1990), “unique information content” that each individual contributes to distinguishing itself from others (Burghardt 2008), and “the degree of visibility of permanent marks” related to the “ease of individual identification” (Forcada & Aguilar 2000). Across species, distinctiveness is commonly age or sex dependent. Most amphibians, pinnipeds, and birds, for instance, fail to develop a stable pattern until they reach sexual maturity because juveniles often exhibit different pigmentation patterns than adults (e.g. Forcada & Aguilar 2000). Similarly, with
morphological marks, it is typically older manatees, and female otters that have the highest degree of scarring available to photograph (Langtimm et al. 2004; Finerty, Hillman, & Davis 2007). Within a species there is natural variation in morphology and/or pigmentation among individuals (Friday et al. 2000). The more divergent the pattern or shape between two photos, the easier it is for an observer or a computer to identify, code, or rank the individuals as different (Agler 1992). Lastly, if not accounted for, gradations in distinctiveness can potentially cause both human and computer processing errors or bias in matching (Friday et al. 2000; Harting, Baker, & Becker 2004).

1.5.3 Stability

Another factor affecting probability of recapture is stability (or permanence) of the pattern, defined as the propensity for change in a mark over time. Stability is a major assumption in CMR studies. Individuals with completely static marks will have constant and equal recapture probabilities across sampling occasions. Individuals with dynamic or “evolving natural marks” may change to such an extent that a pattern photographed on a previous occasion is unidentifiable as a recapture (Yoshizaki et al. 2009). Changing natural marks, whether pigmented or morphological, present a mechanism of misidentification analogous to the loss of an applied tag (Stevick et al. 2001). Rather than being matched to its original tag (photograph), an individual is tagged (photographed) again creating a separate capture history for a “new” individual (Yoshizaki et al. 2009). Researchers concerned about the issue of pattern change should consider the life-span of their study species with respect to the duration of the study (Hammond 1986), their estimated rate of recapture (e.g. Sherley et al. 2010), environmental or social factors that may accelerate pattern changes (Auger-Méthé &
Acquired morphological marks such as scars, nicks, or scrapes are, by their nature, more dynamic than a pigmented mark (Dufault & Whitehead 1995). While injuries that create visual patterns might be useful over short sampling intervals, such as within a season, these types of marks tend to heal over longer time periods (as is the case with humpback whales, Blackmer, Anderson, & Weinrich 2000) making them less reliable. Nevertheless, species like the sea lion (McConkey 1999) or manatee (Langtimm et al. 2004) which have their flippers and dorsal surfaces exposed to damage on a regular basis tend to accumulate scar tissue over time, necessitating a cataloguing system that allows for sequential mark development. Pigmentation can also be directional in its change over time, tending to lighten/darken, expand/contract, or appear/disappear (e.g., humpback whales, Carlson, Mayo, & Whitehead 1990; sperm whales, Dufault & Whitehead 1995; bottlenose whales, Gowans & Whitehead 2001; eastern tiger salamanders, M. Chesser, unpubl. data). Because stability relates to the distinctiveness of each individual through time, all matching methods are negatively impacted by the issue of pattern change as it adds a potential source of error. More complex mark types would be expected to be more difficult to assess by eye for slight (or major) changes, and thus some type of computer-assistance might be recommended in these situations (Kelly 2001; Anderson et al. 2010).

1.5.4 Photographic quality

Another source of processing and human error in recapturing individuals relates to the quality of the photographic image. In general, quality has been broadly defined with respect to several secondary factors: 1) clarity, sharpness, or focus (Hammond 1986; Friday et al. 2000), 2) contrast or the degree of difference between blacks and whites (Hammond 1986; Friday et al.
2000), 3) noise or the amount of unnecessary background information in the frame with the individual of interest (Burghardt et al. 2004; Gamble, Ravela, & McGarigal 2008), 4) resolution or the amount of available pixel information (Hammond 1986; Markowitz, Harlin, & Wursig 2003), 5) glare or specularity (Arzoumanian, Holmberg, & Norman 2005; Gamble, Ravela, & McGarigal 2008), and 6) relative size of the animal of interest in the picture frame (Hammond 1986; Sears et al. 1990). Weather, water depth or turbidity, patterns in the background environment, and ambient light are just a few natural environmental variables that affect photographic quality (Markowitz, Harlin, & Wursig 2003; Langtimm et al. 2004). Photographing individuals in their natural environments typically precludes the controlled lighting, uniform background, and limited movement that would otherwise improve photographic quality. In fact, it is this unpredictability of photographic quality and its complex relationship with distinctiveness that continues to prevent widespread use of fully automated matching systems (Ardovini, Cinque, & Sangineto 2008; Sherley et al. 2010).

Friday et al. (2000) succinctly describe this confounded relationship between quality and distinctiveness:

“As the quality of the photograph decreases, the information in the natural markings becomes obscured, and it becomes increasingly difficult to recognize the represented individual. Less distinctive individuals are more difficult to recognize than more distinctive individuals. The use of poor quality photographs further exacerbates this problem because very distinctive individuals can be more readily recognized from poorer-quality photographs.”

If CMR studies do not take into account the potentially confounding effects of quality and distinctiveness, probability of recapture and error type/rate can be affected (Hammond 1986;
Therefore, it has been recommended that photographic quality and distinctiveness be judged separately (Hammond, Mizroch, & Donovan 1990).

Photographic quality is often judged when images are initially input into the database. Depending on the study, images may be given a categorical (e.g., excellent to poor) or numerical (e.g., 1-6) ranking of quality (Gowans & Whitehead 2001; Mizroch & Harkness 2003). Though all photo-id methods ideally use high quality images, in some cases software programs enable researchers to access pattern information from even low quality images through brightness and contrast manipulation (Mazzoil et al. 2004; Sherley et al. 2010). When possible, some studies have attempted to standardize lighting, extraneous background noise, and camera angle for each image to reduce differences among images in photographic quality (e.g. Gamble, Ravela, & McGarigal 2008). Whether through the photographic process or image categorization at input, we encourage researchers to report the quality of images included in their analyses. Knowing which categories were used in the matching/analysis (e.g., only “good and excellent”, Gowans & Whitehead 2001; Auger-Méthé 2008; equal numbers of each quality category, Friday et al. 2000; 2008; or all possible images, Mizroch & Harkness 2003) cues readers to look for possible effects of image quality on the significance and reliability of reported results. For example, were the differences in photographic quality addressed as a source of error or bias (within individual observers, between individual animals, between sampling occasions, or as a factor affecting the size of the database searched for matches)? Additionally, explicitly stating the quality of images used in an analysis clarifies the procedure, allowing for easier replication of and comparison between studies in the future.
1.5.5 Measurability

In many circumstances animals have patterns on a three-dimensional surface where valuable pattern information extends beyond a single plane of vision. “Measurability” refers to the ability of researchers to capture and extract this mark from the photograph (Gunnlaugsson & Sigurjonsson 1990; Burghardt 2008). One aspect of measurability is camera angle or orientation. Any time the camera is non-perpendicular to the animal’s surface, the observer can experience a non-linear deformation of pattern information and a decreased ability to accurately match images (Speed 2006; Burghardt 2008). Researchers in wildlife photo-id have two main options to deal with this issue: 1) develop a computerized three-dimensional model of the surface of the study species, or 2) take multiple images from different angles/sides per individual per capture event.

The concept of a 3-D model was first introduced to the field of photo-id in 1990 by Hiby and Lovell as a method of describing a pattern on a particular section of an animal such that the information would be invariant to the effect of ‘camera orientation and the posture of the animal at the moment the photograph was taken’ (Hiby & Lovell 1990). First accomplished with gray seals, but subsequently applied to a wide variety of species (cheetahs, Kelly 2001; harbor seals, Hastings, Small, & Hiby 2001; tigers, wildebeest, crested newts, sand lizards, chital, leopards and more, Hiby 2011), this complex and costly method constructs a 3-D mathematical model of the surface of the animal which is then projected onto the image. The model dictates the degree of computational transformation necessary to account for distortion related to camera viewpoint prior to extracting the identifying marks for each individual (Hiby & Lovell 1990). The 3-D approach is a special computer-assisted method of pre-processing images before storing unique numerical (vector/matrix) representations of patterns for each individual.
Taking multiple photographs of each individual in the field offers a cheaper and less technically demanding method for capturing pattern information of multiple planes of a surface. Though ideal, it is not necessarily realistic to expect that each individual will position itself such that their entire pattern viewed from multiple planes can be photographed on each sampling occasion (Wursig & Jefferson 1990; Van Tienhoven et al. 2007). Complete sets of images may not be obtained until subsequent sampling occasions, or may remain incomplete, possibly affecting recapture error rate (Karanth 1995; Haddad and M. Chesser, unpubl. data).

As with photographic quality, we encourage researchers to report as much methodological information as possible related to measurability. Information might include, but is not limited to: the minimum number (and quality) of photos required to positively identify an individual (of a given distinctiveness), the number of photos actually taken per individual per occasion, and the expected recapture rate (which relates to the number of expected images/individual across time) (Stevick et al. 2001). In addition, knowing the total number of images included in the database used for analysis (which is often different than the total size of the database of the study) and their relationships to one another is critical in determining and comparing efficiency across studies and/or between trials of matching algorithms. Whether photographing an individual once or multiple times per sampling occasion, many studies select a single “representative” photograph from a capture history to serve as a reference when comparing that individual to the rest of the image library in the future. When this happens, the entirety of each capture history is never directly compared to the entirety of other capture histories in the database, rather, only the representative photographs of each individual are compared. While time-saving because of the dramatic reduction in the total number of comparisons necessary, this technique can be problematic in terms of how it relates to overall accuracy and efficiency. Capture histories may contain mistakes, patterns/shapes may change.
over time, or multiple images may be required for a complete “measure” of distinguishing characteristics. Not viewing all combinations of images for an individual and the rest of the catalog reduces the level of experiential learning and information content potentially available from each additional comparison image (Patton & Jones 2008), as well as excludes the possibility of catching any mistakes due to slight pattern changes which are more readily visible when viewing capture histories in consecutive years (Carlson, Mayo, & Whitehead 1990).

With computer-assisted matching, Kelly (2001), Van Tienhoven et al. (2007), and Hiby et al. (2009) found that while slightly more time consuming, including more reference images per individual in their queried databases increased the accuracy and efficiency of programs, although Forcada and Aguilar (2000) observed no increase in accuracy with increasing images per individual for monk seals. With manual matching approaches, including more images in the catalog might possibly lower overall precision by increasing the likelihood that a viewer sees other individuals with similar patterns or shapes to the reference image in question, thus making it more challenging and time consuming to search for a correct match (Ardovini, Cinque, & Sangineto 2008). Similarly, if there are differences in photographic quality within the multiple images/individual, the inclusion of low quality images can complicate the ability of either humans or computer programs to isolate the mark in the photograph. In many cases this results in lower similarity coefficients between matching images, and thus negatively affects accuracy (Whitehead 1990; Kelly 2001). Including these types of information in publications will facilitate methodological understanding, replication, and comparisons between studies in the future.

1.5.6 Error rate estimation

Among researchers, error rate is perhaps the single most utilized criterion to compare (and chose) among photo-id methods, yet estimating error rates is one of the most confusing
issues because of the numerous ways to define and measure error. All possible combinations of natural mark types and methods of photo-id are subject to the same two sources of error: 1) missing a match (also known as a false negative, false rejection, or an omission error) and 2) making an incorrect match (also known as a false positive, false acceptance, or a commission error) (Hammond, Mizroch, & Donovan 1990). It is important to recognize that these two types of errors are separate and can have divergent impacts. False negatives (failing to match two photographs of the same individual), for instance, effectively lead to reduced survival and recapture estimates (Morrison et al. 2011) and inflated estimates of population size because fewer marked individuals are “recaptured” than should be the case (Hammond 1986; Gunnlaugsson & Sigurjónsson 1990). False positives (incorrectly assigning a match between two different individuals) effectively overestimate survival and recapture rates, and underestimate population size because more marked individuals are “captured” than should be the case (Hammond 1986).

In an attempt to reduce the occurrence of both types of errors, numerous studies have suggested protocol changes including, but not limited to: 1) confirming all matches and any potential new individuals by multiple (preferably experienced) observers and/or programs (Sears et al. 1990; Stevick et al. 2001; Beekmans et al. 2005), 2) comparing all photographs to the entire catalogue several times (Forcada & Aguilar 2000; Friday et al. 2008), 3) reviewing catalogues periodically to check for duplicates (Hammond, Mizroch, & Donovan 1990), 4) avoiding long matching sessions (>2-3 hrs) to prevent observer fatigue (Hammond, Mizroch, & Donovan 1990; Sears et al. 1990), 5) exercising caution (or eliminating altogether) using poor quality photographs in matching libraries (Stevick et al. 2001; Beekmans et al. 2005; Friday et al. 2008), 6) double marking animals with a second type of tag (Beekmans et al. 2005; Friday et al. 2008), 7) matching using multiple identifying features (Kniest, Burns, & Harrison 2010), or 8)
excluding all capture histories with only a single photograph (Morrison et al. 2011). While many studies have demonstrated that excluding poor quality photographs from the catalogue greatly reduces the rate of false negatives (Friday et al. 2008), addressing the impact of false positives has remained relatively uncommon in the literature (exceptions include Gunnlaugsson & Sigurjónsson 1990; Stevick et al. 2001).

Reporting of error rates associated with photo-id has been inconsistent and ambiguous at best. Studies have previously reported the number of errors (or conversely the number or known matches) found for a fixed depth of the image database (e.g., within the top 20 images based on image similarity, Gamble, Ravela, & McGarigal 2008; Morrison et al. 2011- calculated as a rate by dividing by the total number of true matches found), the number of errors for a fixed percentage of the database (e.g., the top 10% based on image similarity, Kelly 2001; Gamble, Ravela, & McGarigal 2008), the number of incorrect suggestions that rank higher than the actual match (variable depth and percentage) (Hillman et al. 2003; Mizroch & Harkness 2003), or, lastly, some studies fail to clearly report specifics regarding the number of images used to calculate error rate making interpretation of the reported error rates ambiguous and potentially even misleading. For example, Mizroch and Harkness (2003) seem to have conflicting numbers of images in their catalog: they state that a random 0.5% draw from their database results in 116 images (indicating a catalog size of approximately 23200 images), but then subsequently state that the computer-assisted matching program searched for a match until about 5% of the database had been searched (citing 1250 images- which indicates an overall catalog size of 25,000 images). Additionally, readers should exercise caution in interpreting results that state “Overall, matches were found for 74 of the 116 [sampled] photographs, and on average the first match was found in the top 0.0054 of the database (about 130 photographs) (SD= 0.0073)”; the caveat to notice here is that their protocol describes truncating
the catalog search at 1250 images. If a match existed for the sampled images at a depth deeper than the 1250 images viewed, they would not have found it, and thus the numbers they report should be presented with this in mind. Additionally, many computer-assisted methods do not specify enough details relating to the creation of the ranked list. Readers should be able to tell if this list was generated based on an absolute threshold similarity score (such that the length of the list was free to vary in length depending on how many other similar images were found in the database above a specified similarity value) (Kelly 2001; Hastings, Hiby, & Small 2008), or if the ranked list was restricted to a specific size/depth regardless of the number of images with similar scores reported beyond these limits (Auger-Méthé 2001; Gamble, Ravela, & McGarigal 2008).

Further complicating the interpretation of results is the fact that there is no consistency or standard for reporting the number of matching images (the number of images per individual) included in the searchable database. The type of relationship established between images (e.g., 'linked/identity propagation', such that if image A=B, and B=C, then A=C without ever directly comparing A and C, or 'independent' such that A is always compared to both B and C) and the numbers of images included in the database for each individual affect database size and composition, which in turn may affect error rate type and frequency (Gamble, Ravela, & McGarigal 2008). In addition, more often than not, error rates are calculated for a small subset of the entire database, and then extrapolated out to estimate the overall error rate for the entire database. However, the entire database is rarely assessed, thereby resulting in uncertainty associated with the scalability of the estimates. These varied sources of discrepancies in the literature create ample opportunities for the improper use of some statistical procedures and lead to great confusion for readers.
Specifically reporting both types and potential source(s) of error (e.g., pattern distinctiveness, photographic quality, or differences in observers) is essential when comparing methods and reliability of results between studies (Gunnlaugsson & Sigurjónsson 1990; Stevick et al. 2001; Gamble, Ravela, & McGarigal 2008). Moreover, reporting detailed methods for estimating error rates is of paramount importance for interpreting results. For the purposes of comparison across studies we recommend reporting error rates for a fixed depth (in absolute numbers) of the database (e.g., the top 100 ranked images) or, ideally, for a range of depths. We recommend considering graphically displaying the results such that the reader can visualize the relationship between both false positive and negative error rates as a function of catalogue size and/or depth (Fig. 3 errors and observer experience- Carlson, Mayo, & Whitehead 1990; also Fig. 3 proportion of library searched- Hastings, Hiby, and Small 2008). When possible, displaying the relationship between photo quality, number of photos per individual and/or pattern distinctiveness, and error rate and/or abundance estimates (e.g. see Fig. 1 and 2, Stevick et al. 2001) also offers valuable information to the future reader considering the outcomes and implications of a study.

1.6 Conclusions and final recommendations

Researchers entering the field of photographic identification can choose from three main methods of visual recognition: manual, alpha-numeric encoding, and computer-assisted. Careful consideration of the advantages and disadvantages of each (briefly summarized below) will ensure selection of the most appropriate method for the goals and limitations of the study system.

The manual method of handling images has proven an uncomplicated, yet effective approach for photographically recapturing an extensive array of animals in diverse
environmental conditions. Though some exceptionally large databases have utilized the manual approach, these were generally before the onset of the computer age. Most researchers today use the manual approach for simple to moderately complex patterns, and for smaller datasets (<200-800 images) with one (or a few) image(s) per individual. The fewer images an observer needs to view for an individual, the faster the image processing will be for identifying either a new or a recaptured identity within the library. Optimization of photographic quality is particularly important for manual methods where overall error rate and per comparison time tend to increase with the inclusion of poorer quality photographs. Similarly, accounting for differences in observer(s) experience level as well as distinctiveness of individuals is important as these also impact errors and can bias the results.

Alpha numeric encoding provides an expedited method to process medium to large datasets. Encoding requires minimal technical skills, but does depend on consistent classification and organization for the method to operate effectively. This method is generally restricted to images with simple, readily visible patterns that can be easily categorized or labeled. In circumstances with highly distinctive patterns, lower quality photos can often still be matched using the alpha-numeric code or category. However, it is important to recognize that the encoding and matching process are strongly dependent on the assumption that the mark is stable over time. Even small changes in the pattern from the first occasion an individual was captured could result in an individual being coded differently at the second capture occasion. Since the code/category serves as the first filter for the image library, improper coding (whether caused by pattern change or by differences in interpretation between observers) can result in missed matches (false negatives) and creation of duplicate capture histories when images are processed. Researchers electing to use this method should be aware of the repercussions associated with each key issue with regard to error rate and bias.
Though it is the most technically challenging of the photo-id methods, computer-assisted approaches can easily handle large image libraries. The time and cost associated with developing this method are usually offset with long-term datasets, in cases where multiple images are required per individual per sampling occasion, or in circumstances where the recapture rate is anticipated to be low. Algorithms and programs can be designed such that once established, individual identities are propagated throughout the database, reducing the list of potential matches when a new image is presented to the database. Researchers using computers to match images are encouraged to report detailed information on how the matching algorithm was designed (e.g., describe the characteristics of the subset of images used to design and test the procedure), per-comparison time for each image pair, and the efficiency in terms of the number and type of errors observed in a fixed depth (or range of depths) of the library.

It is our hope that this basic classification framework provides a useful contextual background for researchers considering photographic identification. Keeping in mind the potential mark types, methods, and key issues during all phases of the study should reduce the confusion that has historically been a problem in a very diverse field. Future efforts involving double tagging are needed to address important issues such as stability or error rates. Attaining high-quality images should remain a top priority for all researchers regardless of method or mark type. The introduction of digital cameras into mainstream society places the field of photographic identification in a unique position to expand rapidly by incorporating citizen scientists in data collection or even crowd-sourcing image comparison tasks. Increasing concern for animal welfare of many vulnerable species ensures that minimally invasive mark-recapture approaches like photographic identification will remain at the forefront of applied ecology in the future.
CHAPTER 2  
ANALYSIS OF HUMAN ERROR RATES RELATED TO PHOTOGRAPHIC IDENTIFICATION IN  
ECOLOGICAL DATABASES: IMPLICATIONS OF INCORPORATING CITIZEN SCIENTISTS  

2.1 Abstract  
Two emerging fields of study- public participation in scientific research and using photographic identification in capture-mark-recapture studies- both remain dependent on the collection of data by observers. The absence of rigorous assessment of observer error-rates causes many scientists to resist citizen science or incorporating citizen-scientist collected data into analyses. Photo-id studies mirror the same need for consistent measurement and documentation of the type and frequency of errors. We pilot a crowdsourcing approach to distributed photographic image analysis by a large number of untrained observers. A specially designed website offers diverse audiences access to images containing matching and non-matching salamander photographs. Observers were asked to make judgments on pigmented pattern information of marbled salamanders (*Ambystoma opacum*). All participants remained blind to the fact that the true response to each pair of images they viewed was already known, and that the study actually sought to evaluate untrained human-error rates in ecological photo-id. False negative errors occurred at a higher rate (16.69%) than false positives (0.09%), but overall all observers combined correctly assess 99.6% of all compared images. The probability of making a false positive error was strongly driven by an observer’s experience (the number of images previously viewed) and an interaction with the number of matches available in the catalog. The probability of making a false negative error was strongly driven by the size of the catalog of images they viewed, the time interval between matching images, the age of the observer, and the number of matches available in the catalog. Synthesis and applications. We
recommend expanded use of citizen science methods as an effective alternative or complement to computer-assisted photo-id. Improved training protocols and incorporation of feedback mechanisms during matching should further reduce the frequency of both error types. Error rate improvements should allow scientists to more readily accept data collected by untrained observers as valid, and will also contribute to improved accuracy of ecological analyses of population size, vital rates, and overall conservation management of threatened or endangered species. The general public will benefit from expanded opportunities to engage with and learn about the scientific process.

2.2 Introduction

Recent reviews of the rapidly expanding field of “Public Participation in Scientific Research” (PPSR) – also called simply “Citizen Science” (Bonney 1996) have praised the production of co-created knowledge between scientists and the general public in projects of unprecedented scope and highlighted the benefits of such activities for scientists as well as the public, but also called attention to important knowledge gaps and future challenges. Academic hesitance to embrace citizen science projects as a valid method of collecting scientific data historically rests on one of these challenges: the question of reliability and observer (data) quality (Oberhauser and Prysby 2008; Silvertown 2009; Dickinson, Zuckerburg, and Bonter 2009). Bonney et al. (2009-BioScience) have proposed that the collection of high quality data hinges on three main components: clear protocols, clear data forms, and providing a support network to participants.

In many cases, technological advances and public access to high speed internet have facilitated the data collection process by enabling the use of online platforms and sharing centers (Bonney et al 2009-BioSci). In fact, the “distributed thinking” concept has capitalized on
the similar idea of “crowd-sourcing” to tap into great expanses of human (and computer) computing power (Hand 2010). Today, a wide array of sophisticated internet-based scientific projects like Foldit, Stardust@home, and Galaxy Zoo solicit community support to successfully solve real-life problems like protein folding, sorting NASA images in search of interstellar dust particles, and classifying new galaxies (Hand 2010). However, training and monitoring observers’ pattern recognition abilities often proves a difficult, but critical task, particularly when determining how to pool the responses of many different volunteers into an accurate, consensus solution (Hand 2010). Dickinson, Zuckerburg, and Bonter (2010) second this challenge and go further to outline the pressing need for “wider assessment of data quality and clarification of the independent effects of professional training, task training, experience with the task, observer age, training duration, mode of training (in person vs. the internet), and variation in [species] detection probability”. Additionally, they raise important concerns specific to projects that deliver materials to observers over the internet, particularly the need to estimate the amount of experience required before data are reliable and the importance of standardizing the amount of effort/time between samples or observers as a mechanism for reducing biases in results (Dickinson, Zuckerburg, and Bonter 2010).

Similar calls for increased accountability and reporting of observer error are mirrored within another rapidly growing field: wildlife photographic capture-mark-recapture (CMR). Increasingly popular as a cheap, non-invasive alternative to applied tagging, photographic documentation (marking) and subsequent identification of individuals by their unique morphological or pigmentation patterns as a method almost always remains dependent on human-visual verification of matches (see Sherley et al. 2010 for exception). This reliance on

---

6 Crowdsourcing has been defined as ‘getting an undefined public to do work, usually directed by designated individuals or professionals’ (Dickinson, Zuckerburg, and Bonter 2010)
human observers that are usually, but not always, trained makes many of the issues common in internet-based and distributed-thinking citizen science projects directly applicable to photo-id.

Of primary concern in both fields are observation errors. When comparing two images during photographic identification an observer can make two potential types of error: a false positive (incorrectly identifying two different individuals as the same identity) and a false negative (incorrectly identifying two images of the same individual as different). Consistent measurement and publication of the type and frequency of errors occurring in most photo-id studies has been lacking (Chesser and McGarigal unpublished; Stevick et al. 2001). Historically, false positives have either been overlooked entirely or dismissed because they were estimated to occur at such low rates as a result of strict processing protocols in wildlife photo-id (Stevick et al. 2001); additionally, they are relatively easy to discover and correct (Huele et al. 2000). False negatives comprise the bulk of the discussions and are the biggest cause for concern because they can only be found by re-searching the entirety of the image catalog, a much more labor-intensive task (Huele et al. 2000; Kelly 2001).

Both types of error create inaccuracies in capture histories that can lead to significant divergent effects on vital rates (e.g., survival), movement (e.g., dispersal), population estimates, and eventually conservation strategies. False positives artificially increase the number of recaptures of tagged individuals in the population, leading to reduced population estimates (Stevick et al. 2001). False negatives fail to document a true recapture opportunity, and thus create additional, incorrect “ghost” identities of the same individual, leading to inflated population estimates (Hastings, Small, and Hiby 2001). There are many plausible (often interacting) causes of errors in photo-id, including: photographic quality (Friday et al. 2000), distinctiveness (Friday et al. 2000; Friday et al. 2008), “tag-ability (markability)” (Hammond 1990), changes in pattern stability over time (Carlson and Mayo 1990; Blackmer, Anderson, and
measurability (roughly dealing with how the pattern is viewed—does it require multiple planes/images?) (Karanth 1995), and observer judgment errors often attributed to fatigue (Sears et al. 1990; Kelly 2001) or inexperience (Agler 1992). Photo-id studies examining error rates usually do so with regard to the size of the image library searched, the number of matches available to find, and documented features of the images used during analysis (Gamble, Ravela, and McGarigal 2008). It is widely accepted that error rates likely increase as image catalogs grow (Morrison et al. 2011).

Particularly with long-term studies, but also dependent on population size and anticipated recapture rates, image libraries have reached sizes too large for humans to reasonably and accurately inventory in search of the same individuals (Arzoumanian, Holmberg, and Norman 2005). Studies confronted with large image catalogs as well as the inevitable time, effort, and budgetary constraints have historically turned to interdisciplinary relationships with computer-scientists and the development of computer-assisted ranking algorithms as the primary solution, and with the goal of improving accuracy (Kelly 2001; Arzoumanian, Holmberg, and Norman 2005; Gamble, Ravela, and McGarigal 2008; Hastings, Hiby, and Small 2008; Sherley et al. 2010). This increase in specialization and complexity is not only expensive financially, but it also tends to isolate and remove untrained observers as well as the general public from subsequent phases in the scientific process (e.g., image handling and processing).

In response to calls from the broader scientific community to expand upon existing citizen science projects and to more thoroughly investigate the ability of observers to collect accurate data over the internet, and in reposition to the increasingly expensive computer-assisted ranking algorithms emphasized in the field of wildlife photo-id, we designed the current study. Our driving question of interest was: can the general public (untrained observers) provide an effective alternative (or complement) to computer-assisted image processing using highly
trained observers? We asked this with practical implications in mind. Our long-term amphibian metapopulation monitoring currently fits the description for a ‘contributory level’ citizen science project. Each summer and fall during juvenile emigration and adult breeding migration events, respectively, of our focal species, the state-threatened marbled salamander (*Ambystoma opacum*), student and community volunteers assist in collecting images and metadata (e.g., length, weight, sex etc.) for each salamander. Outside these intense windows of activity, however, the public are not involved in the scientific process. Our study sought to determine what might happen if these participants were incorporated into the subsequent steps of image comparison and analysis. By offering a series of simple, binary (match/non-match) questions and images to a very large number of people, we hoped to engage a diverse audience and demonstrate the feasibility of citizen science as an effective method in photo-id analysis.

2.3 Materials and methods

2.3.1 Collection of image library

Between 1999 and 2010 more than 12,000 images of individual marbled salamanders were collected as part of a CMR study in western Massachusetts. Salamanders were captured during immigration (pre-breeding) and emigration (post-breeding) events using pairs of pitfall traps located every 10 m. along drift fence arrays that completely encircled each of 14 seasonal ponds (also referred to as vernal pools; see Jenkins, McGarigal, and Gamble 2003 for detailed field protocol). One picture per adult individual per capture event was collected using a camera stand and lightbox designed to improve image quality (see Gamble, Ravela, and McGarigal 2008 for detailed image collection protocol). For the purposes of this investigation into human-errors, we only used a subset of our total image library: images had to be taken during or post-lightbox development (approximately year 2000), and had to have been previously processed by our
computer-assisted matching algorithm (at the time of this study’s design we had only processed through year 2006). These constraints ensured we eliminated drastically poorer quality images taken prior to year 2000 (achieving a rough standardization of photo quality for all the images in this study), and that we had complete, human-verified capture histories for any image selected from this group. From this subset we randomly selected images for incorporation into trials. This resulted in approximately 4,000 unique images being chosen for the work we describe here. An observer completing their entire set of 15 trials (see below) would see a total of 2640 unique images. Each image contained only one salamander, but an observer could potentially see many different images of the same individual salamander within, or across trials.

2.3.2 Study design

Our objective was to display images to participant observers in a manner that allowed us to quantify the effects of a series of predictor variables on overall error type and frequency. We approached study design with an Analysis of Covariance (ANCOVA) framework in mind, hoping that our simple categorical results and interactions would offer structured guidelines/trends for designing future crowd-sourcing studies that maximize accuracy and efficiency of observers. Two of the most frequently cited limitations (sources of error) to wildlife photo-id studies are 1) catalog/library size (the number of images an observer must sort through while searching for matches) and 2) number or proportion of matching images in the library (Kelly 2001; Van Tienhoven et al. 2007). These two factors are directly related to population size, recapture rate and tag-ability, as well as study duration.

Sears et al. (1990) and others have demonstrated that observer fatigue likely contributes to increased error rates if more than two hours is spent viewing images. With this in mind, and also attempting to account for variability in viewing speeds across human
observers, we limited the number of images to be shown in a single sitting to 300. For a set with
the maximum of 300 images, a participant could potentially spend up to 24 seconds on each
image and still finish within the hypothetical maximum time of 2 hours. We recognize that
many observers could easily view far more images than 300 in this allotted time, but previous
experience (M. Chesser unpublished data) informed us that in a large enough sample, some
observers would take significantly longer than a few seconds to make a decision.

With catalog size and number of matches as our primary predictors for error rate, we
designed a fully factorial (3 x 5) set of 15 “trials”- built from every combination of three levels of
catalog size (75, 150, and 300 images) and five levels of number of matches in the catalog
(0,1,2,4,8). Each “trial” (set of images) contained one static reference image that was displayed
alongside a single, randomly selected image from the catalog. With each paired comparison
(reference + 1 catalog image), it was up to the observer to determine if the salamanders were
the same individual (a match) or not (non-match) on the basis of their pigmented pattern
information.

2.3.3 Variables measured

For each observer we recorded age, gender, major, paid/volunteer, whether or not they
had any previous experience comparing salamander pictures (yes vs. no), self-identified learning
style (4 levels, e.g., visual, auditory etc.) and comfort level on the computer (e.g. would prefer to
print images out and look at them vs. comfortable or very comfortable viewing images on a
screen). For each trial we recorded number of matches in the catalog (0, 1, 2, 3, 4, 5, 8)\(^7\), catalog

\(^7\) The study was conceived as only embedding 0,1,2,4, and 8 matches in the catalog, but
unanticipated bugs in the code controlling the way images were viewed on the website resulted
in some participants actually only seeing 3 or 5 (rather than the intended 4 or 8 matches) in one
or more of their trials. Six trials appeared with only 3 matches, and two trials appeared with only
5 matches before the bug was corrected.
size (75, 150, 300), and the proportion of false positive and false negative responses. For each 
comparison we recorded the observer’s response (match vs. non-match), current experience 
level (the number of images viewed prior to the comparison at hand), time to decision 
(measured as the time difference in seconds between when the server executed the command 
to show a different image and the time that the “next” button was clicked again to advance), 
and interval between images (measured as the time difference in days between when the 
reference image and catalog image were collected in the field).

After the trials were completed we documented that no correct “matches” were 
identified when the observer spent less than four seconds to make a decision. Taken at face-
value, this seemed to indicate that correctly identifying a match required a minimum of four 
seconds (median time to correct decision was approximately 15 seconds). However, without 
measuring internet speed for every instance of computer use (and it was known to vary among 
instances/locations), unfortunately we could not be sure if this “time to decision” variable 
accurately measured the time an observer spent viewing the pair of images, or if, in fact, it also 
included the time it took their computer to load the images. Therefore, we decided to drop 
“time to decision” as a predictor variable in the statistical models below.

In addition, ideally the “observer’s current experience” would have had a maximum of 
2640 images viewed (15 trials with one reference image each = 76x5+ 151x5+ 301x5); however, 
we were forced to assign new “make-up trials” to some participants after discovering a bug in 
the code that controlled the way images were viewed on our website. Consequently, after 
completing their “make-up trial(s)”, some observers viewed more than 2640 images and some 
of the trials ended up with three and five matches in the catalog. The number of make-up trials 
assigned varied depending on how far along through their original set an observer was at the 
time we discovered the bug in the code.
2.3.4 Platform design

In order to provide observers with easy access to these sets (trials) of images and to record their responses, we designed and built a website, mandermatcher.com. The beta version of mandermatcher.com was not intended to be an educational tool (it did not incorporate K-12 or university level curricula, games, or user feedback), but was simply the easiest method of sharing a large number of images with a large number of people on radically different schedules. Being freely available whenever and wherever internet was available maximized participation in the study by allowing participants to work from the comfort of their homes, coffee shops, or on campus, and at all hours of day or night when their schedules allowed. Almost all participants praised how easy it was to do this work, so much so that many requested additional opportunities to look for matches after termination of the study.

During their first visit to mandermatcher.com each participant generated a login and password that were used to assign them sets of images and to track their responses through the system. Users then took a simple background questionnaire, and similar to the protocol used by Westphal et al. (2006) with their Stardust@home program, our participants read over 5 brief lessons while viewing accompanying examples of matching salamander images (e.g. “Lesson 2: Salamanders might experience weight gain or loss between photographs. This means that you should focus specifically on the pattern itself rather than body size differences (or how zoomed in the animal appears) between two images. Because these animals are photographed on their way into and out of the vernal pools in which they breed, weight loss in females is largely attributed to egg deposition. However both sexes can experience weight loss as a result of reduced feeding behavior during this active time.”). After reading over the lessons, observers completed an exercise called “Practice Matching” in which they viewed and made guesses on 10 pairs of images. During this practice session the screen operated identically to the “Real
Matching” of their 15 actual trials, except for two important differences: 1) during the practice each pair contained a different reference image, and 2) after submitting their response, observers received instantaneous feedback telling them if they had gotten the response correct or incorrect. At the end of the practice session observers were told their overall score, and it was submitted via email to website administrators. All users viewed identical practice matches, and only upon completion of BOTH the questionnaire and the practice matching were they granted access to the “real matching” portion of the website. Unlike Westphal et al. (2006), who only allowed people scoring 8/10 or better to access their images, we did not discriminate (hoping to model the full range of varying ability levels in the broader population) and allowed everyone to advance to “Real Matching”.

Behind the scenes, we constructed a series of folders, each containing all the images necessary for a trial. For each participant, the 15 trials in our factorial block were assigned in random order. At the beginning of each trial, observers were forced to review the instructions page that reminded them of 1) instructions for navigating around the website and submitting responses, 2) their task for each pair of photos- “to determine whether the two salamanders displayed are the same (a match) or different (non-matching) individuals”, 3) the importance of taking their time and being confident in their decision before they clicked the “next” button to advance (that this was not a race- we were recording time for the sole purpose of being able to create more appropriate future trial sizes for given time periods), 4) the importance of completing a trial in a single sitting, and 5) to use the “Pause” button at the bottom of the screen if they needed to take a restroom break.

At the top of the screen in each trial (Fig. 2.1) observers could track their progress through the trial in which they were working (e.g. “Viewing image 7 of 75”), as well as see (upon finishing a trial) how many trials remained to be completed (e.g. “You just completed set 10/15;
Five sets remain to be completed.

Within each trial, catalog images were displayed to the observer in random order (the reference image was always visible on the left side of the screen). Because the body of the salamander in a catalog image was not necessarily displayed parallel to that of the reference, a “Rotate” button at the top of each image allowed users to rotate either the reference or the catalog image for ease of comparison. Though we suggested this viewing modification and reminded them of the rotate feature at the onset of each trial, we have no way of knowing if a participant actually changed the orientation of either image during viewing. To submit their decision for each comparison, observers noted a match by placing a “check” in the “Match” box underneath the pair of images; non-matching pictures of different salamanders were noted simply by advancing to the “Next” catalog image. Observers were told to take as much time as necessary in order to make the correct decision because there was no possibility of going backwards to double-check a response after submitting an answer. We intentionally designed the system without a “Back” button in order to capture observers’ initial (and final) responses. At the conclusion of a set of images, observers were given the option of logging out of the system or beginning another trial.

2.3.5 Participants

In total, 63 unique participants contributed to the data we report here. These observers were a combination of work-study and volunteer students. Our work-study students came from a wide variety of majors, including Japanese, math, dance, psychology, and “undecided” just to name a few, and several were non-traditional students who had returned to school after working for several years. Volunteers came to us through word of mouth and in response to solicitation (for résumé boosting activities) at various undergraduate classes in the life-science departments of UMass Amherst. Participants varied in age from 17 to 47, with an average age of
20.87 years, and a median age of 19 years. Forty-three participants were female, and 20 were male. Forty-five participants were paid (for at least a portion of their time) and 18 participants were volunteers. In all cases, correspondence was initiated with a brief background email explaining the long-term mark-recapture study, its goals, and conservation implications. No information was shared with students about the true goal of the study - to estimate the error rate associated with untrained observers analyzing images. At all times, students worked under the impression they were finding “novel” (new) matches in our extensive photo library; they remained blind to the fact that we already knew the correct answers to all of the comparisons they were examining. To ensure our observers felt supported and in contact with the researchers they would be assisting remotely, we gave participants our contact information and instructed them to email us if they encountered any questions or problems (Bonney et al. 2009-BioSci).

Upon beginning the 15 trials, 24 of our 63 participants had some previous experience working on the computer to verify matching salamander images using our ranking algorithm. The remaining 39 participants had zero previous experience with salamander photo-id when they began our study. A small number of these ‘novice’ participants requested to meet in person before beginning work online. Because we wanted all new participants to start from the same baseline (zero experience with photo-id), when we met with them for approximately 15 minutes, we deliberately avoided speaking about photographic mark-recapture, electing instead to only speak to them about vernal pools and amphibians in general before giving them the web address for mandermatcher.com. Like most citizen science studies, in every instance participants operated independently and wholly without supervision while using mandermatcher.com (Trumbull et al. 2000). Researcher feedback was not available during matching. In spite of this, several students self-recognized their own mistake (usually failing to
acknowledge a match) and contacted us the “scientists” so that we could correct the error in their specified trial. When this happened, most students told me they had been moving too quickly through images, and did not click the “Match” box before advancing. Though the email was noted, no action was taken to correct the observer’s response because we were interested in recording the overall “initial response” by a large number of untrained observers and we wanted to ensure that all observer responses were treated uniformly. By comparing an observer’s response to the correct answer we determined their type and frequency of errors.

2.3.6 Statistical analysis

Extenuating circumstances prevented some participants from completing their full block of trials. Rather than exclude incomplete blocks from analysis, we elected to shift our statistical approach from an ANCOVA to one of logistic regression, which is more flexible in this regard. In addition, logistic regression allowed us to assess error rates using both the trial and individual image comparison as the observational unit within the same modeling framework. To assess the factors influencing error rates, we subset the raw data into two sets: 1) false positive (FP) set, which contained only catalog images that were true non-matches (and thus could result in potential false positive errors); and 2) false negative (FN) set, which contained only catalog images that were true matches (and thus could result in potential false negative errors). All analyses were done separately for these two data sets. For each data set (FP and FN), we conducted, mixed effects logistic regression analysis at two levels: 1) comparison level, in which each pairwise comparison of reference image and catalog image was treated as an observation; and 2) trial level, in which each trial, consisting of a single reference image and a set of catalog images, was treated as an observation. At both levels the error was binomial. At both the trial and comparison levels the complete list of all possible predictors included variables associated
with the observer: age, gender, major, learning style, comfort level on the computer, if the person was volunteering or paid, and if they had previous experience viewing salamander images before. At the trial level, the response was proportional (# errors given # comparisons), and additional predictors included catalog size, number of matches in the catalog, and the average of an observer’s current experience. At the comparison level, the response was binary (error vs. correct), and additional predictors included catalog size, number of matches in the catalog, observer’s current experience, observer time to decision, and interval between images. Note, because the interval between images variable was calculated on a pairwise image basis, averaging across all images in each trial would have been nonsensical (and would likely yield roughly the same values since all images for each trial were drawn at random from the greater image library); thus, we did not incorporate interval between images as a predictor at the trial level.

All analyses were conducted using the open-source, statistical software package “R” (R Development Core Team 2008) and the generalized linear mixed effect modeling package “lme4” (Bates, Maechler, and Bolker 2011). At both levels, we used a modified top-down strategy to select the optimal set of variables from the complete list (see above) (Diggle et al. 2002; Zuur et al. 2009). Briefly, because it was impractical (i.e., too many explanatory variables, trouble with interactions, and numerical problems) to fit the most complex model, we started with a model that included as many possible variables and interactions as possible. Prior work with linear models informed us that accounting for random differences between observers was paramount, and thus our baseline generalized linear mixed effect models always included a random effect for observer (a varying intercept model).

At both levels, we began by focusing on the fixed effects component of the model and incrementally adding predictors and logical interactions (Zuur et al. 2009). We built up from
what we deemed to be our “main predictors” and their possible interactions: catalog size, number of matches in the catalog, and observer’s current experience (or average current experience at the trial level). At every level and error type, we attempted a three-way interaction among these first (but always failed to fit the model). We then tried combinations of two-way interactions, and adding additional predictors (e.g., interval, age of observer, etc.). We failed to fit models that included major and learning style as fixed effect predictors without creating egregious correlations with other variables. Once we had the fixed effect component of the model as complex as possible, we began adding predictors into the random component of the models. Binary predictors at the observer level (e.g., gender, previous observer experience, volunteer/paid) were incorporated into the random intercept component of the model, while all other predictors were incorporated into the slope component of the random effects (e.g., catalog size and number of matches in the catalog).

The fixed component of the fullest model possible for FP errors included catalog size, observer time to decision, previous experience, age, gender, and an interaction between an observer’s current experience and the number of matches in the catalog; the random component included only an observer effect (varying intercept). The fixed component of the fullest model for FN errors included an interaction between observer’s current experience and the number of matches in the catalog, as well as an interaction between catalog size and observer’s age, and interval between images, previous observer experience, volunteer/paid compensation, and comfort level on the computer; the random component included: observer’s current experience and the number of matches in the catalog (varying slope components) and an observer effect (varying intercept).

From these full (most complex models) we then manually dropped variables in an attempt to minimize AIC (Akaike 1974) and find the most parsimonious model that contained all
significant predictors. When necessary, we centered variables by subtracting the mean or square-root transformed them to aid the numerical optimization. Upon reaching the optimal combination of variables, we refit the model using restricted maximum likelihood and plotted the normalized residuals against both the fitted values and each predictor variable in order to assess the model.

We identified several sources of uncertainty in our analytical strategy. First, we recognized the uncertainty associated with evaluating error rates at the *comparison* level versus the *trial* level. Both levels sought to assess the factors influencing error rates, but did so at different scales of observation, and neither scale was deemed inherently better than the other. Second, we recognized that it was possible to treat both catalog size and number of matches in the catalog as continuous variables in the analysis when in fact they were designed to be categorical variables (factors), since they are both inherently continuous phenomena. Note, treating these design variables as continuous allowed us to more clearly visualize the random effects associated with unique observers and to portray the relationships with error rates as continuously varying, which is ultimately more intuitive for these data, but at an unknown cost. Lastly, we also noted that two observers (#34 and #13) were responsible for 80% of the false positive errors and that two separate observers were considerably younger or older than the rest of the observers. Given these sources of uncertainty, we analyzed the data under a variety of modeling scenarios and assessed the robustness of the results by looking for consistencies across scenarios in terms of whether variables consistently remained in the best models as well as the direction and magnitude of their coefficient estimates. Modeling scenarios included all possible combinations of: 1) building models using either the *comparison* or the *trial* level data, 2) treating the design variables (catalog size and number of matches in the catalog) as either continuous or categorical (factors), 3) including or excluding the two observers (#34 and #13).
that were responsible for 80% of the false positive errors, and 4) including or excluding the highest and lowest aged observers. For simplicity, we present the detailed results of the first modeling scenario (i.e., first choice in each of the factors above) and reserve the other scenarios for an assessment of the robustness of the results.

2.4 Results

We recorded a total of 144,373 responses from 63 unique observers. Correctly identified non-matching images accounted for 98.2% of the data, and correctly identified matching images accounted for 1.4% of the data for a combined total of approximately 99.6% correct across all observers (Fig. 2.2). All observers combined committed 533 total errors. False positive errors occurred only 0.09 percent of the time (127/(127+141815)). False negative errors occurred 16.69 percent of the time (406/(406+2026)). Though they had the potential to occur much more frequently (98.3% of comparisons were true non-matches), false positives actually occurred less frequently than false negatives in our data, indicating that most observers seemed to be conservative in their responses of “match”. Viewing the total number of errors by type for each observer reveals dramatic differences between individuals, and begins to point toward the importance of incorporating observer differences into the random effect component of our models. Two observers (ID #34, 20 FP errors and #13, 81 FP errors) account for the vast majority (80%) of all the false positives in the dataset (Fig. 3.3). Outside of these two, perhaps overly confident, observers, FP errors occurred rarely and at low frequencies for each person. On the other hand, all but 11 observers committed at least one FN error, and false negatives tended to occur at higher frequencies per person than false positives (Fig. 3.3).
2.4.1 False positive errors

\[ FP \sim 1 + \text{sqrtobs.expC} \times \text{matchesC} + \text{preexp} + (1|\text{obs}) \]

The model given above was selected as the "best" model based on our model selection procedure for FP at the comparison level treating catalog size and number of matches in the catalog as continuous variables and including all observations. In this model, the fixed effects included an interaction between the observer's current experience (centered and square-root transformed) and number of matches in the catalog (centered), and the observer's previous experience (binary). There was also a random effect for observer, which allowed the intercepts of the fixed effects to vary among observers. The results indicate that observers are overall more likely to commit a FP error the fewer the number of images they have seen previously (i.e., the lower their current experience level) \( (P<0.001) \). In other words, each additional image seen dramatically reduces the probability of an observer incorrectly stating that two different salamanders are in fact the same. The results also indicate that the number of matches in the catalog (matchesC) has a weak overall negative effect on the probability of false positives \( (P=0.04) \). The more true matches included in the trial, the fewer false positive errors committed. However, there was significant \( (P=0.001) \) negative interaction between observer current experience and number of matches in the catalog, which indicates that at low levels of observer current experience (i.e., when observers first begin matching images) the probability of making a FP error increases with the number of true matching images in the catalog, but that after observers gain a lot of experience they make the greatest number of errors when the catalog contains the fewest number of matches. Observer's previous experience (preexp, which measured if an observer had any prior experience working with photo-id at the start of their trials) did not have a significant predictive relationship with probability of false positives on its own \( (P=0.2) \), but it remained in the final model because the model's overall AIC value was more
than two-units improved (lower) over a model in which it was removed. This indicates that the presence of the “preexp” variable may synergistically improve the performance of one of the other variables in the model. Lastly, when predicting false positives, the best model was one that incorporated a random intercept for each unique observer (obs).

### 2.4.2 False negative errors

\[
FN \sim 1 + \text{matchesC} + \text{catalogC} + \text{ageC} + \text{intrvlC} + (1 + \text{sqrtobs.expC} \mid \text{obs})
\]

The model given above was selected as the "best" model based on our model selection procedure for FN at the comparison level treating catalog size and number of matches in the catalog as continuous variables and including all observations. In this model, the fixed effects included number of matches in the catalog (centered), catalog size (centered), observer age (centered) and the interval between reference and catalog images (centered). The model also incorporated a random slope (observer current experience, centered and square-root transformed) and intercept for each observer. The results indicate that observers have a higher probability of failing to see a match the higher the number of comparison images they view in a single sitting (larger catalog size) (P<0.001). The results also indicate that the amount of time that has elapsed between when an animal was first captured, and when it was photographed again is strongly predictive of the ability of an observer to recognize the individual as a match; in other words, the longer the time interval between matching images, the more difficult it is for an observer to recognize the match (P<0.001). For our sample of 63 participants, the age of an observer was also a significant predictor of their probability of committing a FN error (P=0.0028). Older observers were less likely to miss matches than younger participants. However, it should be noted that we had a very unequal age distribution in our sample; 53 of 63 observers were age 21 or younger with only 10 observers representing ages 22 through 47.
Similar to the trend seen in the FP model, the number of matches in the catalog was also weakly predictive of the number of FN errors committed (P=0.019). The higher the number of matches available, the smaller the probability an observer would commit a FN error. When predicting FN, the best model was one that incorporated a random slope (observer experience) and intercept for each observer.

2.5 Discussion

2.5.1 Predictors of errors using untrained observers

2.5.1.1 False positives
An interaction between observer current experience and the number of matches in a trial (i.e., opportunities to view the same pattern again) is the primary predictor for increased probability of making a FP error. We can think of observer current experience (number of images viewed) as roughly translating into an exposure level to a diversity of patterns. Greater exposure to a wide variety patterns leads to an improved ability to recognize subtle variations in patterns that may once have appeared the same. Following this same line of reasoning, we can begin to understand one possible factor contributing to this interaction. Without experience viewing salamander photographs, an observer lacks an appreciation for the broad spectrum of patterns that exist in a population. Under this initial condition, being exposed to high numbers of true matches in their catalog seems to trick observers into thinking (or simply reacting) that more matches exist than actuality - any pattern that appears remotely similar to the reference image in question, might receive a response of “match”. This period of “trigger-happiness”, when an observer clicks “match” before adequately identifying the comparison, quickly subsides as an observer views more images. As observers become quite experienced, they tend to only make FP errors at the lowest number of matches in the catalog (indicating they want to find
matches even when none exist). The fact that observers who began our 15 trials with some baseline “experience” using photographic identification still followed this same inverse relationship between observer experience (measured during the trials) and number of matches in the catalog is indicative of a secondary cause leading to this relationship. If experience was the only driver of this trend, we would expect observers beginning with a baseline understanding to be immune to this phenomenon. However, this is not the case, perhaps indicating that the relationship we see is also reflecting some sort of adjustment period to the nature of the trials and web-interface. Numerous studies have observed similar “first year” or “learner” effects during data collection (Dickinson, Zuckerberg, and Bonter 2010). After their first trial, observers (regardless of prior experience) know more of what to expect from the system.

2.5.1.2 False negatives

The two most significant predictors of false negative errors were catalog size and time interval between paired reference and catalog images. Due to the nature of this work being unsupervised, we cannot be certain that observers viewed all images in a trial in a single bout of work (as they were instructed). Thus, we must exercise caution when considering the positive relationship between catalog size and number of FN errors. However, previous studies have observed a similar trend, and often attribute this higher probability of making mistakes to fatigue (Sears et al. 1990; Kelly 2001). It is also possible that because there is potential confounding between the number of images viewed (catalog size) and observer current experience, the increase in probability of making a FN error we see could be due in part to differences in how observers respond to increasing amounts of experience (some make fewer errors, others make more).
The time interval variable is likely representative of two sources of difficulty for the observer: 1) potential changes in photographic quality and 2) potential for pattern change (Yoshizaki et al. 2009). Rapid technological advancements over the past decade have meant yearly improvements to digital (and film) cameras. To give an example from the context of our own long-term study, images taken prior to 2003 were collected using a Mavica digital camera (Model MVC-FD83) with a 0.8 megapixel capacity. Between 2003-2009 images were collected using Canon Powershot A70 and A520 cameras with 3.2 and 4 megapixels. Today, images (not included in this study) are collected using an Olympus Stylus 1030 SW (shock and waterproof) camera with 10.1 megapixel capacity. However, despite these technological advances, we believe that even the oldest camera used in this study (0.8 megapixels) was more than sufficient to capture both the fine and coarse features of the marbled salamanders’ patterns (Gamble, Ravela, and McGarigal 2008). Moreover, we feel that our decision to only use photographs taken after the development of the lightbox in our study provided a rough standardization of photographic quality in terms of lighting, removal of extraneous background information, and angle to the subject, eliminating the bulk of potential sources of error related to photo quality. Accepting this line of reasoning, we are left with pattern change as the cause for failure to identify the same salamander across time.

An animal’s pattern may change for a variety of reasons. For marbled salamanders, it appears that injury and the subsequent healing process are a primary cause of change to the pigmentation pattern. Anecdotal evidence from large numbers of long-term recaptures of adult marbled salamanders in our image library indicates the more severe the injury (and less ‘clean’ the wound, e.g. an animal appears chewed on by a predator), the less likely an animal is to heal quickly and without impact to their pigmented pattern. Deep scar tissue generally lacks contrast (appearing an opaque gray rather than the black or white pigments characteristic of this species)
and can remain visible for many years, whereas small scrapes or cuts appear to heal quickly and leave little lasting alteration in pattern (M. Chesser, unpublished data). Studies with marine mammals in particular (sea otters, manatees, dolphins etc.) have been quick to recognize the challenges scar tissue may pose to individual identification (Langtimm et al. 2004; Gilkinson et al. 2006). Growth, genetic, and environmental triggers are also likely candidates affecting pattern change in animals (Forcada and Aguilar 2000; Gowans and Whitehead 2001). For example, aging individuals in a population can show directional shifts in coloration or scarring patterns (Mediterranean monk seals (Monachus monachus) - Forcada and Aguilar 2000; Alaskan harbor seals (Phoca vitulina richardii) - Hastings, Hiby, and Small 2008; and eastern tiger salamanders (Ambystoma tigrinum tigrinum) M. Chesser pers.obs.). Ultimately, matching images taken more closely in time usually positively affect recapture rates and reduce probability of observer errors (Blackmer, Anderson, and Weinrich 2000; Gowans and Whitehead 2001; Hastings, Hiby, and Small 2008).

The moderately significant negative relationship between observer age and probability of FN errors should be viewed with caution because our sample was skewed by high numbers of young participants. For this pilot study, our selection of participants could be described as opportunistic. Specifically, each observer was a student of the University of Massachusetts; we did not attempt to find or incorporate participants from a broader age base and/or educational background. We recommend that future studies hoping to measure the effect of age should consider a stratified random sample within each age bracket, and should consider expanding to well below (possibly even middle school level) and above the age range (possibly including tech-savvy senior citizens) of participants in our study (following in the footsteps of many of the projects available to the public through citizenscience.org).
Lastly, though all catalog images were randomly displayed to observers during their trials, a higher number of true matches in the catalog likely built a level of familiarity with a particular pattern (Duncan and Humphreys 1989). Having the opportunity to seeing the same animal (pattern) multiple times appears to marginally reduce the probability an observer will miss a match.

2.5.2 Importance of accounting for observer differences

When predicting FP and FN errors, the best models always included, at a minimum, “observer” as a random intercept effect. Building a model that allowed differences between our 63 observers to capture some of the variance around our predicted values usually reduced overall AIC values by half. It is likely that the small number of FP errors prevented us from being able to keep any other variables in the random effects component of this model. With more abundant FN data, random differences in how individuals were affected by their current observer experience throughout the trials (random slope component) helped to remove additional unexplained variance from the best model. Our decision to analyze and display catalog size and number of matches as continuous variables (rather than categorical) was partly due to the clarity provided in graphing predicted values and relationships with respect to each individual as a separate line. Visualizing the (at times) dramatic differences in how observers react to different image library conditions is a powerful reminder that future studies should account for this variation prior to conducting ecological analyses. Deciding how to select and/or eliminate responses from particular observers will be important for maintaining accuracy when using untrained observers to collect photo-id data. Future studies incorporating a broader spectrum of participants (ages, comfort levels on the computer, backgrounds etc) from the general public can expect an even wider range of variation. To ensure a high level of accuracy,
researchers might consider designing systems built around a “majority rule” taken from a group of people or by developing a scoring system to weight responses by an observer’s previous experience or error rate before incorporating them into final recapture data.

2.5.3 Prediction scenarios

Using our best models as predictive tools, we created a series of hypothetical scenarios to examine the effect of changes in the image library and observer current experience. In the first scenario, we examined differences in the probability of making a FP error as a result of changing observer current experience (# of previous images viewed) for both novice and previously experienced participants under conditions favoring the largest number of FP errors -- a catalog containing zero matches (Fig. 2.4). Having prior experience with photo-id (vs. being a novice) appears to reduce the overall probability of making a FP error. Recall that previous experience was not a significant variable in the model even though it was retained in the best model; it likely affects FP error rate through a more immediate measure of the observer’s current experience (number of images seen) during this study. Note that the overall FP error rate for the population is incredibly low, with most novice observers having less than a 2.5% chance of making a FP during the initial and most error-prone trial of the study. A small number of novice participants would be expected to perform much worse, potentially making FP mistakes 27% of the time in the beginning. But, the probability of making a FP decreases rapidly with each passing image even for these underperformers (overly confident participants) (Fig. 2.4).

In the second scenario, we examined differences in the probability of making a FP error as a result of changing the number of matches in the catalog for a hypothetical novice observer (i.e., no prior experience) with no current experience and having already viewed an average
number of images (Fig. 2.5). Though we have elected to show only the results for a hypothetical novice observer, the plots for observers with prior experience follow the same general trend but over a much smaller range of probabilities. The difference between the plots of no current experience and average current experience (Fig. 2.5) reveals the interaction that exists between observer current experience and number of matches in the catalog. Specifically, with no current experience an observer has a greater probability of making a FP error as the number of matches in the catalog increases (Fig. 2.5). However, this reverses (and becomes trivial) by the time an observer has reached average levels of experience viewing images (Fig. 2.5). The steepness of the learning curve for observers with little or no current experience suggests that more thorough training (e.g., giving more examples of false positives including highly similar patterns in different individuals during practice matching), or incorporating feedback early on may be two potential mechanisms for preventing false positive errors in future photographic identification studies. One practical option for providing early or periodic feedback would be to intersperse known matches among novel comparisons, so that at a set interval (e.g., every 5 or 20 images) the observer could receive feedback. Receiving information that confirms a response is correct or incorrect would provide an opportunity to inform “trigger-happy” or overconfident observers of their mistakes earlier on, allowing them to alter their behavior accordingly.

In a third set of scenarios, we examined the probability of making a FN error under the best and worst case scenarios (Fig. 2.6). Under the worst-case scenario (Fig. 2.6 left side), the variable plotted on the x-axis was varied systematically while keeping all other variables at the worst match-finding conditions: largest catalog size (300 images), smallest number of matches (1, or a 0.3% recapture rate), and longest time interval between matching images (1845 days). Similarly, under the best-case scenario (Fig. 2.6 right side), the variable plotted on the x-axis was varied systematically while keeping all other variables at the optimum match-finding conditions:
smallest catalog size (75), largest number of matches (8, or a 10.67% recapture rate), and shortest time interval between matching images (0, indicating individual salamanders entered and exited pool basins within the same 24 hour period). In both scenarios, we assumed that observers were untrained (i.e., had no prior experience), had average current observer experience, and average age.

Under the worst-case scenario with varying time interval between images, the population probability of making a FN error more than triples (from 0.10 to 0.36) as the time between images lengthens (Fig. 2.6A left), and for many observers, this increase is only pronounced at relatively long intervals, adding support for the theory that observers can more readily identify changes occurring over short periods than changes occurring over longer periods between photographs (Blackmer, Anderson, and Weinrick 2000). Under optimum match-finding conditions, the increase in probability of making a FN error is much smaller (from 0.04 to 0.17) as the time between images lengthens and maintains a very shallow curve over a much longer time span (Fig. 2.6A right), suggesting that higher recapture rates or restricting the number of images an observer sees at a time might allow for images from longer-term studies to be viewed more accurately.

Under the worst-case scenario with varying number of matches (which translates into recapture rates varying from 0.33 to 2.6%), the population probability of making a FN error decreases 10% from 0.37 to 0.27) as more matches are available (Fig. 2.6B left). Under the best-case scenario, the decrease is only 2% (from 0.06 to 0.04%)(Fig. 2.6B right). These results indicate that the number of available matching images is more important to accuracy when catalog sizes are large and time between matching images is expected to be long. Also, all observers can be expected to commit fewer FN errors the higher the recapture rate per individual in the population (i.e., the greater the number of matching images per individual).
Under the worst-case scenario with varying catalog size, the population probability of making a FN error increases nearly 13% (from 0.24 to 0.36) as catalog size quadruples from 75 to 300 images (Fig. 2.6C left). Under the best-case scenario, the increase is only 3% (from 0.04 to 0.07)(Fig. 2.6C right). These results indicate that catalog size is a more important variable to consider when the number of matches in the catalog (and thus recapture rate) is expected to be low, and the time interval between recaptures may be relatively long. However, regardless of number of matches or time between them, all observers can be expected to commit more FN errors the larger the catalog of images they are expected to view in a single sitting.

With varying observer current experience, there is no population-level change in the probability of making a FN error predicted under either scenario (Fig. 2.6D). In both scenarios, observers respond differently to increases in experience: some observers improve, reducing their probability of committing a FN error as they see more and more images over time; some observers have virtually no change in their probability of committing a FN error; and other observers get worse, increasing their probability of committing a FN error the more images they view. Under worst-case scenario conditions the population is expected to commit a false negative 35.9% of the time, with some observers approaching nearly 100% probability of making a FN error. The population-level probability of making a FN error is much lower (4%) under best-case conditions. These results indicate that a smaller catalog size, higher number of matches, and shorter time intervals between matches are conditions that can drastically reduce the probability of untrained observers committing FN errors. However, the fact that there is no population-level change in probability of making a FN error suggests that regardless of observer experience, all observers are equally capable of committing this type of error.
2.5.4 Robustness analysis

Our robustness (or modeling uncertainty) analysis examined the impact of three major decisions regarding the statistical model: 1) conducting the analyses with comparison- or trial-level data, 2) treating the design variables as continuous or categorical, and 3) deciding whether to include or exclude extreme values with regard to observer age and/or number of false positive errors committed.

For our analysis of FP errors, the best models constructed using comparison- and trial-level data had coefficients in the same direction and nearly identical in strength for observer current experience, number of matches in the catalog, and previous experience. At the comparison-level, treating number of matches as a factor (rather than continuous) and excluding the two observers who committed 80% of the FP errors reduced the significance of observer experience. At the trial level, excluding these extreme individuals removed the significance of observer current experience altogether in the model. Dropping age outliers had no effect on the models at either level. Treating number of matches as a factor at the trial level lowered its own significance as a predictor. Overall, the results were remarkably consistent among the modeling scenarios.

For our analysis of FN errors, the best models constructed using comparison and trial level data differed with regard to which variables remained in the fixed effects component of the model; at the comparison level, these included: number of matches, catalog size, observer age, and interval between images; at the trial level, these included: observer current experience, catalog size, and observer age. Both comparison and trial levels included observer experience and observer as the only random effect terms in the model. Note, we omitted interval between images as a predictor at the trial level because it is meaningless when averaged across a trial. The variables in common to the models at both levels had coefficients in the same direction and
nearly identical in strength. Due to the different terms in the models at these two levels, we also ran the best trial-level model (terms) using comparison-level data, and found no significant differences in the strength or direction of predictive relationships. Dropping age outliers from the models at the comparison level had no effect on the significance or direction of any term in the model; however, when the model built using the trial-level data was run using the comparison-level data without the age outliers (17 and 47 year olds) there was a slight increase in the strength of the predictive relationship of catalog. Overall, the results were remarkably consistent among the modeling scenarios.

2.5.5 Website design and future modifications

Building from what we have learned through this study about likely predictors of FP and FN errors, we believe that future work attempting to refine accuracy of untrained observers should attempt to follow a logistic regression framework. Designing a website and data recording platform that can measure an observer’s responses for each incremental increase in catalog size (especially at less than 75 images, which is where crowd participants are likely to be participating) should increase the amount of data collected (more people are likely to commit to viewing fewer images) and facilitate statistical analyses. In addition, we recommend designing a single-comparison “back” button (blocking users from clicking multiple times to check the whole library) which would allow an observer to correct a single mistake if they recognize it immediately after making an incorrect decision. We also recommend incorporating more background information on the front-end of the website, and potentially building in games and/or features that meet educational curricula objectives into more of a collaborative (rather than a contributory) citizen science project.
2.5.6 Applied research and management implications

Untrained observers in our study correctly identified 99.6 percent of paired comparisons of marbled salamander images, an accuracy level that definitely merits future expanded use of citizen scientists in the image-analysis phase of photo-id. Similar to previous studies, FN errors occurred more frequently (16.69%) than FP (0.09%) for our observers. Since FN are much more costly to correct (because an observer must re-survey the entire catalog) than FP (where an observer would simply need to re-examine the matches in a single capture history), renewed effort should be given to designing future conditions to limit their probability. One possible suggestion would be to design new “training modules” to expose participants to images that clearly demonstrate the potential for pattern change and alert them to be on the lookout for this phenomenon during matching. Our model for FN errors suggests that catalog size (number of images shown to an observer at a time) and time interval between matching images are the two most significant variables affecting probability of a FN error. The number of matches, age of an observer, and observer characteristics (e.g., experience level) are also significant predictors. Smaller catalog sizes, higher numbers of matches, shorter time intervals between matching images, and older observers decrease the probability of FN. The relationship with observer experience varies at the individual level, with some people doing remarkably better (zero errors) than others (87 errors), but with no statistical difference from zero at the population level, we are hopeful that observer experience is not a prerequisite for high accuracy in photo-id. Designing future citizen science based photo-id studies that improve training for pattern change, limit the number of images an observer views at a time, and that contain recaptures of individuals on a short-term (yearly) basis should reduce the amount of FN in the data.

FP errors occurred at an extraordinarily low level (only a 0.28 probability in the worst case) in the population of observers, but three observers in particular accounted for 87.4% of all
FP errors, indicating that there may be variation either in confidence level when responding or in psychological approaches to pattern recognition. Our best model indicated an interesting interaction between observer current experience and high numbers of matches in the catalog. Inexperienced observers (or those who had only viewed very small numbers of images) committed more FP errors if shown more true matches in their catalog. At medium and high levels of observer current experience, this relationship was reversed, with fewer FP being made the more true matches were available in the catalog. This reversal in relationship suggests the need for improved training with/exposure to more examples of FP prior to beginning official viewing of images. This could be achieved by modifying the number and type of “practice” matches displayed to interested participants, or possibly designing a feedback mechanism at the early stages of image-recognition.

Our overall results demonstrate the potential for this method of incorporating large numbers of untrained observers to provide a low-cost alternative (or addition) to development of a computer-assisted ranking algorithm. This pilot study can serve as a controlled precursor to future crowd-sourcing applications as we were able to demonstrate great success in making images readily available online for users to access at their leisure, and the relative ease with which a binary (match/non-match) pairwise decision could be answered by a great variety of different users.

The original full-factorial block, randomized assignment of trials, and careful notations of covariates in our study allowed us to model changing probabilities of FN and FP errors. Quantifying the conditions that proved challenging for members of the public to contribute to image analysis in long-term photographic CMR studies was an important first step in determining the feasibility of collaborative citizen science projects of this nature. Our findings offer great promise (99.6% correct) and offer a framework of variables for researchers to
consider when building future studies involving pattern recognition, computer-access, and contributory/collaborative scientific research. As our project looks to move from a contributory to a collaborative level project by incorporating public support and input into a broader range of steps in the scientific process, we hope our successes will inspire other researchers to consider the scientific and societal education benefits possible from allowing the general public to participate in real scientific projects.
Figure 1.1: Screenshot from “real matching” on mandermatcher.com. With each comparison an observer determined if the salamanders were the same individual on the basis of their pattern information. The reference image on the left was constant throughout an entire trial. The comparison image on the right changed every time the user clicked the “Next” button at the bottom of the page.
Figure 2.2: Raw observer response data subdivided by the known, true relationship between the reference and comparison images. For 143841 (141815 + 2026) out of a possible 144373 comparisons, observers correctly identified the comparison image (99.63%); in 533 (127 + 406) comparisons, observers made an error (0.37%). In order to display all results on the same figure, please note that we have inserted a gap in the y-axis.
Figure 2.3: The total number of errors of each type observed per unique observer (all trials combined). False negatives are much more common and occur more frequently than false positives, which occur rarely except in higher frequencies for observers #34, and #13.
Figure 2.4: Hypothetical scenarios for observers from the general public committing false positive (FP) errors while participating in ecological photographic identification. We examined differences in the probability of making a FP error as a result of changing observer current experience (# of previous images viewed) for both novice and experienced participants. The thick blue line (very close to zero) represents the population average, and thinner multi-colored lines represent individual observers.
Figure 2.5: Hypothetical scenarios for observers from the general public committing false positive (FP) errors while participating in ecological photographic identification. We examined differences in the probability of making a FP error as a result of changing the number of matches in the catalog for a hypothetical novice observer with no current experience and having already viewed an average number of images. The probability of making a false positive after viewing the maximum number of images is essentially invariant from the average number of images. The thick blue line (near zero) represents the population average, and thinner multi-colored lines represent individual observers. Additionally (not shown) these same hypothetical scenarios with an experienced observer show similar trends over a much smaller range of probabilities.
Figure 2.6: Hypothetical best and worst case scenarios for observers from the general public committing false negative errors while participating in ecological photo-id. Under a worst-case scenario (on the left), variables not plotted on the x-axis were chosen to reflect poor match-finding conditions: largest catalog size (300 images), smallest number of matches (1), and longest time interval between matching images. Under a best-case scenario (on the right), variables not plotted on the x-axis were chosen to reflect optimum match-finding conditions: smallest catalog size (75), largest number of matches (8), and shortest time interval between matching images. In order to plot both these best and worst-case scenarios we used average age and average current observer experience values. We examined differences in the probability of making a false negative error as a result of changing A) time interval between matching images, B) number of matches in the catalog, C) size of the catalog viewed in a single sitting, and D) observer experience (# of previous images viewed). The thick black line in each plot indicates the population average; thinner multi-colored lines represent individual observers.
LITERATURE CITED


Bates, D., Maechler, M., and Bolker, B. (2011). lme4: Linear mixed-effects models using S4 classes. R package version 0.999375-41. http://CRAN.R-project.org/package=lme4


