Herpetological Conservation and Biology 19(1):117–128.
Submitted: 29 September 2023; Accepted: 20 March 2024; Published: 30 April 2024.

Development and Validation of a Scoring System for Abnormalities in the Gopher Frog (*Rana capito*)

Kiersten N. Nelson [1,2,4], ***Adam J. McFall*** [1,2], ***E. Tucker Stonecypher*** [1,2], ***Christian S. Swartzbaugh*** [1,2], ***Matthew C. Allender*** [3], ***and Stacey L. Lance*** [2]

[1] *Odum School of Ecology, University of Georgia, 140 East Green Street, Athens, Georgia 30602, USA*
[2] *Savannah River Ecology Laboratory, University of Georgia, Aiken, South Carolina 29802, USA*
[3] *Wildlife Epidemiology Lab, University of Illinois College of Veterinary Medicine, 2001 South Lincoln Avenue, Urbana, Illinois 61802, USA*
[4] *Corresponding author, e-mail: k.nelson@uga.edu*

Abstract.—**Headstarting efforts are thought to be critical in supplementing populations of the at-risk Gopher Frog (*Rana capito*); however, recent efforts have occasionally resulted in juveniles with developmental abnormalities. In response, we developed a scoring system to collect quantitative data on the presence and severity of these developmental abnormalities. Our objective was to describe and validate the abnormality scoring system so that it can be used by all Gopher Frog headstarting facilities. The scoring system covers five primary conditions encompassing commonly observed abnormalities. Two groups of participants with different levels of prior experience working with Gopher Frogs assigned scores to a set of images presented to them for each condition. We used intra-class correlation coefficients (ICC) to test the scoring system for inter- and intra-rater agreement as well as agreement with the benchmark standard (established by the authors). We found high ICC values for inter-rater agreement, intra-rater agreement, and agreement to the benchmark standard indicating either excellent or good reliability for all five conditions and for all raters when grouped together. These findings support the reliability and validity of the proposed developmental abnormality scoring system. Gopher Frog headstarting facilities can implement this scoring system to assist in tracking the frequency and severity of abnormalities observed in future headstarting efforts. We hope that by creating a reliable scoring system for Gopher Frogs, it can provide an overall framework and serve as a valuable resource to evaluate abnormalities across any amphibian species.**

Key Words.—abnormality; amphibian; conservation; headstarting; inter- and intra-rater agreement; scoring system

Introduction

The Gopher Frog (*Rana capito*) is a medium-sized terrestrial anuran that is native to the southeastern U.S. Coastal Plain (Conant and Collins 1991; Palis and Fischer 1997; Enge et al. 2014). Gopher Frogs are associated with the Longleaf Pine (*Pinus palustris*) ecosystem where adult frogs spend most of the year in upland habitat (Humphries and Sisson 2012) relying on vital refugia such as stump holes and Gopher Tortoise (*Gopherus polyphemus;* Franz 1986) or small mammal burrows (Lee 1968; Richter et al. 2001; Roznik et al. 2009). Gopher Frogs typically breed in isolated open-canopied ephemeral wetlands free of predatory fish (Bailey 1991; Enge et al. 2014). The Longleaf Pine ecosystem is considered critically endangered and has been reduced by more than 98% from its historic extent (Noss and Scott 1995), contributing to the decline of many associated wildlife species, including several threatened or endangered species (Noss and Scott 1995; Means 2006). As is the case with many amphibian and reptile species that rely on the Longleaf Pine ecosystem (Dodd 1995), the Gopher Frog has experienced steady population declines throughout its range that can be primarily attributed to habitat loss of suitable breeding ponds and terrestrial uplands (Jensen and Richter 2005). The Gopher Frog is a species of conservation concern throughout its range as they are currently listed as Vulnerable in Florida, Imperiled in Alabama, Georgia, and North Carolina, and Critically Imperiled in South Carolina and Tennessee (Nature Serve. 2023. NatureServe Network Biodiversity Location Data. NatureServe, Arlington, Virginia. Available from https://explorer.natureserve.org/ [Accessed 8 September 2023]).

In response to widespread Gopher Frog declines, headstarting programs have been implemented as a conservation strategy to assist with population augmentation and reintroduction efforts. Currently, headstarting efforts support many amphibian species of conservation concern across the globe, and the number of programs continues to increase (Harding et al. 2016). Headstarting involves collecting eggs from the wild, rearing the larvae in outdoor mesocosms, and subsequently releasing the newly metamorphosed frogs back into natural habitats. The goal of Gopher Frog headstarting efforts is to produce and contribute to viable self-sustaining populations in the wild. Headstarting is expected to facilitate population persistence by protecting the vulnerable larval stage and increasing the probability of survival to metamorphosis (Dodd 2005). The number of Gopher Frog headstarting programs has expanded over the decade. Currently, several institutions and federal agencies in Georgia, North Carolina, and South Carolina have established headstarting programs. These concerted conservation efforts are thought to be critical for long-term population persistence across the native range of the species. Yet, Gopher Frog headstarting efforts have occasionally resulted in frogs emerging with a suite of developmental abnormalities (McFall et al. 2023). The cause of these developmental abnormalities is currently undetermined and under investigation.

In 2021, we reared Gopher Frogs at the Savannah River Ecology Laboratory of the University of Georgia, USA, and 99% of them emerged with a wide range of developmental abnormalities (McFall et al. 2023). The most common developmental abnormalities observed included cutaneous hypopigmentation, microphthalmia, brachygnathia, edema, and exposed gill slits (McFall et al. 2023). After observing these abnormalities in 2021, we were able to confirm that researchers from multiple facilities involved in headstarting efforts have previously observed similar developmental abnormalities in Gopher Frogs (John Maerz and Dustin Smith, pers. comm.) along with two closely related species, Crawfish Frogs (*Rana areolatus*; Stiles et al. 2016) and Dusky Gopher Frogs (*Rana sevosus*; Joe Pechmann, pers. comm.). While developmental abnormalities have occasionally been observed in previous Gopher Frog headstarting events, we observed abnormalities that were more prevalent and severe than previous observations (McFall et al. 2023). Yet, researchers have not collected any quantitative data on the developmental abnormalities observed in previous headstarting events. Thus, we developed a scoring system to collect quantitative data on the prevalence and severity of the developmental abnormalities. Moving forward, it is critical to collect consistent quantitative data across headstarting facilities to monitor abnormalities and compare data. Ideally, we need a validated scoring system that can be used by individuals with and without experience in amphibian husbandry, as headstarting efforts involve teams of individuals with different backgrounds.

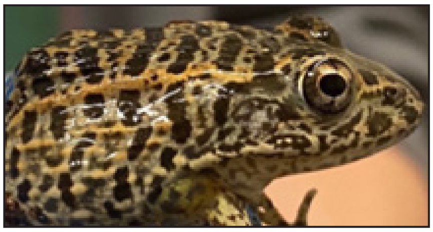
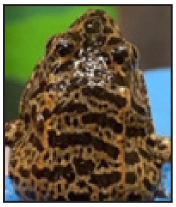
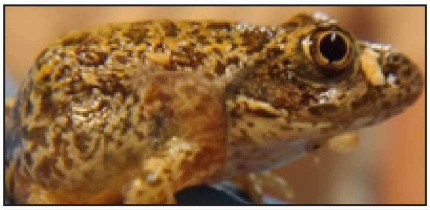
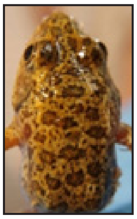
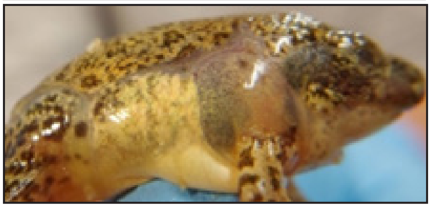
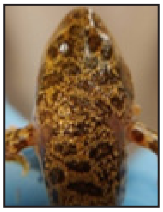
Scoring or rating systems have been used by a wide range of disciplines including human health and medicine, veterinary medicine or zoo animal husbandry, and ecological and conservation studies with a variety of applications. Scoring systems have been used to quantify specific animal health or welfare-relevant variables, such as body condition (Joblon et al. 2014; Jayson et al. 2018), disease (McGuirk and Peek 2014) or overall health monitoring (Wells et al. 2004; Ali et al. 2016), injuries or lesions (Hubert et al. 1999; Nielsen et al. 2010; Foddai et al. 2012; Chan and Karczmarski 2019; Lovallo et al. 2021), lameness (Garner et al. 2002; Flower and Weary 2006), alopecia of captive animals (Honess et al. 2005; Bechard et al. 2011), and pain assessments (Fitzpatrick et al. 2006; Evangelista et al. 2019; Evangelista and Steagall 2021). Scoring systems based primarily on visual assessments provide a non-destructive method to assess body condition, which can be particularly important for rare or endangered species where destructive methods or extensive handling times are undesirable or prohibited (e.g., Thomson et al. 2009; Clements and Sanchez 2015; Jayson et al. 2018). A scoring system that is practical, non-intrusive, repeatable, and inexpensive can serve as a valuable tool to monitor the health of an individual. Prior to widespread use, it is necessary to assure reliability (i.e., the overall reproducibility of the scores) and validity (i.e., the ability of the scale to measure what it is intended to measure) of the scoring system.

The purpose of our study was to describe and test the reliability and validity of our proposed Gopher Frog developmental abnormality scoring system. To do this, we assessed the agreement of the score definitions within and among two groups of participants that varied in expertise. While developed with Gopher Frogs in mind, our hope included creating a reliable scoring system that can provide a framework to quantitively evaluate developmental abnormalities across any amphibian species.

Materials and Methods

Gopher Frog headstarting.—We collected partial Gopher Frog egg masses on 5 March 2021 from

Table 1. Definitions, descriptions, and photographic examples for the finalized scoring model of cutaneous hypopigmentation for Gopher Frogs (*Rana capito*). Photographs were cropped and lightened as needed to enhance contrast. (Photographed by Christian Swartzbaugh).

Score	Description	Example Photographs: Lateral View	Dorsal View
0	No abnormalities present. Skin appears to have developed normally.		
1	Affected skin (clear/gray pigmentation) appears to only surround the front limb.		
2	Affected skin (clear/gray pigmentation) appears to surround the front limb, run down the lateral surface, and may continue onto the dorsal surface.		

a wetland in Barnwell County, South Carolina, USA. We reared tadpoles in outdoor mesocosms at the Savannah River Ecology Laboratory (SREL) of the University of Georgia until they reached metamorphosis as part of an experimental headstarting program. Once tadpoles began to metamorphose, we saw clear signs of developmental abnormalities. Initially, it was unclear if a pathogen may have been involved and thus, we implemented proper biosafety protocols and treated all individuals as if they had an unknown disease. Additionally, the South Carolina Department of Natural Resources determined that the frogs were unsuitable for release. To better understand the abnormalities and how they progressed, we transferred recently metamorphosed frogs to the SREL animal care facility and reared them in individual 0.47 L (16-ounce) Pro-Kal® deli containers (Fabri-Kal Corp., Kalamazoo, Michigan, USA) with a layer of moist Spagmoss (Besgrow Limited, Bishopdale, Churchchrist, New Zealand). Soon after (about 1–2 mo post tail resorption), we transferred frogs to 25.55 L (27-quart) latchable containers (Sterilite Corp., Townsend, Massachusetts, USA) with soil substrate and artificial burrows for several additional months to further understand the progression of the abnormalities and the effects on survival (McFall et al. 2023).

Scoring system.—To categorize and quantitatively evaluate the observed developmental abnormalities in the Gopher Frog metamorphs, we developed a scoring system to record presence/absence and severity of the different conditions. The scoring system covered five primary conditions: cutaneous hypopigmentation (Table 1), microphthalmia (Table 2), gill retention (Table 3), edema (Table 4), and brachygnathia (Table 5). Brachygnathia, edema, and gill retention conditions had bimodal outcomes (presence or absence), while cutaneous hypopigmentation and microphthalmia conditions included degrees of severity. We evaluated brachygnathia, edema, and gill retention conditions as bimodal outcomes as we did not observe a range in severity levels. We scored newly metamorphosed individuals upon tail resorption and re-scored periodically.

Photographic documentation.—Once we observed the first frog that metamorphosed with developmental abnormalities, we began scoring and photographing all subsequent emerging frogs. We used an LG G7 ThinQ (LG Electronics Inc., Seoul 150–721, South Korea) with a Xenvo Clarus 15× Macro Lens (Xenvo, Matawan, New Jersey, USA) attachment and took four photographs per frog: (1) left side; (2) right side; (3) dorsal view; and (4)

Table 2. Definitions, descriptions, and photographic examples for the finalized scoring model of microphthalmia for Gopher Frogs (*Rana capito*). Photographs were cropped and lightened as needed to enhance contrast. (Photographed by Christian Swartzbaugh).

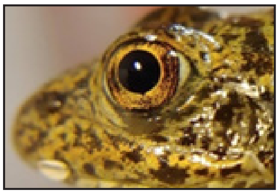
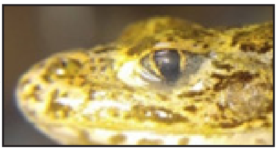
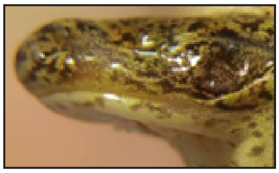
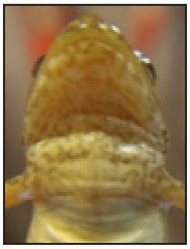
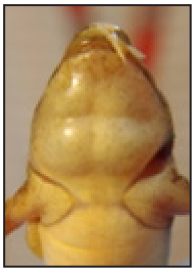
Score	Description	Example Photographs
0	No abnormalities present. Eye appears to have developed normally and is fully emerged from the skull.	
1	Eye partially emerged from skull. Eye may be concealed by a thin layer of skin but is still partially visible.	
2	Eye is not emerged from the skull. Eye appears to be completely absent/ not visible.	

Table 3. Definitions, descriptions, and photographic examples for the finalized scoring model of gill retention for Gopher Frogs (*Rana capito*). Photographs were cropped and lightened as needed to enhance contrast. (Photographed by Christian Swartzbaugh).

Score	Description	Example Photographs
0	No abnormalities present. No open slits/wounds and no visible external gills are present directly above the front limbs.	
1	Open slits/wounds and/or visible external gill present directly above the front limbs. A visible slit/gill may be present on one or both sides.	

ventral view. To capture all four photographs, one person properly handled and repositioned the frog while another person took each desired photograph.

***Participant recruitment*.**—We sought participants from two categories defined by their level of expertise in rearing amphibians. The first category, specialists, consisted of professionals who were directly involved in Gopher Frog research and/or headstarting (including rearing tadpoles and/or releasing metamorphs). We excluded personnel directly involved in this project, including the authors, as participants in the evaluation. The second category, graduate students, consisted of graduate students in the life sciences at the University of Georgia who had no previous experience with amphibian rearing and/or research. We chose these participant groups to determine if training or previous experience working with Gopher Frogs is required to use this scoring system because employees and interns at headstarting facilities often have a background in the life sciences, but no specific experience rearing amphibians. We recruited participants by sending emails to 25 specialists and 46 graduate students (mailing list of current graduate students affiliated with SREL). Anyone who expressed interest in participating in the study was sent a website link to a Google Drive (Google Corporation, Mountain View, California, USA) folder containing the evaluation materials. In total, seven specialists and 11 graduate students completed the evaluation.

***Scoring evaluation*.**—We created a presentation in Microsoft PowerPoint (Microsoft Corporation, Redmond, Washington, USA) containing a series of photographs of the conditions. The presentation was broken down into five sections, one for each condition (see Tables 1–5). For each condition, we selected 20 photographs (15 photographs were unique and five were randomly selected repeats) for each defined score. We used the 15 unique photographs to evaluate the inter-rater agreement (i.e., the level of agreement between evaluators for the score of each photograph) and we used the five repeat photographs to evaluate the intra-rater agreement (i.e., the level of agreement amongst the same evaluator scoring the exact same photograph). Before distributing the presentation, we used a random number generator to randomize the order of the conditions and photographs within each condition. We edited each photograph in the presentation as needed by cropping and enhancing the brightness to focus on the corresponding condition. We gave each participant the same presentation.

Table 4. Definitions, descriptions, and photographic examples for the finalized scoring model of edema for Gopher Frogs (*Rana capito*). Photographs were cropped and lightened as needed to enhance contrast. (Photographed by Christian Swartzbaugh).

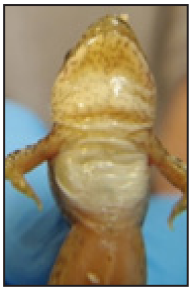
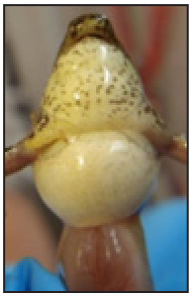
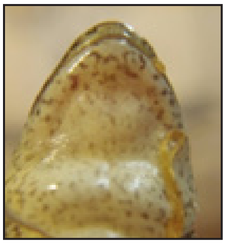
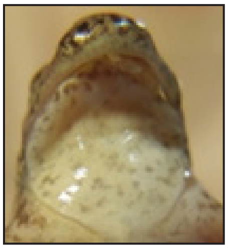
Score	Description	Example Photographs
0	No abnormalities present. No bloating in any portion of the body and/ or extremities.	
1	Appears bloated in any portion of the body and/ or extremities.	

Table 5. Definitions, descriptions, and photographic examples for the finalized scoring model of brachygnathia for Gopher Frogs (*Rana capito*). Photographs were cropped and lightened as needed to enhance contrast. (Photographed by Christian Swartzbaugh).

Score	Description	Example Photographs
0	No abnormalities present. Jaw appears to have developed normally.	
1	Reduced lower mandible causing a noticeable gap between the upper and lower jaw.	

After agreeing to participate, an evaluator received an instruction document. We asked participants to read through the instruction document and contact the primary author in case of questions. The instruction document included a description of the scoring system, example photographs of individual Gopher Frogs with and without developmental abnormalities, and instructions for how to complete the evaluation. Once evaluators read through the instructions, we asked them to close the instruction document and begin the evaluation. Per the instructions, we requested that the evaluators maximize the brightness of their computer screen to reduce variation between viewing the photographs on different computers. Each of the evaluators assigned a score to each photograph in the presentation and submitted the scores via a Google Form (Google Corporation, Mountain View, California, USA) that was provided. We requested that the evaluators not discuss their results with any other participant.

Statistical analysis.—To assess the validity and reliability of our scoring system, we evaluated agreement in three ways: (1) inter-rater agreement (i.e., agreement of the score assigned to the same photograph between different raters); (2) intra-rater agreement (i.e., agreement of the score assigned to the same photograph on two separate occasions by the same rater); and (3) agreement to the benchmark standard (i.e., agreement between the score of the photograph assigned by the raters and the score assigned by the authors that were considered the benchmark standard). We used intra-class correlation coefficients (ICC; Shrout and Fleiss 1979; McGraw and Wong 1996) to assess inter- and intra-rater agreement, as well as agreement to the benchmark standard for all conditions and within groups (i.e., specialist or graduate students) and across all participants. We assessed inter-rater agreement for each of the conditions described in the scoring system using a Two-way Random Effects ICC for absolute agreement. For intra-rater agreement, we used a Two-way Mixed Effects ICC for absolute agreement. For agreement to the benchmark standard, we used a Two-way Mixed Effects ICC for absolute agreement. In addition, we calculated the proportion of the matched scores between the raters and the benchmark standard across all raters for each condition. The ICC estimates can be reported as a single measure or the average of k measures, where ICC single represents the score reliability of a single rater and ICC average represents the mean value of multiple raters (Koo and Li 2016). In general, ICC average values tend to be higher than the more conservative ICC single values (Hallgren 2012; Koo

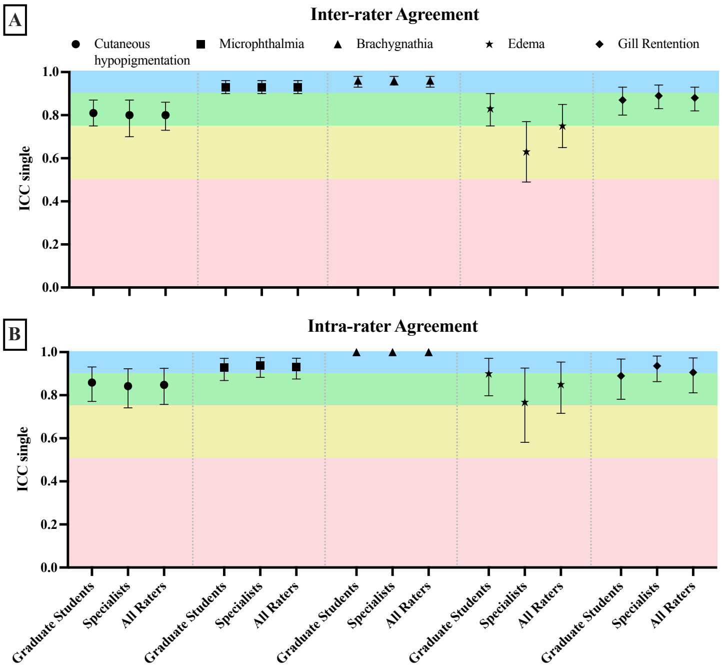


Figure 1. (A) Inter-rater agreement and (B) intra-rater agreement for each condition described in the Gopher Frog (*Rana capito*) developmental abnormalities scoring system. Intra-class correlation coefficient (ICC) was used to assess the inter- and intra-rater agreement using a two-way random effects ICC with absolute agreement (people rating: n = 11 for graduate student group, n = 7 for specialist group, n = 18 for all people [i.e., all raters]). Estimates for ICC single measures and the 95% confidence intervals are presented. Interpretation of rating agreement was as follows: ICC < 0.5 = poor (pink), 0.5–0.75 = moderate (yellow), 0.75–0.9 = good (green), and > 0.90 = excellent (blue).

and Li 2016). Both ICC single and average indicate scale performance but depend upon the desired application; thus, both estimates can be found in the Supplemental Information file (Tables S1-S3) along with their 95% confidence intervals (95% CI). In the results of our study, we focus on presenting the ICC single measures as it is a more accurate reflection of how the scoring system will typically be used (single rater, assigning a single score). We based our interpretation of agreement of ICC single measures on previously published guidelines as follows: (1) ICC < 0.5 = poor; (2) 0.5–0.75 = moderate; (3) 0.75–0.9 = good; and (4) > 0.90 = excellent (Koo and Li 2016). We followed previously published methods and guidelines to select the appropriate ICC form and reporting of ICC parameters for each test (Shrout and Fleiss 1979; McGraw and Wong 1996; Koo and Li 2016). ICCs are a commonly used statistic for assessing rater reliability for ordinal, interval, and ratio variables (Hallgren 2012); however, there is an underlying assumption of normality. Thus, we tested for violations of normality by visually examining histograms displaying the distribution of these data and determined that the normality assumption was not violated as we did not observe any extreme skewness. Additional statistical tests for normality were not appropriate because these data are ordinal and often bivariant. We performed all statistical analyses using R open software (R Development Core Team 2022) using the irr package (Gamer et al. 2012).

Results

Inter-rater agreement was high (i.e., at or above an ICC good level of agreement) across all conditions and levels of expertise (Fig. 1; Supplemental Information Table S1). Similarly, intra-rater agreement was high across all conditions and levels of expertise (Fig. 1; Supplemental Information Table S2). The inter- and intra-rater level of agreement was similar between the two levels of expertise across four of the five conditions (cutaneous hypopigmentation, microphthalmia, brachygnathia, and gill retention). When evaluating edema, the graduate students had a slightly higher level of both inter- and intra-rater agreement compared to the specialists. There was a high level of agreement between the score assigned by the raters compared to the score determined by the benchmark standard across all conditions and all raters (Supplemental Information Table S3). The

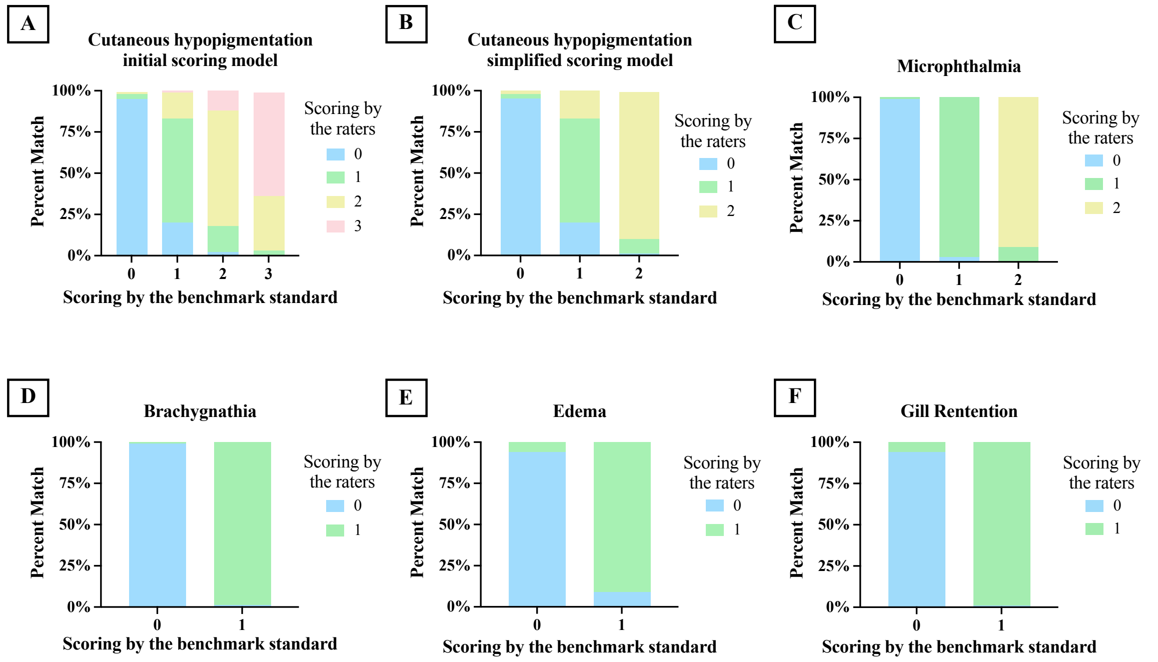


Figure 2. Accuracy of people (i.e., raters) assigning the abnormality scores for each condition described in the Gopher Frog (*Rana capito*) developmental abnormalities scoring system. Accuracy based on the percentage of scores assigned by people (i.e., raters) that matched the benchmark standard for each of the conditions: (A) cutaneous hypopigmentation for the initial scoring model (i.e., the originally proposed scoring model including the score categories of 0, 1, 2, 3), (B) cutaneous hypopigmentation for the simplified (i.e., finalized) scoring model (combining score categories of 2 and 3), (C) microphthalmia, (D) brachygnathia, (E) edema, and (F) gill retention.

percentage of occasions when the raters assigned the same score as the benchmark standard was relatively high (> 90%) across all scores for four out of the five conditions (microphthalmia, brachygnathia, edema, and gill retention; Fig. 2; Supplemental Information Table S4). For cutaneous hypopigmentation, the percentage of occasions when the raters assigned the same score as the benchmark standard was much more variable (percentage matched < 75%), particularly for scores 2 and 3 (Fig. 2). Consequently, we re-analyzed these data with the scores of 2 and 3 combined into a single category. With this simplified scoring system for cutaneous hypopigmentation with only categories 0, 1, and 2, the percentage of occasions when the raters assigned the same score as the benchmark standard increased to 89% (Fig. 2).

Discussion

Headstarting is a vital conservation tool for Gopher Frogs, particularly as we enhance our understanding of long-term population persistence and sensitivity to the complex factors contributing to declines across the range of the species, such as drought (Crawford et al. 2022) and anthropogenic stressors (Paulukonis et al. 2021). Conservation planning efforts are and will continue to be a priority for many species, as 41% of amphibian species globally are considered threatened with extinction according to the International Union for Conservation of Nature (IUCN) Red List of Threatened Species (IUCN 2022). The Amphibian Conservation Action Plan (ACAP) was developed by the IUCN Species Survival Commission (IUCN/SSC) Amphibian Specialist Group as a response to global amphibian declines and specifically highlights the immediate need to establish long-term conservation programs (Gascon et al. 2007). Despite the importance of conservation programs to many amphibian species, there are significant knowledge gaps related to amphibian husbandry, health, and basic nutritional requirements that complicate the success of these programs (Ferrie et al. 2014; Olea-Popelka et al. 2014). In addition, to successfully tackle the complex factors contributing to global amphibian declines, a broad set of stakeholders must be involved in amphibian conservation efforts (Gascon et al. 2007). Yet, efforts involving a range of stakeholders have resulted in different technical approaches resulting in a general lack of standardization in techniques, tools, measurement units, and protocols (Ferrie et al. 2014). This lack of standardization represents a significant challenge when attempting to compare and evaluate parameters related to amphibian health for species involved in conservation programs.

Previous efforts have proposed potential standardized approaches and methods for several aspects related to amphibian health (Olea-Popelka et al. 2014). The developmental abnormalities observed in Gopher Frog headstarting efforts exemplify many of these complications and challenges. Currently, Gopher Frog headstarting efforts are occurring in Georgia, North Carolina, and South Carolina and involve multiple stakeholders and agencies. Personnel involved in Gopher Frog headstarting efforts have observed developmental abnormalities in captive-reared animals since at least 2017 and anecdotal evidence suggests that the frequency has since increased. Up until this point, no quantitative data have been collected on the developmental abnormalities observed in headstarting events. Previous efforts to evaluate abnormality severity in wild amphibians used a scoring system that involved calculating the sum of the number of abnormalities on each abnormal individual (Johnson et al. 2001). Using this scoring model results in a single value to represent the overall abnormality severity for the individual but does not provide detailed information about the severity of multiple abnormal conditions. Currently, there is a lack of understanding concerning the cause of the abnormalities occurring in headstarting and whether they occur in the wild. Thus, to improve efforts to track the potential impacts of environmental conditions on the abnormalities we observed, we wanted a scoring model that assessed the severity of each abnormality.

By creating and implementing our own scoring system, we were able to collect quantitative data to compare the severity of abnormalities across rearing tanks and document the progression of conditions (McFall et al. 2023). This scoring system has not been previously used by other facilities; thus, we were unable to compare our data to those of other facilities that also experienced abnormalities. Based on the results of this study, we now have a reliable and valid standardized scoring system that can be widely used across headstarting facilities. This will aid in our ability to track the presence and severity of abnormalities across headstarting facilities and throughout time. Widespread use of this scoring system will provide researchers and conservation practitioners with consistent data and information that will aid further investigations into the potential causes of the abnormalities. Additionally, collecting quantitative data across Gopher Frog headstarting facilities will provide valuable insights into the potential role of captive rearing environments in contributing to the observed developmental abnormalities.

We found clear support that our Gopher Frog developmental abnormality scoring system can be used by individuals with no prior experience in amphibian husbandry. Overall, inter-rater agreement was either excellent or good for four of the five conditions. Evaluators were less consistent in judging whether or not frogs had edema. When evaluating edema, the graduate students had a higher ICC single value indicating an excellent level of agreement compared to the specialist who had a good level of agreement. This is encouraging because often seasonal technicians assisting with headstarting efforts have less experience than specialists in the field. Edema is difficult to evaluate with a photograph compared to an in-person visual examination, which may explain this differing level of agreement between the graduate students and specialists. Importantly, intra-rater agreement was excellent or good across all conditions. This suggests that the definitions proposed in the described scoring system are clear and repeatable on an individual level. Thus, implementation of this scoring system across headstarting facilities should provide consistent scoring within and across facilities.

The inter- and intra-rater agreements assess how consistent the scores are within and between raters, but they do not reflect accuracy (i.e., the scores could be reliable between raters and among the same rater but consistently wrong and not aligned with the defined scoring system). By comparing the agreement of the scores from each group with the scores assigned by the authors that were considered the benchmark standard, we were able to determine the accuracy of the evaluators. Across most conditions, the raters correctly scored the abnormality more than 90% of the time. The exception was for cutaneous hypopigmentation, where the raters were correct 70% of the time. Initially, cutaneous hypopigmentation had the most complex scoring system with four possible categories (described in Supplementary Information Table S5). When we evaluated the inter- and intra-rater agreement for cutaneous hypopigmentation, we examined agreement with this original scoring model with the four possible categories (scores 0, 1, 2, 3). While we found a high level of inter- and intra-rater agreement for cutaneous hypopigmentation, this does not reflect accuracy. When we compared the agreement of the scores from the evaluators to the benchmark standard for cutaneous hypopigmentation, raters were not able to reliably distinguish between a score of 2 and 3. When we re-analyzed the agreement between the scores assigned by evaluators and the benchmark standard with only three categories (scores 0, 1, 2), accuracy improved. Thus, moving forward we have combined categories 2 and 3 for a finalized validated scoring system (described in Table 1).

The present study has a few limitations that should be taken into account. First, the scoring system was developed by evaluating live specimens while this study evaluated the scoring system based solely on image analysis. It is possible that an in-person examination could lead to a different level of agreement between evaluators. Though, it is unlikely that we would have been able to represent the entire range of the abnormalities in adequate numbers at a single time for an in-person evaluation. For the same rationale, other similar studies examining rater agreement in scoring systems for injuries (Mejdell et al. 2010), lesions (Foddai et al. 2012), or disease symptoms (Webb et al. 2020) in different species have also used images or videos rather than live specimen in-person evaluations. Additionally, when the scoring system was developed the condition scores included a description of the location of the abnormality (i.e., left, right, or bilateral). The abnormalities were not always symmetrical (e.g., microphthalmia may have received a score of 0 on the left eye and a score of 2 on the right eye). To simplify this study, we chose not to include this component in the evaluation of the reliability of the scoring system. Though, we recommend that the finalized scoring system include an additional classification for the location and symmetry of the abnormality.

In conclusion, this study provides support for the validation of the Gopher Frog developmental abnormality scoring system with high score agreement across conditions and evaluators. We recommend that facilities involved in headstarting Gopher Frogs and other amphibians use this developmental abnormality scoring system to track the frequency and severity of the abnormalities observed in future headstarting efforts. Wide use of the developmental abnormality scoring system will be a valuable tool for collecting standardized data and will provide researchers with beneficial information to further investigate potential causes of the abnormalities. Additionally, we hope that this will provide a framework for future management or conservation efforts for other imperiled species that may encounter the need to monitor similar health and welfare characteristics.

Acknowledgements.—The authors thank Padraic Robinson and Ezmie Trevarrow for providing feedback on the evaluation, University of Georgia graduate students (Carolyn Cummins, Carson Pakula, Chris Leaphart, Chris Smaga, Colin Richter, Corinne Sweeney, Emily Martin, Jennifer McDaniel, Jordan Youngmann, Kelly Holland, Samantha Bock, Shane Guenin, and Susanna Mann) and Gopher Frog headstarting facility personnel (James Henne, Tobias Landberg, Casey Perkins, Carol Price, Ryan McAlarney, Chris Shupp, Kabryn Mattison, and Jeff Hall) for their participation in the evaluation process. This material is based upon work supported by the U.S. Department of Energy Office of Environmental Management under Award Number DE-EM0005228 to the University of Georgia Research Foundation, and by the U.S. Fish and Wildlife Service (F21AC02583). We collected animals under South Carolina Department of Natural Resources permit SC-43–2021 and all procedures followed Institutional Animal Care and Use Committee guidelines approved by the University of Georgia (AUPs A2020 12–010-Y2-A2 and A2021 06–014-Y2-A3).

Disclaimer: This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government.

Literature Cited

Ali, A.B.A., M.A. El Sayed, M.Y. Matoock, M.A. Fouad, and C.R. Heleski. 2016. A welfare assessment scoring system for working equids - a method for identifying at risk populations and for monitoring progress of welfare enhancement strategies (trialed in Egypt). Applied Animal Behaviour Science 176:52–62.

Bailey, M.A. 1991. The Dusky Gopher Frog in Alabama. Journal of the Alabama Academy of Science 62:28–34.

Bechard, A., R. Meagher, and G. Mason. 2011. Environmental enrichment reduces the likelihood of alopecia in adult C57BL/6J mice. Journal of the American Association for Laboratory Animal Science 50:171–174.

Chan, S.C.Y., and L. Karczmarski. 2019. Epidermal lesions and injuries of coastal dolphins as indicators of ecological health. EcoHealth 16:576–582.

Clements, J., and J.N. Sanchez. 2015. Creation and validation of a novel body condition scoring method for the Magellanic Penguin (*Spheniscus magellanicus*) in the zoo setting. Zoo Biology 34:538–546.

Conant, R., and J.T. Collins. 1991. A Field Guide to Reptiles and Amphibians, Eastern and Central North America. Houghton Mifflin Company, New York, New York, USA.

Crawford, B.A., J.C. Maerz, and V.C.K. Terrell. 2022. Population viability analysis for a pond-breeding amphibian under future drought scenarios in the southeastern United States. Global Ecology and Conservation 36:e02119. https://doi.org/10.1016/j.gecco.2022.e02119.

Dodd, C.K., Jr. 1995. Reptiles and amphibians in the endangered Longleaf Pine ecosystem. Pp. 129–131. *In* Our Living Resources: A Report to the Nation on the Distribution, Abundance, and Health of U.S. Plants, Animals, and Ecosystems. LaRoe, E.T., G.S. Farris, C.E. Puckett, P.D. Doran, and M.J. Mac (Eds.). U.S. Department of the Interior, National Biological Service, Washington DC, USA.

Dodd, C.K., Jr. 2005. Amphibian conservation and population manipulation. Pp. 265–270. *In* Status and Conservation of U.S. Amphibians. Lanoo, M.J. (Ed.). University of California Press, Berkeley, California, USA.

Enge, K.M., A.L. Farmer, J.D. Mays, T.D. Castellón, E.P. Hill, and P.E. Moler. 2014. Survey for winter-breeding amphibian species. Final Report 9242216399, Florida Fish and Wildlife Conservation Commission, Gainsville, Florida, USA.

Evangelista, M.C., and P.V. Steagall. 2021. Agreement and reliability of the Feline Grimace Scale among cat owners, veterinarians, veterinary students and nurses. Scientific Reports 11:5262. https://doi.org/10.1038/s41598-021-84696-7.

Evangelista, M.C., R. Watanabe, V.S.Y. Leung, B.P. Monteiro, E. O'Toole, D.S.J. Pang, and P.V. Steagall. 2019. Facial expressions of pain in cats: the development and validation of a Feline Grimace Scale. Scientific Reports 9:19128. https://doi.org/10.1038/s41598-019-55693-8.

Ferrie, G.M., V.C. Alford, J. Atkinson, E. Baitchman, D. Barber, W.S. Blaner, G. Crawshaw, A. Daneault, E. Dierenfeld, M. Finke, et al. 2014. Nutrition and health in amphibian husbandry. Zoo Biology 33:485–501.

Fitzpatrick, J., M. Scott, and A. Nolan. 2006. Assessment of pain and welfare in sheep. Small Ruminant Research 62:55–61.

Flower, F.C., and D.M. Weary. 2006. Effect of hoof pathologies on subjective assessments of dairy cow gait. Journal of Dairy Science 89:139–146.

Foddai, A., L.E. Green, S.A. Mason, and J. Kaler. 2012. Evaluating observer agreement of scoring systems for foot integrity and footrot lesions in sheep. BMC Veterinary Research 8:65. https://doi.org/10.1186/1746-6148-8-65.

Franz, R. 1986. *Gopherus polyphemus* (Gopher Tortiose). Burrow commensals. Herpetological Review 17:64

Gamer M., J. Lemon, I. Fellows, and P. Singh. 2012. Irr: various coefficients of interrater reliability and agreement. R package version 0.84.1. https://cran.r-project.org/packages=irr.

Garner, J.P., C. Falcone, P. Wakenell, M. Martin, and J.A. Mench. 2002. Reliability and validity of a modified gait scoring system and its use in assessing tibial dyschondroplasia in broilers. British Poultry Science 43:355–363.

Gascon, C., J.P. Collins, R.D. Moore, D.R. Church, J.E. McKay, and J.R. Mendelson, III (Eds.). 2007. Amphibian Conservation Action Plan Proceedings: IUCN/SSC Amphibian Conservation Summit 2005. International Union for Conservation of Nature, Species Survival Commission, Amphibian Specialist Group, Gland, Switzerland.

Hallgren, K.A. 2012. Computing inter-rater reliability for observational data: an overview and tutorial. Tutorials in Quantitative Methods for Psychology 8:23–34.

Harding, G., R.A. Griffiths, and L. Pavajeau. 2016. Developments in amphibian captive breeding and reintroduction programs. Conservation Biology 30:340–349.

Honess, P., J. Gimpel, S. Wolfensohn, and G. Mason. 2005. Alopecia scoring: the quantitative assessment of hair loss in captive macaques. Alternatives to Laboratory Animals 33:193–206.

Hubert, G.F., G.K. Wollenberg, L.L. Hungerford, and R.D. Bluett. 1999. Evaluation of injuries to Virginia Opossums captured in the EGGTM trap. Wildlife Society Bulletin 27:301–305.

Humphries, J.W., and M.A. Sisson. 2012. Long distance migrations, landscape use, and vulnerability to prescribed fire of the Gopher Frog (*Lithobates capito*). Journal of Herpetology 46:665–670.

International Union for the Conservation of Nature (IUCN). 2022. The IUCN Red List of Threatened Species. Version 2022-2. https://www.iucnredlist.org.

Jayson, S., L. Harding, C.J. Michaels, B. Tapley, J. Hedley, M. Goetz, A. Barbon, G. Garcia, J. Lopez, and E. Flach. 2018. Development of a body condition score for the Mountain Chicken Frog (*Leptodactylus fallax*). Zoo Biology 37:196–205.

Jensen, J.B., and S.C. Richter. 2005. Gopher frogs, *Rana capito*. Pp. 536–538 *In* Amphibian Declines: The Conservation Status of United States Species. Lannoo, M (Ed.). Univeristy of California Press, Berkley, California, USA.

Joblon, M.J., M.A. Pokras, B. Morse, C.T. Harry, K.S. Rose, S.M. Sharp, M.E. Niemeyer, K.M. Patchett, W.B. Sharp, and M.J. Moore. 2014. Body condition scoring system for delphinids based on Short-beaked Common Dolphins (*Delphinus delphis*). Journal of Marine Animals and Their Ecology 7:5–13.

Johnson, P.T.J., K.B. Lunde, E.G. Ritchie, J.K. Reaser, and A.E. Launer. 2001. Morphological abnormality patterns in a California amphibian community. Herpetologica 57:336–352.

Koo, T.K., and M.Y. Li. 2016. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. Journal of Chiropractic Medicine 15:155–163.

Lee, D.S. 1968. Herpetofauna associated with central Florida mammals. Herpetologica 24:83–84.

Lovallo, M.J., H.B. White, J.D. Erb, M.S. Peek, and T.J. Deliberto. 2021. Welfare performance of three foothold traps for capturing North American River Otters. Journal of Fish and Wildlife Management 12:513–519.

McFall, A.J., K.N. Nelson, E.T. Stonecypher, C.S. Swartzbaugh, M.C. Allender, C.E. Burrell, M.J. Yabsley, and S.L. Lance. 2023. Morphological abnormalities in the Gopher Frog (*Lithobates capito*) during a headstarting event. Herpetological Conservation and Biology 18:436–449.

McGraw, K.O., and S.P. Wong. 1996. Forming inferences about some intraclass correlation coefficients. Psychological Methods 1:30–46.

McGuirk, S.M., and S.F. Peek. 2014. Timely diagnosis of dairy calf respiratory disease using a standardized scoring system. Animal Health Research Reviews 15:145–147.

Meagher, R.K. 2009. Observer ratings: validity and value as a tool for animal welfare research. Applied Animal Behaviour Science 119:1–14.

Means, D.B. 2006. Vertebrate Faunal Diversity of Longleaf Pine Ecosystems. Pp. 157–213 *In* The Longleaf Pine Ecosystem: Ecology, Silviculture, and Restoration. Jose, S., E.J. Jokela, and D.L. Miller (Eds.). Springer New York, New York, USA.

Mejdell, C.M., G.H.M. Jørgensen, T. Rehn, K. Fremstad, L. Keeling, and K.E. Bøe. 2010. Reliability of an injury scoring system for horses. Acta Veterinaria Scandinavica 52:68. https://doi.org/10.1186/1751-0147-52-68.

Nielsen, A.M.W., S.S. Nielsen, C.E. King, and M.F. Bertelsen. 2010. Classification and prevalence of foot lesions in captive flamingos (Phoenicopteridae). Journal of Zoo and Wildlife Medicine 41:44–49.

Noss, R.F., and J.M. Scott. 1995. Endangered Ecosystems of the United States: A Preliminary Assessment of Loss and Degradation. U.S. Department of the Interior, National Biological Service, Washington DC, USA.

Olea-Popelka, F., G.M. Ferrie, C. Morris, A.P. Pessier, K. Schad, M.A. Stamper, R. Gagliardo, E. Koutsos, and E.V. Valdes. 2014. Leaping forward in amphibian health and nutrition. Zoo Biology 33:586–591.

Palis, J.G., and R.A. Fischer. 1997. Species profile: Gopher Frog (*Rana capito* spp.) on military installations in the southeastern United States. Technical Report SERDP-97-5, U.S. Army Engineers Waterways Experimental Station, Vicksburg, Mississippi, USA.

Paulukonis, E.A., B.A. Crawford, J.C. Maerz, S.J. Wenger, and N.P. Nibbelink. 2021. Prioritization of vulnerable species under scenarios of anthropogenic-driven change in Georgia's coastal plain. Journal of Fish and Wildlife Management 12:273–293.

R Development Core Team. R: A language and environment for statistical computing. R

Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Richter, S.C., J.E. Young, R.A. Seigel, and G. N. Johnson. 2001. Postbreeding movements of the Dark Gopher Frog, *Rana sevosa* Goin and Netting: implications for conservation and management. Journal of Herpetology 35:316–321.

Roznik, E.A., S.A. Johnson, C.H. Greenberg, and G.W. Tanner. 2009. Terrestrial movements and habitat use of Gopher Frogs in Longleaf Pine forests: a comparative study of juveniles and adults. Forest Ecology and Management 259:187–194.

Shrout, P.E., and J.L. Fleiss. 1979. Intraclass correlations: uses in assessing rater reliability. Psychological Bulletin 86:420–428.

Stiles, R.M., M.J. Sieggreen, R.A. Johnson, K. Pratt, M. Vassallo, M. Andrus, M. Perry, J.W. Swan, and M. J. Lannoo. 2016. Captive-rearing state endangered Crawfish Frogs *Lithobates areolatus* from Indiana, USA. Conservation Evidence 13:7–11.

Thomson, J.A., D. Burkholder, M.R. Heithaus, and L.M. Dill. 2009. Validation of a rapid visual-assessment technique for categorizing the body condition of Green Turtles (*Chelonia mydas*) in the field. Copeia 2009:251–255.

Webb, J.K., K.A. Keller, K. Welle, and M.C. Allender. 2020. Evaluation of the inter- and intraindividual agreement of pododermatitis scoring model in Greater Flamingos (*Phoenicopterus roseus*). Journal of Zoo and Wildlife Medicine 51:379–384.

Wells, R.S., H.L. Rhinehart, L.J. Hansen, J.C. Sweeney, F.I. Townsend, R. Stone, D.R. Casper, M.D. Scott, A.A. Hohn, and T.K. Rowles. 2004. Bottlenose Dolphins as marine ecosystem sentinels: developing a health monitoring system. EcoHealth 1:246–254.

Supplemental Information: http://www.herpconbio.org/Volume_19/Issue_1/Nelson_etal_2024_Suppl.pdf

Kiersten N. Nelson (left) is a Ph.D. student in the University of Georgia (UGA) Odum School of Ecology in Athens, Georgia, USA. She is primarily located at the Savannah River Ecology Laboratory (SREL) where she is conducting her dissertation research focusing on Gopher Frog conservation. She received a B.S. in Ecology, Evolution, and Environmental Biology at Purdue University in West Lafayette, Indiana, USA. Her research interests are herpetology, conservation, habitat management, ecosystem ecology, and disease ecology. **Adam J. Mcfall** (middle) is a Biologist with the U.S. Geological Survey at the Columbia Environmental Research Center in Columbia, Missouri, USA. He received his B.S. in Biology in 2019 from the University of South Carolina Aiken, USA, and his M.S. in Integrative Conservation and Sustainability in 2023 from the UGA Odum School of Ecology, Athens, USA. His research interests are amphibian conservation, behavioral ecology, and habitat restoration. **E. Tucker Stonecypher** (second from left) has a Master's degree in Integrative Conservation and Sustainability from the UGA Odum School of Ecology and is a Research Professional at the SREL. His research interests are wetland restoration, wetland plant communities, disturbance ecology, and amphibian conservation. **Christian Swartzbaugh** (second from right) is a doctoral student in the UGA Odum School of Ecology. He studies dynamics of freshwater fish communities at the SREL. His research interests include behavior, biodiversity, and community ecology of fishes and aquatic environments, as well as community response to biological stress. **Stacey Lance** (right) is a Senior Research Scientist at the SREL. She received her B.S. in Biological Sciences at the University of Connecticut, Storrs, USA, and her Ph.D. in Zoology at the University of Maryland, College Park, USA. Her research in the Lance Lab at SREL focuses on the conservation and management of freshwater vertebrates, the impact of global change on isolated wetlands and pond-breeding amphibians, and aquatic pollution and evolutionary toxicology. (Photographed by Sophia Zaslow).

Matthew C. Allender is a wildlife and zoo Veterinarian at the Brookfield Zoo of the Chicago Zoological Society, Illinois, USA, and is the Director of the Wildlife Epidemiology Lab at the University of Illinois, Urbana, USA. He received his B.S. in Ecology, Ethology, and Evolution, D.V.M., M.S., and Ph.D. from the University of Illinois, Urbana, USA. His research focuses on health and pathogen investigations of free-ranging and managed populations of wildlife. (Photograph courtesy of the Chicago Zoological Society/Brookfield Zoo).