
The Consistency of Extinction Risk Classification Protocols

TRACEY J. REGAN,* MARK A. BURGMAN,† MICHAEL A. MCCARTHY,‡ LAWRENCE L. MASTER,§
DAVID A. KEITH,** GEORGINA M. MACE,†† AND SANDY J. ANDELMAN‡‡

*The Ecology Centre, School of Life Sciences, The University of Queensland, Brisbane QLD 4072, Australia,
email t.regan@uq.edu.au

†School of Botany, The University of Melbourne, Parkville, Victoria 3010, Australia

‡Australian Research Centre for Urban Ecology, Royal Botanic Gardens Melbourne, School of Botany, University of Melbourne,
Victoria, 3010, Australia

§NatureServe, 11 Avenue de Lafayette, 5th Floor, Boston, MA 02111, U.S.A.

**New South Wales National Parks and Wildlife Service, Hurstville, New South Wales 2220, Australia

††Institute of Zoology, Zoological Society of London, Regents Park, London NW1 4RY, United Kingdom

‡‡National Center for Ecological Analysis and Synthesis, University of California, Santa Barbara, CA 93101, U.S.A

Abstract: *Systematic protocols that use decision rules or scores are seen to improve consistency and transparency in classifying the conservation status of species. When applying these protocols, assessors are typically required to decide on estimates for attributes that are inherently uncertain. Input data and resulting classifications are usually treated as though they are exact and hence without operator error. We investigated the impact of data interpretation on the consistency of protocols of extinction risk classifications and diagnosed causes of discrepancies when they occurred. We tested three widely used systematic classification protocols employed by the World Conservation Union, NatureServe, and the Florida Fish and Wildlife Conservation Commission. We provided 18 assessors with identical information for 13 different species to infer estimates for each of the required parameters for the three protocols. The threat classification of several of the species varied from low risk to high risk, depending on who did the assessment. This occurred across the three protocols investigated. Assessors tended to agree on their placement of species in the highest (50–70%) and lowest risk categories (20–40%), but there was poor agreement on which species should be placed in the intermediate categories. Furthermore, the correspondence between the three classification methods was unpredictable, with large variation among assessors. These results highlight the importance of peer review and consensus among multiple assessors in species classifications and the need to be cautious with assessments carried out by a single assessor. Greater consistency among assessors requires wide use of training manuals and formal methods for estimating parameters that allow uncertainties to be represented, carried through chains of calculations, and reported transparently.*

Key Words: conservation status, classification protocols, threatened species lists, uncertainty, operator error, IUCN Red List, NatureServe

La Consistencia de los Protocolos de Clasificación del Riesgo de Extinción

Resumen: *Los protocolos sistemáticos que utilizan reglas o puntuaciones de decisión mejoran la consistencia y transparencia de la clasificación del estatus de conservación de especies. Al aplicar estos protocolos, típicamente se requiere que los asesores tomen decisiones con estimaciones de atributos que inherentemente son inciertos. Los datos de entrada y las clasificaciones resultantes generalmente son tratadas como si fueran exactas y, por lo tanto, sin error de operario. Investigamos el impacto de la interpretación de datos sobre la consistencia de protocolos de clasificación del riesgo de extinción y diagnosticamos las causas de discrepancias*

cuando ocurrieron. Probamos tres protocolos de clasificación sistemática ampliamente utilizados por World Conservation Union, NatureServe y Florida Fish and Wildlife Conservation Commission (FFWCC). Proporcionamos información idéntica de 13 especies a 18 asesores para que infirieran estimaciones de cada uno de los parámetros requeridos por los tres protocolos. La clasificación de amenaza de varias de las especies varió desde riesgo bajo hasta riesgo alto, dependiendo de quien hizo la evaluación. Esto ocurrió con los tres protocolos investigados. Los asesores tendieron a coincidir en la ubicación de especies en categorías de riesgo más altas (50-70%) y más bajas (20-40%), pero hubo poco acuerdo en las especies que debían ubicarse en las categorías intermedias. Más aun, la correspondencia entre los tres métodos de clasificación fue impredecible, con gran variación entre asesores. Estos resultados resaltan la importancia de la revisión por pares, del consenso en las clasificaciones de especies entre múltiples asesores y de la necesidad de tener cautela con las evaluaciones realizadas por un solo asesor. Para que haya mayor consistencia entre asesores se requiere del uso extensivo de manuales de entrenamiento y de métodos formales para la estimación de parámetros que permitan la representación de incertidumbres mediante cadenas de cálculos, y que sean reportados con transparencia.

Palabras Clave: error del operario, estatus de conservación, incertidumbre, listas de especies amenazadas, Lista Roja IUCN, protocolos de clasificación, NatureServe

Introduction

Government and nongovernmental organizations around the world use a broad range of classification protocols to evaluate the level of threat to species (Millsap et al. 1990; Master 1991; Lunney et al. 1996; Keith 1998; Carter et al. 2000; Baldi et al. 2001; IUCN 2001). These protocols vary from qualitative assessments, where threat level is inferred from information on the species such as geographic range, population size, number of populations, and trends in these attributes based on expert judgment (Master 1991), to more objective methods that use rules or point-scoring approaches which infer threat from these attributes (Millsap et al. 1990; Lunney et al. 1996; Carter et al. 2000; IUCN 2001). Some protocols aim only to classify species according to their extinction risk, whereas others aim to rank species for conservation actions based on their extinction risk as well as other factors. Assessments are often summarized in lists of threatened species. These lists are used to assess potential adverse impacts on species, to determine whether there is a need for legal or other protections of species, to set management priorities for resource allocation (e.g., reserve design and recovery planning), and for state of the environment reporting (Possingham et al. 2002).

Developers and proponents of classification protocols believe these protocols will result in classifications that are consistent, transparent, and relatively accurate. For example, Mace and Lande (1991:149) stated in reference to the IUCN protocol that "...it is essential that the system used to establish the level of threat be consistent and clearly understood..." Partners in Flight developed a point-scoring scheme motivated by the need to "...develop clear and consistent priorities among several hundred land bird species based on their vulnerability to extinction" (Carter et al. 2000: 541). Similarly, Millsap et al. (1990: 8) developed a point-scoring scheme to assess wildlife in Florida and stated their method provided a "...

means for ensuring consistency and accuracy in the evaluations."

Application of the protocols typically requires assessors to estimate attributes such as geographic range, population size, trends, and threats. Information regarding these attributes is inherently uncertain because available data are often fragmentary. The uncertainty in applying the protocols can be categorized as two major types: epistemic and linguistic (Regan et al. 2002). Epistemic uncertainty is the uncertainty associated with our knowledge of the state of a system. It includes measurement error, natural variation over time and space, and the subjective differences that arise when experts interpret data or make judgments about criteria. Linguistic uncertainty is associated with use of language. It includes vocabulary that is vague, ambiguous, context dependent, or not sufficiently specific (Regan et al. 2002). It arises in threat classification protocols in the way different assessors interpret definitions of criteria, despite attempts to make the definitions exact (IUCN 2001). The uncertainties are not mutually exclusive, and when assessing species many sources of uncertainty are in play simultaneously. The effect they have on threat classifications is largely unknown.

Classification protocols provide a simple and flexible method for determining species' vulnerability to extinction. This flexibility may come with costs. Parameters are uncertain, yet this is acknowledged rarely. There are no guarantees that different assessors will produce consistent classifications, and no empirical studies have examined this issue. We investigated the effect of data interpretation on the consistency of threat classifications and highlight causes of discrepancies when they occurred. We tested three widely used classification protocols employed by the World Conservation Union (IUCN 2001), NatureServe (Regan et al. 2004), and the Florida Fish and Wildlife Conservation Commission (FFWCC) (Millsap et al. 1990).

We also investigated whether the classification protocols produced consistent relative ranks of species. A common goal of classification protocols is to highlight those species that are of high conservation concern (Keith 1998; Mace & Hudson 1999). Yet protocols vary in structure and data requirements, and these structural differences are compounded by uncertainty in the information available and how it is interpreted. Two studies (Burgman et al. 1999; O'Grady et al. 2004) compared different classification protocols and highlighted the structural differences between the methods. In both studies, however, the protocols were operated by a single assessor and thus the studies did not address the effect of variations in interpretation of the data and criteria (i.e., by different assessors) on the resulting correlations.

Methods

Threatened Species Classification Protocols

The three threatened species classification protocols we evaluated—IUCN, NatureServe, and FFWCC—encompass a broad range of parameters and structural features that are representative of those used in a variety of other protocols adopted by other agencies (Andelman et al. 2004). The IUCN Red List criteria and NatureServe methods were designed to estimate the relative risk of extinction, whereas the method adopted by FFWCC is predominantly used for setting management priorities. For the protocols to be comparable we delimited the variables of the FFWCC system to those that reflect extinction risk. This is explained in detail in the section on the FFWCC protocol.

IUCN Red List Criteria

The IUCN Red List Criteria are a set of decision rules that have five criteria (A–E), which require ecological data on range size, population size, rates of decline, and other parameters. The threshold values for each parameter differ for each of the threat categories, critically endangered (CR), endangered (EN), and vulnerable (VU). Meeting any one of these criteria qualifies a taxon for listing at that level of threat. We ignored criterion E, which involves a quantitative analysis of the risk of extinction.

All species that appear in the IUCN Red List database (<http://www.redlist.org/>) are assessed under criteria A–E by experts or species specialist groups through discussion and peer review (Lamoreux et al. 2003). We omitted the peer-review process to illustrate the issues associated only with interpretation of data and the effect this can have on threatened species assessments. In addition, the IUCN developed (in May 2003) a set of detailed guidelines for using categories and criteria (<http://www.iucn.org/themes/ssc/redlists/Redlistguidelines2003.pdf>) that were not available when we conducted our study.

NatureServe (Formerly Heritage and Association of Biodiversity Information)

The NatureServe protocol assesses species according to 12 biological and external factors that may affect their persistence (Master 1991). Each factor is scored into quantitative ranges, and then experts use this information in a qualitative manner to derive a “conservation status rank,” a numeric code that reflects the relative risk of extinction of an element (a species or an ecological community or system) (Master et al. 2003). A G1 ranking implies an element is critically imperiled at a global scale, and a G5 ranking implies the element is globally secure and abundant. The method in current use applies quantitative thresholds, but the relative weight given to each factor may differ depending on the interpretation of individual assessors. We used an objective conditional point-scoring approach developed by Regan et al. (2004) that approximates the current method.

The Florida Fish and Wildlife Conservation Commission

The FFWCC has the responsibility of conserving Florida's wildlife species. Millsap et al. (1990) designed a point scoring prioritization scheme (referred to here as the FFWCC) to rank Florida's vertebrate fauna according to their conservation needs. Scores are assigned based on biological, action, and supplemental variables. Biological variables reflect the population status or life history of the species, action variables reflect the current knowledge and status of conservation effort, and supplemental variables reflect biogeographic, systematic, and political attributes. In their use of only biological variables, Millsap et al. (1990) indicate thresholds for threat categories. A score ≥ 33 places species in the EN category, between 33 and 29 a species is deemed as threatened (TH), and between 29 and 24 points a species is considered of special concern (SSC). A score of < 24 points indicates the species is thought to be nonthreatened (NT). To enable a comparison with the other two protocols we considered only biological variables.

Species and Assessors

We chose 13 species to represent a broad spectrum of taxonomic groups and life histories, including vascular plants, birds, invertebrates, mammals, a reptile, and a fish (Fig. 1). In several cases only an isolated portion of the species' total range was considered in the assessment: the keeled snail, grass tree, Florida Scrub-Jay, and desert tortoise (scientific names provided in Fig. 1). As a result our assessments are different from previously published assessments for the full species ranges.

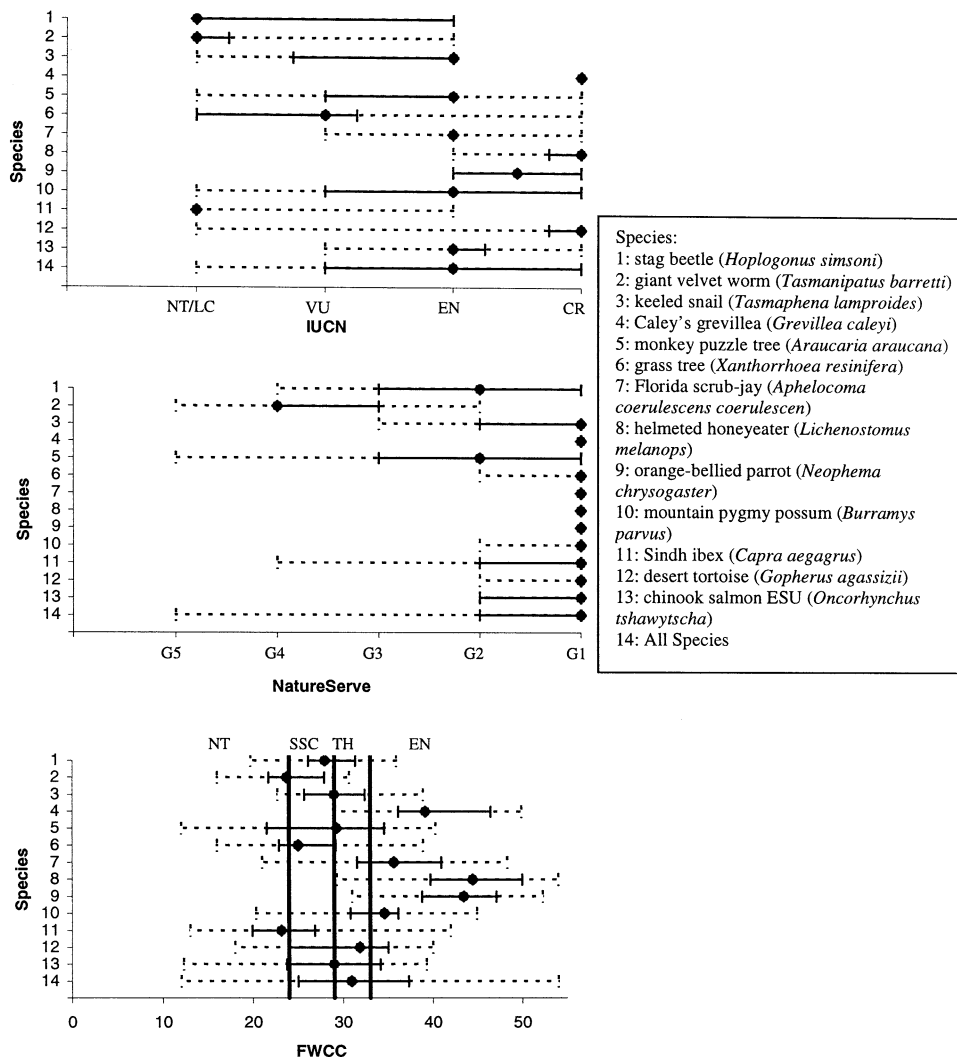


Figure 1. Threat classifications for 13 species performed by 18 assessors according to three protocols: IUCN, NatureServe, and FFWCC. Points on graph represent median classifications, dark lines represent interquartile ranges, and dashed lines represent breadth of categories.

Eighteen assessors took part in the experiment. Each assessor inferred estimates for each parameter used in the three classification protocols for each species. All assessors were ecologists. The experts ranged from those who regularly use the classification protocols to assess species, to those who were experts on individual species, to those who had a working knowledge of the protocols and could apply them to data sets. In all cases, the participants had sufficient knowledge that they could do the tasks credibly. We categorized assessors into two groups: those very familiar with the protocols (8 assessors) and those initially unfamiliar with either the protocols or the species (10 assessors). Of the assessors who were familiar with the protocols, two regularly use the protocols to assess species.

Eliminating Two Sources of Uncertainty

We eliminated uncertainty that derives from experts having different information at their disposal. Thus, the results underestimate the breadth of uncertainty that would

typically occur between assessors. To ensure equivalent information, T.J.R. compiled species profiles by collating available information from published and gray literature. The profiles included information on life-history traits, geographic range, population size, trends in range size and population numbers, threats, and management activities. Assessors used only the information in these species profiles to derive parameter estimates.

Another source of uncertainty enters most classifications when assessors make mistakes interpreting thresholds and criteria. We automated the application of each classification protocol to ensure implementation would be consistent across all assessors. Each assessor was required only to interpret the definitions of the criteria and the available information from the species profiles and make parameter estimates for each of the data requirements for each method. Fifty-three estimates were required for each species to provide assessments under the three protocols. Algorithms programmed into a spreadsheet organized parameter estimates, matched them to the relevant protocol, and assigned risk categories for

each species for each of the protocols according to the methods described in the relevant literature (Millsap et al. 1990; IUCN 2001; Regan et al. 2004). We structured data collection in this way so that differences could be attributed to interpretation of data and definitions rather than to differences in the underlying data or to operator error resulting from application of the protocols. This effectively eliminated some of the variation that would be present in most real-world applications.

Results

Consistency among Assessors

For the three classification protocols there was large variation in species classifications, with many assessments spanning more than one category of threat (Fig. 1). For several species the classification spanned all categories of threat, from “safe” to the most threatened categories.

Under the IUCN Red List criteria, assessors agreed unanimously on the classification of only one species, Caley’s grevillea, a small Australian shrub that has a very restricted range of <30 ha. All parameter estimates by all assessors resulted in the species being classified as CR, largely because of its restricted range. All other species’ classifications spanned more than one category of threat, with four species’ classifications spanning all categories from near threatened/least concern (NT/LC) to CR (Fig. 1). The interquartile range indicated several species were assessed within a similar category, whereas the assessment of seven species resulted in an interquartile range that spanned two or more categories.

Eleven species had a median classification in the most threatened category (G1) under the NatureServe protocol. Nine species spanned more than one category of risk with six of these species’ interquartile range spanning two or more categories. Under the FFWCC protocol nine species spanned all categories, from NT to EN, and there were no species for which there was complete agreement among all assessors. Similarly, the interquartile range of 10 species under the FFWCC protocol spanned more than two categories with 3 species’ interquartile ranges spanning all categories of threat. Overall the IUCN Red List criteria tended to rank species as EN, NatureServe as G1, and FFWCC as TH.

The distribution of species’ classifications among assessors varied depending on the species and the protocol implemented (Fig. 1). Under the IUCN the most varied classifications were for the monkey puzzle tree, grass tree, and mountain pygmy possum. Under the NatureServe protocol the most varied classifications were for the stag beetle, giant velvet worm, monkey puzzle tree, and the Sindh ibex. In general, the classifications of species resulting from the FFWCC protocol were the most varied, with only three species having reasonably consistent classifications

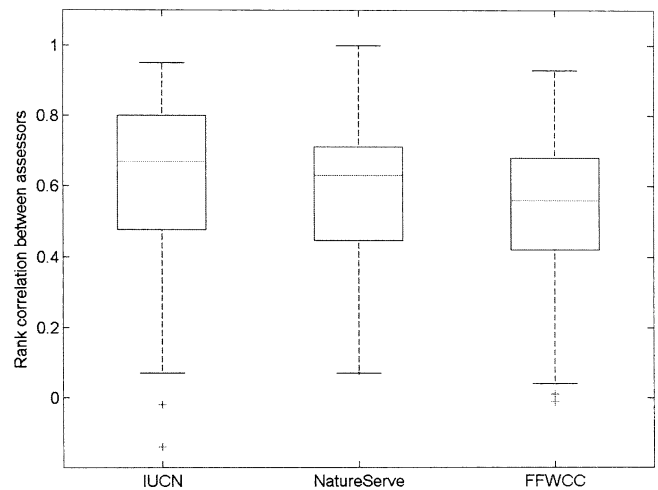


Figure 2. Box and whisker plots for rank correlation coefficients for pair-wise comparisons of species ranks between 18 assessors for the three classification protocols: IUCN, NatureServe, and FFWCC.

across all the assessors. Three species—Caley’s grevillea, Helmeted Honeyeater, and Orange-bellied Parrot—were consistently placed in the highest (NatureServe) or two highest categories (IUCN, FFWCC). All had very small range sizes and/or very small population sizes.

Rank correlations for pair-wise comparisons of species ranks between the 18 assessors under each of the classification protocols were highly variable for all three protocols (Fig. 2). The pair-wise comparisons ranged from small negative values to high positive values. The median rank correlation for pair-wise comparisons of assessors’ species ranks for the IUCN was 0.67 with an interquartile range of 0.48–0.8. For NatureServe and FFWCC the median rank correlations were 0.63 and 0.56 with interquartile ranges of 0.45–0.71 and 0.42–0.68, respectively.

Rank correlations provide insight into the similarity of the relative rank of species among assessors. An alternative measure of consistency is to determine how often two assessors categorize a particular species in the same risk category (represented as the proportion of species in a risk category that are the same for two assessors). For all classification protocols, assessors tended to agree on their placement of species in the highest (50–70%) and lowest risk categories (20–40%), but there was poor agreement on which species should be placed in the intermediate categories (Table 1).

Consistency among Classification Protocols

The consistency among the protocols was determined using pair-wise comparisons for each protocol combination for the species ranks of each assessor (Fig. 3). The median correlation coefficient between the IUCN and NatureServe was 0.46 (range 0 to 0.69), depending on the

Table 1. The average proportion (SD) of species in a risk category* that are the same for pair-wise comparisons between assessors.

Classification protocol	Average proportion			
	CR	EN	VU	NT/LC
IUCN	0.52 (0.22)	0.31 (0.16)	0.09 (0.17)	0.42 (0.27)
NatureServe	G1	G2	G3	G4/G5
	0.71 (0.13)	0.14 (0.19)	0.10 (0.23)	0.35 (0.40)
FFWCC	EN	TH	SSC	NT
	0.53 (0.17)	0.09 (0.16)	0.18 (0.20)	0.21 (0.20)

*Categories: CR, critically endangered; EN, endangered; VU, vulnerable; NT/LC, not threatened/least concern; G1–G5, range of values from critically imperiled at G1 to globally abundant and secure at G5; TH, threatened; SSC, species of special concern; NT, not threatened.

individual assessor. Similarly the median correlation coefficient for IUCN and FFWCC was 0.58 and the range was from 0.17 to 0.82. For NatureServe and FFWCC the median correlation was 0.42 with a range of 0.08 to 0.80. All correlations were positive except for one correlation between IUCN and NatureServe of zero.

Cause of Discrepancies

Discrepancies among Assessors

Discrepancies among assessors could not be attributed to particular individuals or the experience of the assessor. All assessors deviated from the most common categories several times, depending on the protocol. Some assessors deviated more than others for particular protocols, but this was also variable across the protocols.

We attributed causes of discrepancies in species assessments to the following general reasons: unwillingness of some assessors to make inferences about a particular parameter given the information available, variation in parameter estimates spanning more than one category, and differences in opinion regarding the influence of some factors on a population (e.g., whether the population was experiencing extreme fluctuations). Other minor causes for discrepancies were inconsistent logic and mistakes entering data on the spreadsheet.

The effect of discrepancies in parameter estimates on the classification of a species differed depending on the structure and driving parameters of each protocol. To indicate the effect of these causes of discrepancies, we analyzed the classification protocols separately.

The majority of discrepancies in classifications under the IUCN system occurred when the variation in parameter estimates among assessors invoked alternative criteria, slightly different parameter estimates fell on either side of category thresholds, or assessors made mistakes in calculations or did not infer parameter estimates even if sufficient information was provided. For example, the mountain pygmy possum had the most varied IUCN classification, covering the breadth of the four categories (NT/LC–CR). Differing interpretations of the data for continuing

declines was one of the main reasons for the discrepancies between assessors. From the information available, five assessors did not believe there was a continuing decline. All assessors who listed this species as NT/LC or VU would have listed EN or CR under criterion B if they answered “yes” to continuing decline. Two of the assessors who answered “no” entered other parameters (estimated continuing decline and/or the type of continuing decline) that were inconsistent with their assumption of no continuing decline. Rectifying these inconsistencies would have upgraded these two classifications to EN under criterion B. Second, the parameter estimates for geographic distribution (extent of occurrence [EOO] or area of occupancy [AOO]) were on either side of the threshold between EN and CR (EOO < 100 km²; AOO < 10 km²). This occurred even though the estimates for these parameters by different assessors spanned a relatively small range of values.

The driving parameter in NatureServe assessments is the number of element occurrences (i.e., populations). The value of this parameter determines an initial rank. If this parameter is not estimated, then the element is

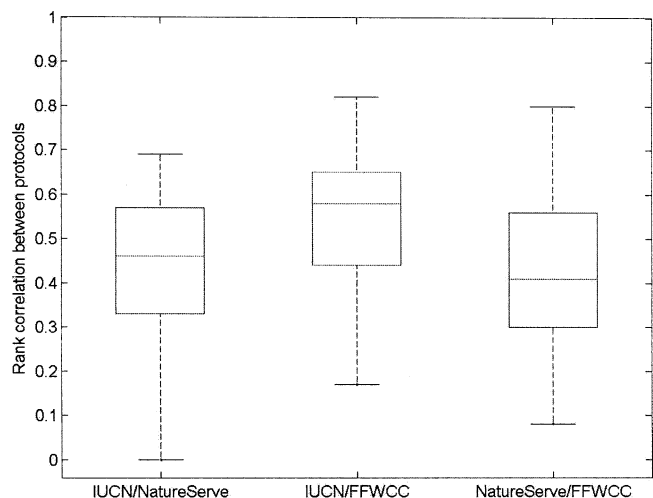


Figure 3. Box and whisker plots for rank correlation coefficients for each protocol combination for the species ranks of each assessor.

given an initial rank of G3.5. The variation in assessor estimates for the number of element occurrences was the primary cause of discrepancies in assessments. Coupled with this, the evaluation of any realized threats based on the scope, severity, and immediacy caused the species to be ranked in different categories. Other minor discrepancies were due to short-term trends and estimates for population size.

Assessments based on FFWCC were the most varied. All species were classified over more than one category with many spanning the full breadth of the threat categories (NT-EN). The protocol has a point-scoring structure whereby points for attributes are added together to give an overall score. This implies that species are affected by the range of factors working in concert rather than one or two threatening processes. Protocols with this type of structure are hindered by data availability and whether assessors are willing to make inferences based on available information because better known species can be classified as more threatened. Species that are less known will tend to be classified as less threatened. Another main cause of discrepancies in FFWCC assessments was that threat categories were delineated by small differences (5 points). Thus even small differences in scoring between assessors had a large effect on the category of threat a species was assigned. Generally the parameters that had the most varied scores (0–10 points) across the assessors for all species were population concentration, ecological specialization, percent change in AOO over the last 200 years, extent of occurrence, and number of individuals.

Discrepancies among Protocols

The differences in the ranks of species among methods were due to the structural differences of each of the methods. The methods have different data requirements and arbitrary thresholds that delineate categories. The differences in how the data were combined and weighted made consistent assessment difficult.

NatureServe and the IUCN protocols have similar data requirements. They differ in that the NatureServe protocol considers threats explicitly whereas the IUCN protocol considers threats implicitly through estimation of continuing decline and reductions in the future. They also differ in the way data are combined and weighted. The IUCN uses decision rules whereby a species' rank is ultimately determined by either one piece of information (e.g., criterion D, small population size) or a combination of interacting attributes (e.g., criterion B, restricted range coupled with a combination of continuing decline, severe fragmentation, or extreme fluctuations). The criterion that determines the rank is the one that returns the highest risk estimate from available data, and this varies from case to case. In contrast, NatureServe considers all available information for ecological attributes to derive a rank, but some attributes are primarily used as surro-

gates for missing data (e.g., environmental specificity is considered only if number of occurrences and AOO are unknown). Also the number of element occurrences (i.e., subpopulations) is given a large weighting that drives the overall ranking.

The FFWCC differs in its data requirements from the IUCN and NatureServe protocols because it does not use any information about current threats, future population size, future distributions, or future trends in population and distribution. The difference in the structure of the FFWCC protocol from IUCN is that all information for attributes is used. Missing data are ignored and given a value of zero, implying an optimistic value for the attribute.

Discussion

Our results show that substantial operator errors may be involved when risk classification protocols are applied by different assessors in isolation. Classifications can range from low risk to high risk of extinction depending on the assessors' interpretation of the available data. These operator errors may produce inconsistent and unpredictable classifications of species' extinction risk, despite the strict guidelines and explicit definitions incorporated into the protocols to improve their consistency. This emphasizes the importance of discussion and review to reduce inconsistencies and misinterpretations in the assessment process.

Although there were high positive correlations among assessors' classifications of species, this was not consistent across assessors. The average correlations among assessors were moderate for all three protocols but varied substantially. There tended to be better agreement between assessors for species in the highest and lowest risk categories rather than the intermediate risk categories. Higher levels of agreement under the NatureServe protocol for species placed in the highest-risk category were predominantly due to the method producing precautionary assessments (Regan et al. 2004).

Many of the discrepancies in species' classifications could be rectified with detailed discussion or consensus meetings between individuals regarding their interpretation of definitions and how they estimated parameter values. Inconsistencies in logic, varied interpretations, and errors could be resolved through this process, producing more consistent assessments. This approach has been adopted by the IUCN in their Species Survival Commission (SSC), which comprises a network of >7000 expert volunteers who carry out assessments in Specialist Groups for particular taxonomic groups. NatureServe uses an informal peer-review process in which global conservation status assessments are reviewed by experts. Group decisions, however, do not necessarily reflect agreement. It may submerge honest disagreements

about parameter estimates. Instead, for discrepancies that remain, the development of standard methods for incorporating, communicating, and reporting differences among several individuals directly is required.

One method for incorporating uncertainty is to use fuzzy boundaries rather than sharp cut-off values. In this approach, species classifications can span more than one category of threat (Akçakaya et al. 2000; Regan & Colyvan 2000). This approach has been adopted by the IUCN for single-assessor situations but could be extended to incorporate uncertainty for multiple assessors. NatureServe uses range ranks that span more than one category depending on the degree of uncertainty, but the application is done in a somewhat qualitative manner rather than adopting formal methods for incorporating uncertainty.

Our results demonstrate that the correspondence among classification protocols is unpredictable. Rank correlations between methods were highly variable depending on the individual performing the assessments. The average correlations among the protocols were moderate (0.42–0.53) but ranged between 0.0 and 0.82. The variation in correlations among protocols encapsulates the results from the studies of Burgman et al. (1999) and O'Grady et al. (2004), in which correlations between the protocols covered the intervals (0.2, 0.5) and (0.58, 0.69), respectively. The disparity in correspondence between the methods is primarily due to the structural differences in each of the methods. The methods have different data requirements and arbitrary thresholds that delineate categories. The differences in how the data are combined and weighted make it difficult to achieve consistent results from different protocols.

Resolving the discrepancies between protocols is an arduous task. The structural differences and data requirements make it almost impossible to have consistent classifications across methods for a wide range of assessors. Peer review and discussion may highlight and eradicate some of the inconsistencies and discrepancies but the fundamental differences among the protocols lie in the structure of the individual methods, and the effect of this on the consistency of classifications will remain unpredictable. A recommendation for dealing with the structural differences in classification protocols is to have experts assess species according to a range of classification protocols and discuss the implications of using one method over another. The process may highlight where one method is performing better than another. For instance, when data are scarce, the rule-based approach may be preferred to a point-scoring approach because it has robust strategies for dealing with unknown data. The parameter weightings that are implicit in particular protocols may suit some purposes more than others.

The classification status of species has an important influence on the allocation of resources and often determines whether there is a need for legal or other protection of species. Our results suggest that the uncertainties

in parameter estimation and the structural differences between the classification protocols can lead to significant differences within and between species' threat classifications that are difficult to control even when the method is systematic. This is of concern because the outputs from these types of protocols are rarely questioned. The uncertainties are rarely communicated in lists and related forms of communication, even when methods exist for incorporating uncertainties from single assessors (e.g., Akçakaya et al. 2000; Master et al. 2000).

A solution to some of these problems lies in developing formal methods for estimating parameters that allow uncertainties to be represented, carried through chains of calculations, and reported transparently. The development and widespread use of training materials (such as the IUCN Red List guidelines released in May 2003) will help ensure greater consistency among assessors. A more open and transparent review process that leads to more confidence in ranks could be achieved by making all the data considered in assessments available, along with the protocol criteria. Classifications of threatened species should not be based on evaluations made by a single individual. Instead, several individuals should separately rank species, before a peer review or consensus meeting that attempts to identify and reconcile any errors, inconsistencies, and differences in interpretation of data and criteria definition. For discrepancies that remain, methods for incorporating and reporting these differences from multiple assessors directly into assessments are required.

Acknowledgments

We thank the many people who gave their time to assess species and are not listed as authors (S. Bekessy, S. Bidwell, C. Bowles, T. Browning, Y. Chee, P. Clarke, C. Drill, P. Lucas, T. Mizarek, H. Possingham, H. Regan, M. Ruckelshaus, and T. Walshe) and those who provided species information. We also thank R. Akçakaya, who provided helpful comments on the manuscript. This work was conducted as part of the Extinction Risk Working Group supported by the Australian Research Council and the National Center for Ecological Analysis and Synthesis, a center funded by National Science Foundation (grant DEB-0072909), the University of California, and the Santa Barbara campus.

Literature Cited

- Akçakaya, H. R., S. Ferson, M. A. Burgman, D. A. Keith, G. M. Mace, and C. R. Todd. 2000. Making consistent IUCN classifications under uncertainty. *Conservation Biology* 14:1001–1013.
- Andelman, S. J., C. Groves, and H. M. Regan. 2004. A review of protocols for selecting species at risk in the context of US Forest Service viability assessments. *Acta Oecologica* 26:75–83.

- Baldi, A., G. Csorba, and Z. Korso. 2001. Setting priorities for the conservation of terrestrial vertebrates in Hungary. *Biodiversity and Conservation* **10**:1283–1296.
- Burgman, M. A., D. A. Keith, and T. V. Walshe. 1999. Uncertainty in comparative risk analysis for threatened Australian plant species. *Risk Analysis* **19**:579–592.
- Carter, M. F., W. C. Hunter, D. N. Pashley, and K. V. Rosenberg. 2000. Setting conservation priorities for land birds in the United States: the Partners in Flight Approach. *The Auk* **117**:541–548.
- IUCN (World Conservation Union). 2001. IUCN Red list categories and criteria: version 3.1. IUCN Species Survival Commission. IUCN, Gland, Switzerland.
- Keith, D. A. 1998. An evaluation and modification of World Conservation Union Red List criteria for classification of extinction risk in vascular plants. *Conservation Biology* **12**:1076–1090.
- Lamoreux, J. H., et al. 2003. Value of IUCN Red List. *Trends in Ecology & Evolution* **18**(5):214–215.
- Lunney, D., A. Curtin, D. Ayers, H. G. Cogger, and C. R. Dickman. 1996. An ecological approach to identifying the endangered fauna of New South Wales. *Pacific Conservation Biology* **2**:212–231.
- Mace, G. M., and E. J. Hudson. 1999. Attitudes toward sustainability and extinction. *Conservation Biology* **13**:242–246.
- Mace, G. M., and R. Lande. 1991. Assessing extinction threats: toward a re-evaluation of IUCN threatened species categories. *Conservation Biology* **5**:148–157.
- Master, L. L. 1991. Assessing threats and setting priorities for conservation. *Conservation Biology* **5**:559–563.
- Master, L. L., B. A. Stein, L. S. Kutner, and G. Hammerson. 2000. Vanishing assets: conservation status of US species. Pages 93–118 in B. A. Stein, L. S. Kutner, and J. S. Adams, editors. *Precious heritage: status of biodiversity in the United States*. Oxford University Press, New York.
- Master, L. L., L. E. Morse, A. S. Weakley, G. A. Hammerson, and D. Faber-Langendoen. 2003. NatureServe conservation status factors. NatureServe, Arlington, Virginia.
- Millspaugh, B. A., J. A. Gore, D. E. Runde, and S. I. Cerulean. 1990. Setting priorities for the conservation of fish and wildlife species in Florida. *Wildlife Monographs* **111**:1–57.
- O'Grady, J. J., M. A. Burgman, D. A. Keith, L. L. Master, S. J. Andelman, B. W. Brook, G. A. Hammerson, T. J. Regan, and R. Frankham. 2004. Correlations among extinction risks assessed by different threatened species categorization systems. *Conservation Biology* **18**:1624–1635.
- Possingham, H. P., S. J. Andelman, M. A. Burgman, R. A. Medellin, L. L. Master, and D. A. Keith. 2002. Limits to the use of threatened species lists. *Trends in Ecology & Evolution* **17**:503–507.
- Regan, H. M., and M. Colyvan. 2000. Fuzzy sets and threatened species classification. *Conservation Biology* **14**:1197–1199.
- Regan, H. M., M. Colyvan, and M. A. Burgman. 2002. A taxonomy and treatment of uncertainty for ecology and conservation biology. *Ecological Applications* **12**:618–628.
- Regan, T. J., L. L. Master, and G. A. Hammerson. 2004. Capturing expert knowledge for threatened species assessments: a case study of NatureServe conservation status ranks. *Acta Oecologica* **26**:95–107.

