Assessing the assessments: evaluation of four impact assessment protocols for invasive alien species

Anne Turbé1†, Diederik Strubbe2,3,4†, Emiliano Mori5, Martina Carrete6,7, François Chiron8, Philippe Clergeau9, Pablo González-Moreno10, Marine Le Louarn9,11, Alvaro Luna7, Mattia Menchetti12, Wolfgang Nentwig13, Liviu G. Părau14, Jose-Luis Postigo15, Wolfgang Rabitsch16, Juan Carlos Senar17, Simon Tollington18,19, Sonia Vanderhoeven20, Anne Weiserbs21 and Assaf Shwartz22

ABSTRACT

Aim Effective policy and management responses to the multiple threats posed by invasive alien species (IAS) rely on the ability to assess their impacts before conclusive empirical evidence is available. A plethora of different IAS risk and/or impact assessment protocols have been proposed, but it remains unclear whether, how and why the outcomes of such assessment protocols may differ.

Location Europe.

Methods Here, we present an in-depth evaluation and informed assessment of the consistency of four prominent protocols for assessing IAS impacts (EICAT, GiSS, Harmonia and NNRA), using two non-native parrots in Europe: the widespread ring-necked parakeet (Psittacula krameri) and the rapidly spreading monk parakeet (Myiopsitta monachus).

Results Our findings show that the procedures used to assess impacts may influence assessment outcomes. We find that robust IAS prioritization can be obtained by assessing species based on their most severe documented impacts, as all protocols yield consistent outcomes across impact categories. Additive impact scoring offers complementary, more subtle information that may be especially relevant for guiding management decisions regarding already established invasive alien species. Such management decisions will also strongly benefit from consensus approaches that reduce disagreement between experts, fostering the uptake of scientific advice into policy-making decisions.

Main conclusions Invasive alien species assessments should take advantage of the capacity of consensus assessments to consolidate discussion and agreement between experts. Our results suggest that decision-makers could use the assessment protocol most fit for their purpose, on the condition they apply a precautionary approach by considering the most severe impacts only. We also recommend that screening for high-impact IAS should be performed on a more robust basis than current ad hoc practices, at least using the easiest assessment protocols and reporting confidence scores.

Keywords biological invasions, confidence, consensus assessment, invasive alien species, invasive species policy, monk parakeet (Myiopsitta monachus), ring-necked parakeet (Psittacula krameri).
INTRODUCTION

Invasive alien species (IAS) are causing significant damage to the economy, society and the environment (Sala et al., 2000; Blackburn et al., 2014). Robust prioritization tools are therefore key to target the limited resources available for biosecurity and mitigation of high-impact IAS. Currently, a plethora of risk assessment protocols evaluating entry, exposure and consequence exists to determine which species are likely to have the strongest impacts in a risk assessment area. Despite the fact that common standards for risk analysis have been issued by international organizations such as the Food Agricultural Organisation, the World Organisation for Animal Health and the World Health Organisation, available protocols differ widely in their purpose, scope and methods, and each has its own way of characterizing IAS impacts (Sandvik et al., 2012; Roy et al., 2014; McGeoch et al., 2015). The demand for proper, unequivocal characterization of IAS impacts is set to amplify in face of increased introductions and global trade, as the magnitude of expected IAS impacts is a key component to decide whether management actions are required. To ensure effective IAS prioritization and smart resource allocation, it is thus essential to know whether the severity of impacts assigned to a given IAS is consistent irrespective of the protocol used (Schrader et al., 2012; Verbrugge et al., 2012). For example, to identify a set of ‘IAS of European concern’, the recent and ambitious IAS legislation in the European Union endorses the use of any risk assessment protocol, and thus of their associated characterizations of impacts, that meets certain minimum requirements (European Union, 2014; Tollington et al., 2015).

Comparative analyses of IAS impact characterizations obtained from different risk and/or impact assessment protocols are still largely missing in Europe (Essl et al., 2011; Verbrugge et al., 2012) but also world-wide (Dahlstrom et al., 2011). Yet, evaluation of IAS impacts is usually fraught with a high level of uncertainty due to a multitude of confounding factors. For example, impact assessments must deal with a lack of data on impacts in the invaded range and an often poor understanding of the underlying mechanisms promoting these impacts (Hulme et al., 2013; Kumschick et al., 2015). Consequently, impact protocols tend to be of a qualitative or semi-quantitative nature and rely heavily on expert opinion (Heikkilä, 2008; Dahlstrom et al., 2011; Strubbe et al., 2011; Leung et al., 2012; Verbrugge et al., 2012). Expert opinion can, however, be prone to a range of biases, yet objective methods for eliciting expert judgement and minimizing cognitive limitations and overconfidence in expert judgements are rarely used (Knol et al., 2010; Morgan, 2014; Sutherland & Burgman, 2015). Assessments also differ in the way impacts are defined, their geographical or temporal scales, the endpoint categories considered, as well as the specific scoring method applied (Verbrugge et al., 2010, 2012; Sandvik et al., 2012). Each of these differences in assumptions may have significant downstream effects on assigned impact scores, and consequently on policy and management actions. In fact, there are no general, transparent and repeatable assessment procedures to ensure the consistent use of existing protocols.

Here, we evaluate the consistency among the impacts estimated from four prominent, generic risk or impact assessment protocols employing both a standard (individual-based) assessment procedure and a consensus (group-based) assessment procedure, as well as two different methods commonly used for scoring impacts. We do this for two parakeet species (Aves, Psittaciformes) at contrasting invasion stages in Europe: the ring-necked parakeet (RNP, Psittacula krameri), which is already widespread in Europe (Parràu et al., 2016) and considered one of the continent’s worst avian invaders (DAISIE, 2009; Kumschick & Nentwig, 2010), and the monk parakeet (MP, Myiopsitta monachus), which is rapidly expanding and reportedly causing damage to infrastructure and agriculture in its native and other invaded ranges (Strubbe et al., 2011; Menchetti & Mori, 2014). This is the first in-depth study of this kind, and by focusing on two prominent avian invaders only, we are able to investigate not only the impact classifications assigned to invaders, but also how elicitation procedures affect the impact assessment process and its outcomes. We apply insights from expert appraisal of the four impact assessment protocols to formulate recommendations for carrying out robust and policy-relevant IAS impact assessments.

METHODS

Impact assessment protocols

We compared four prominent, generic risk or impact assessment protocols that can be applied at international, national or regional level to any taxonomic group and in any environment (EICAT, GISS, Harmonia+ and NNRA, Table 1). We focused on these four generic impact assessment protocols because they are the closest to meeting the EU minimum requirements (Roy et al., 2014) and are thus likely to be increasingly used in the future. Harmonia+ and the NNRA are in fact full risk assessment protocols, as they not only consider IAS impacts, but also evaluate likelihood of introduction and spread. As we focus on IAS impacts here, we only carried out the impact assessment module of these two protocols. EICAT and GISS are impact-only protocols, and these were carried out fully.

Impact assessment protocols typically consider one to three broad categories of impacts (such as environmental, economic and social impacts), each of which addresses a number of different impact mechanisms (e.g. predation, human health) formulated in questions. The Generic Impact Scoring System (GISS) is a generic protocol that measures the environmental and economic impacts of invasive alien species in 12 impact categories, and the protocol has already been widely used to compare IAS impacts in different regions and across taxonomic groups (Kumschick & Nentwig, 2010; Nentwig et al., 2010, 2016; Kumschick et al., 2011; Vaes-
Petignat & Nentwig, 2014; Laverty et al., 2015; van der Veer & Nentwig, 2015). The Environmental Impact Classification for Alien Taxa (EICAT) protocol was recently derived from GISS, but modified to classify species according to the magnitude of their detrimental environmental impacts, using a broad range of impact categories that correspond to the ones used by the IUCN Global Invasive Species Database (Blackburn et al., 2014; Hawkins et al., 2015). It considers 13 different environmental impact mechanisms, and the five levels of impacts are aligned and consistent across mechanisms. Harmonia* is designed to cover all types of IAS impacts in a unified framework aiming to be maximally compliant with international law (D’hondt et al., 2015). It presents 18 questions on impacts, including environmental impacts, plant, animal and human health impacts, as well as impacts on infrastructure. It provides ample and precise guidance with every question and is envisioned to be used in a multi-expert set-up to reach consensus scores for as many criteria as possible. The NNRA was developed for Great Britain (Baker et al., 2008) but has recently been updated and modified to be applicable to the whole EU. It is based on the EPPO risk assessment framework, which is recognized in international plant health regulations and comprises 18 questions on impacts, with a focus on potential biodiversity and ecosystem impacts.

These protocols have comparable scoring scales and coverage of different impact categories (except for EICAT, which covers only environmental impacts). Yet, even though all protocols instruct experts to score impacts under the assumption that the IAS under consideration occupies all suitable habitats in the assessed area (Baker et al., 2008; Kumschick & Nentwig, 2010; Blackburn et al., 2014; D’hondt et al., 2015), protocols differ substantially in their underlying assumptions (Table 1). For example, the NNRA protocol gives emphasis to biodiversity and ecosystem impacts (Baker et al., 2008), while Harmonia* concentrates on the mechanisms through which IAS cause impacts (D’hondt et al., 2015). Protocols also differ in how confidence of evaluators is measured and how total impact scores are aggregated (Roy et al., 2014). NNRA, GISS and Harmonia* consider impacts as additive by default but allow defining total impact based on a species’ most severe impact (i.e. based on the maximum impact score assigned to any question). Such precautionary impact scoring is the default approach in the EICAT protocol.

**Procedure for impact assessments**

A formal expert elicitation procedure was used to carry out a European-level impact assessment for both parakeets during a two-day workshop. As there are no widely accepted procedures for carrying out expert elicitations, in order to present a transparent and repeatable assessment framework, we broadly followed the seven step approach recommended by Knol et al. (2010). The first step in this approach is the ‘characterization of uncertainties’, that is clarifying which type of uncertain information is to be elicited. Here, the issue is the need to robustly classify IAS according to their (expected) impacts before conclusive empirical evidence is available. We apply the invasion of Europe by ring-necked and monk parakeets, as case studies for which substantial uncertainties regarding impacts remain (Kumschick & Nentwig, 2010; Strubbe et al., 2011).

The second step is to decide on the ‘scope and format of the elicitation’. Given our decision to focus on an in-depth evaluation of two avian invaders using four protocols in order to elucidate how elicitation procedures affect impact assessment outcomes, we opted for a two-day workshop to which a number of experts were invited to attend. On the first day, experts were asked to perform impact assessments independently, followed by a group (consensus) elicitation on the second day. We decided to perform the individual assessments during the workshop to standardize the conditions and information among participants and ensure a common starting point for the consensus assessment.

The third step concerns the ‘selection of experts’. Here, a well-balanced panel of 16 experts was put together, consisting of generalists (ecology/conservation, $n = 6$), as well as subject-matter experts (parakeet biology, $n = 6$) and normative experts (IAS policy/impact assessment, $n = 4$). Note that this panel selection included experts with known differing opinions regarding parakeet impacts and the way invasive

<table>
<thead>
<tr>
<th>Impact assessment protocol</th>
<th>Year of publication</th>
<th>Scoring system</th>
<th>Impact</th>
<th>Confidence</th>
<th>Impact categories: no. questions</th>
<th>Applied to: no. species</th>
</tr>
</thead>
<tbody>
<tr>
<td>EICAT*</td>
<td>2014</td>
<td>5 levels</td>
<td>Impact</td>
<td>3 levels</td>
<td>12 Env. 0 Eco. 0 Soc. 415</td>
<td></td>
</tr>
<tr>
<td>GISS†</td>
<td>2010</td>
<td>5 levels</td>
<td>Impact</td>
<td>3 levels</td>
<td>6 Env. 3 Eco. 3 Soc. 350</td>
<td></td>
</tr>
<tr>
<td>Harmonia‡</td>
<td>2015</td>
<td>5 levels</td>
<td>Impact</td>
<td>3 levels</td>
<td>6 Env. 9 Eco. 3 Soc. 5</td>
<td></td>
</tr>
<tr>
<td>NNRA§</td>
<td>2008</td>
<td>5 levels</td>
<td>Impact</td>
<td>4 levels</td>
<td>5 Env. 8 Eco. 4 Soc. 125</td>
<td></td>
</tr>
</tbody>
</table>

*Evans et al. (2016);
†Kumschick & Nentwig (2010); Nentwig et al. (2010, 2016);
‡D’hondt et al. (2015);
§Baker et al. (2008).
species policy should be conducted, thus ensuring the expert panel represented a breadth of perspectives. The panel also comprised experts with varying experience with the impact assessment protocols considered, from experts that had already used all protocols, to those that had only used one, or who knew about the protocols but had never used them, as well as to those who were completely unfamiliar with any of these protocols prior to the workshop.

In the fourth step, the 'design of the elicitation protocol' must be carefully appreciated, to ensure that the format of the elicitation is fit for its purpose. The main design decisions we took were (1) to present the four assessment protocols considered here to the participating experts in a standardized elicitation protocol worksheet and accompanying formal guidance. All experts were provided with the same information prior to the workshop to ensure a uniform knowledge base. This included an overview document and more than 40 relevant publications (both peer-reviewed papers and grey literature such as NGO/governmental reports) as suggested preliminary evidence base. This database was not meant to be exhaustive, and experts were encouraged to complement and share any additional evidence they were aware of. Additionally, we leveraged the ParrotNet (http://www.ke nt.ac.uk/parrotnet/) network to ask a member of each EU country with established parakeet populations to survey the literature on their impacts in their national language and to send a summary of any evidence found to the workshop coordinators.

A sixth step is then the actual 'elicitation of expert judgement'. This is the two-day workshop that took place in March 2015, and which started with a presentation of the scope and purpose of the meeting, informing the experts about what would be expected from them during the 2 days. This involved a clarification of some essential definitions (e.g. risks vs. impacts), a brief presentation of each of the four impact assessment protocols and their main assumptions, as well as highlighting the potential sources of cognitive biases in expert judgement (e.g. anchoring biases, availability biases, representativeness biases) or of other forms of biases (e.g. motivational biases). A summary of evidence of parakeet impacts published in national languages was also presented. On the first day, experts were asked to independently complete as many impact assessments as possible in a given order, so as to ensure an even coverage of the four protocols and two parakeet species. Experts had the possibility to ask questions about each of the schemes to a reference person. On the second day, two separate consensus impact assessments were carried out for the RNP with the Harmonia+ and NNRA protocols, using a modified Delphi process with a structured elicitation procedure (Burgman et al., 2011). For each question, a facilitator presented the distribution of answers from the independent assessments, summarized the available evidence, stimulated discussion and highlighted guidance for scoring impacts and confidence. All experts were then asked to vote simultaneously and anonymously, using 'clickers' (i.e. small handheld devices that record and transmit expert responses). After all evaluators had voted, a live visualization of the answers was presented. Consensus was assumed to be reached when two-thirds of the participants were in agreement. If no consensus was reached the first time, up to another two voting rounds were conducted, before which the facilitator stimulated further discussion to ensure the discrepancies reflected differences in expert judgement and were not due to overlooking or misinterpretation of evidence, nor to misunderstanding of the scoring rules (Knol et al., 2010).

As a seventh, last step, a 'post-elicitation questionnaire' was carried out. Two weeks after the workshop, experts were asked to fill out an online questionnaire to appraise the different protocols following a RACER evaluation framework (i.e. Relevant, Accepted, Credible, Easy and Robust (Lutter & Giljum, 2008)). The RACER framework was developed specifically to assess the value of scientific tools for use in policy-making (Lutter & Giljum, 2008). We devised 14 questions, four for the category Relevant (i.e. closely linked to the objectives to be reached), two for the category Accepted (i.e. by scientists and policymakers), three for the category Credible (i.e. for non-experts, unambiguous), one for the category Easy (i.e. easy to interpret and use) and four questions for the category Robust (e.g. the method can remain effective when applied to a variety of conditions, such as different taxonomic groups, data quality, scopes). Narrative summaries were provided for each RACER category, and respondents allocated a numerical score on a scale of 1 (criterion not fulfilled) to 5 (criterion fully fulfilled) to each question, supporting the visual presentation of the results in an easily readable overview summary table. The full questionnaire can be accessed via a link provided at the end of the acknowledgements section.

**Statistical analyses**

We rescaled data for each question between 0 and 1 to compare impact estimates across protocols. This was done by calculating \((V - V_{\min})/(V_{\max} - V_{\min})\), where \(V\) represents the impact score assigned to a question in the original dataset and \(\min\) and \(\max\) refer to the minimum and maximum scores attainable. For each of the four protocols, impact estimates were then aggregated into impact categories (total impact; environmental, economic and social impacts) in two ways, that is by taking the arithmetic mean (~additive impacts) and the maximum impacts (~impact defined on the basis of the most severe impact only). To test whether impact estimates varied among protocols and impact
categories, we included protocol and impact category and their interaction as fixed effects in a linear mixed model, with ‘rater’ (i.e. expert) as random effect (R library lme4 (R Bates et al., 2014)). Statistical model structure was thus as follows: \( \text{lm}('\text{impact estimate'} \sim '\text{protocol'} + '\text{impact category'} + '\text{protocol: impact category'} + (1|'rater')) \). Tukey’s post hoc tests for multiple comparisons (protocols and categories) were carried out using the R library multcomp (Hothorn et al., 2008). Separate analyses were carried out for arithmetic mean and for maximum impacts, and for RNP versus MP. To test whether impact estimates assigned to ring-necked parakeets differed between the individual and consensus assessment, we applied a linear model that included protocol, category, assessment method (consensus versus individual) and their two- and three-way interactions; models structure was \( \text{lm}('\text{impact} \sim '\text{protocol'} + '\text{assessment method'} + '\text{category'} + '\text{protocol: category'} + '\text{assessment method:category'} + '\text{protocol: category:assessment method'}) \). Note that as the consensus assessments were anonymous, it was not possible to include rater identity as a random factor in this analysis. Agreement between participants was assessed using Cronbach’s (R library psych (Revelle, 2014)) and Krippendorff’s alpha (R library irr (Gamer et al., 2012)). Both statistics vary from 0 to 1 and higher values indicate stronger agreement.

RACER responses were assessed with a linear mixed model with RACER category, protocol and their interaction as fixed effects, using RACER questions nested within experts as random effect. Model structure was \( \text{lmer('RACER response'} \sim '\text{protocol'} + '\text{RACER category'} + '\text{protocol: RACER category'} + (1|'rater'/RACER question')) \). As RACER categories A (‘accepted’) and E (‘easy’) cover only one and two questions, respectively, we present the results of an analysis on the R, C and R categories only below (although results are similar when including all categories, see Appendix S1 in Supporting Information, RACER). For all models described above, normality of model residuals was tested and verified (i.e. Shapiro–Wilk \( W > 0.91 \)).

**RESULTS**

**Consistency of results among protocols**

When considering impacts as additive, we find that across methods, MP and RNP total impact scores (summarizing environmental, economic and social impacts) vary between 0.19 and 0.45 (on a potential scale of 0–1). Comparing these scores against previously published estimates of invasive bird impacts (see Table S1 and Appendix S1) designates both parakeets as low- to mid-level impact species in Europe, in line with previously published national or regional impact assessments (Fig. 1, Table S1 and Appendix S1). For both species, total impact scores derived from individual assessments were similar between NNRA and Harmonia+, and significantly higher than those obtained with the GISS protocol (Figs 1 & 2, Appendix S1). In fact, GISS consistently assigned species the lowest impact scores across impact categories, although not all differences were statistically significant (Fig. 2, Appendix S1). The similarity between NNRA and Harmonia+ total impacts, however, masks differences among impact categories. For example, according to the NNRA protocol, both parakeet species are anticipated to have a relatively high impact on the economy, whereas Harmonia+ considers environmental impacts to be more severe. Note that according to the EICAT protocol (which focuses solely on environmental impacts), both species are considered to have a rather low impact on the environment. More consistent and low scores were obtained for impacts on society across all protocols (Figs 1 & 2, Appendix S1).

These additive impact results, however, contrast with protocol outcomes based only on species’ most severe documented impacts. Employing maximum impact estimates not only resulted in higher impact scores across all impact categories, but also in a much higher consistency across impact protocols (Fig. 2, Appendix S1). The only remaining significant differences among protocols are that economic impacts of both species and the social impacts of MP were estimated to be lower according to GISS than according to NNRA and Harmonia+ (EICAT does not consider these impact categories, Fig. 2, Appendix S1).

Apart from the estimated magnitude of impacts, our results show that the degree of confidence evaluators assigned to their judgements varied between impact protocols, but not between impact categories (Fig. 1, Appendix S1). NNRA confidence levels were consistently lower than those obtained for GISS and Harmonia+ (as well as EICAT, when assessing confidence levels for environmental impacts, Fig. 1). Expert agreement regarding estimated impacts and associated confidence levels was generally high for both species and in particular for the RNP (i.e. Cronbach’s alpha frequently \( > 0.7 \), Table 2, Appendix S1). Yet, according to Krippendorff’s alpha, expert estimates were not reliable enough to allow for strong conclusions (i.e. all alpha \( < 0.8 \), Table 2, Appendix S1).

**Consensus assessment vs. independent assessment**

The consensus assessment did not significantly modify RNP impact scores \((P = 0.69)\) or experts’ confidence levels \((P = 0.23)\) with the NNRA protocol, but it resulted in significantly higher impact scores \((P < 0.001)\) and a reduced degree of confidence in the estimates with Harmonia+ \((P = 0.031, \text{Fig. 1, Appendix S1})\). The agreement among experts increased according to Cronbach’s alpha (all \(\alpha > 0.73\)) and strongly according to Krippendorff’s alpha (all \(\alpha > 0.74\)), although agreement remained low regarding confidence estimates (Table 2, Appendix S1).

**RACER evaluation**

All protocols were viewed by the workshop participants as reasonably comprehensive, but they performed differently across the set of RACER criteria. Harmonia+ performed best, closely followed by GISS, and both protocols were generally thought...
to be more credible than the other two. NNRA earned the lowest appraisal across all categories (Table 3, Appendix S1, all Tukey’s HSD $P < 0.001$). Harmonia$^+$ was considered the most robust protocol, notably in terms of estimation procedures and methodology, while GISS and EICAT were considered the easiest to use (Table 3, Appendix S1).

**DISCUSSION**

The choice of acting against an invasive alien species, including the choice of no action, must be made before conclusive scientific evidence is available, as in many policy-related decisions (Morgan et al., 1992). This requires unambiguous identification of likely high-impact invaders. Using four prominent assessment protocols, we find that the RNP and MP cannot be assigned unequivocally to a consistent impact level. Apparent consistency in total scores masks discrepancies due to contrasting emphasis on different impact categories, how the assessments are conducted and the scores aggregated. Only when assessing species based on their most severe documented impacts, impact protocols yield largely consistent outcomes. Our findings suggest that clear guidelines, closed-
ended questions and the use of a consensus approach can considerably improve consistency among assessment outcomes. Assessment protocols introduce several implicit biases in the evaluation of IAS impacts. First, questions related to a given impact mechanism may not be independent. For example, while GISS and EICAT ask only one question about competition with native species and impact upon agriculture, NNRA involves three closely related questions. Given that the main impacts of parakeets relate to damage to agriculture and competition with native species, this dependence leads to high-impact scores on each of these related questions, resulting in higher arithmetic (i.e. additive) mean scores. Second, each protocol puts a higher emphasis on certain aspects of impacts. For example, while Harmonia+ is similar to GISS and EICAT regarding emphasis on competition and agricultural damage, it stresses the potential consequences of parasite and pathogen transmission more strongly than GISS does, leading to a higher score on environmental impacts. Indeed, both RNP and MP are known to act as hosts for a number of potentially harmful pathogens and parasites (Strubbe et al., 2011; Mori et al., 2015), although actual disease transmission has yet to be observed. Finally, the wording

Table 2 Expert agreement regarding ring-necked parakeet (Psittacula krameri) impacts (and associated expert confidence) across assessment protocols, as measured by Cronbach’s and Krippendorff’s alpha, for independent assessments and when applying a consensus approach. Higher values indicate stronger agreement.

<table>
<thead>
<tr>
<th>Protocol</th>
<th>Cronbach Impact</th>
<th>Cronbach Confidence</th>
<th>Krippendorff Impact</th>
<th>Krippendorff Confidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>RNP</td>
<td>0.58</td>
<td>x</td>
<td>0.74</td>
<td>x</td>
</tr>
<tr>
<td>EICAT*</td>
<td>0.82</td>
<td>x</td>
<td>0.77</td>
<td>x</td>
</tr>
<tr>
<td>GISS</td>
<td>0.68</td>
<td>0.85</td>
<td>0.89</td>
<td>0.97</td>
</tr>
<tr>
<td>Harmonia+</td>
<td>0.91</td>
<td>0.93</td>
<td>0.74</td>
<td>0.97</td>
</tr>
<tr>
<td>NNRA</td>
<td>0.68</td>
<td>0.91</td>
<td>0.74</td>
<td>0.97</td>
</tr>
</tbody>
</table>

*Calculated for environmental impacts only; x: consensus assessment not performed for these protocols because of time constraints (see main text).

Table 3 Summary results from the evaluation of the comprehensiveness of the different protocols according to RACER (Relevant, Accepted, Credible, Easy and Robust) criteria. The results show the weighted average of scores per subcategory and average summary scores per category in bold.

<table>
<thead>
<tr>
<th>Criterion</th>
<th>Relevant</th>
<th>Minimum standard</th>
<th>Coverage</th>
<th>Applicability at EU level</th>
<th>Sensitive</th>
<th>Accepted Academics</th>
<th>Policy-Makers</th>
<th>Credible</th>
<th>Unambiguous</th>
<th>Repeatable</th>
<th>Transparent</th>
<th>Easy</th>
<th>Expertise</th>
<th>Robust</th>
<th>Clear</th>
<th>Robust estimation procedure</th>
<th>Quantification of uncertainty</th>
<th>Timeframes</th>
<th>Summary comprehensiveness</th>
</tr>
</thead>
<tbody>
<tr>
<td>EICAT</td>
<td>3.65</td>
<td>2.9</td>
<td>3.4</td>
<td>4.7</td>
<td>2.8</td>
<td>2.8</td>
<td>3.1</td>
<td>2.5</td>
<td>3.5</td>
<td>2.5</td>
<td>4.3</td>
<td>4.3</td>
<td>3.7</td>
<td>4.3</td>
<td>4.3</td>
<td>3.5</td>
<td>3.5</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td>GISS</td>
<td>3.8</td>
<td>3.9</td>
<td>4.1</td>
<td>4.7</td>
<td>2.5</td>
<td>3.1</td>
<td>3.3</td>
<td>2.8</td>
<td>4.1</td>
<td>3.7</td>
<td>4.4</td>
<td>4.3</td>
<td>3.5</td>
<td>4.2</td>
<td>3.4</td>
<td>3.1</td>
<td>3.8</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>Harmonia+</td>
<td>4.08</td>
<td>3.9</td>
<td>4.6</td>
<td>4.8</td>
<td>3.0</td>
<td>3.6</td>
<td>3.9</td>
<td>3.3</td>
<td>4.4</td>
<td>4.0</td>
<td>4.6</td>
<td>4.6</td>
<td>4.2</td>
<td>4.2</td>
<td>4.6</td>
<td>3.5</td>
<td>4.0</td>
<td>11</td>
<td></td>
</tr>
<tr>
<td>NNRA</td>
<td>3.68</td>
<td>3.8</td>
<td>4.2</td>
<td>3.9</td>
<td>2.8</td>
<td>3.1</td>
<td>2.4</td>
<td>3.7</td>
<td>3.5</td>
<td>3.8</td>
<td>3.8</td>
<td>3.6</td>
<td>3.6</td>
<td>3.6</td>
<td>3.5</td>
<td>4.2</td>
<td>3.5</td>
<td>13</td>
<td></td>
</tr>
</tbody>
</table>


Diversity and Distributions, 1–11, © 2017 John Wiley & Sons Ltd
of the questions may lead to different interpretations between different assessors and bias the assessments.

Our results suggest that a number of simple aspects related to the structure and the language of the impact assessment protocol can improve the consistency of outcomes (see also Verbrugge et al., 2012). Harmonia+ and GISS were most appreciated by the workshop experts and were considered the most ‘credible’, whereas NNRA scored the lowest and consistently produced the lowest confidence scores. We argue that this is probably related to the question form and language clarity, as experts’ confidence tends to be lower with broad, open-ended questions than with more targeted choice questions (Ibabe & Sporer, 2004). Experts may be more reluctant to express strong confidence to general questions regarding impacts upon ‘biodiversity’, such as phrased in the NNRA. In contrast, higher confidence scores may be more easily assigned when questioned on specific impact mechanisms (such as competition with native species). GISS and EICAT protocols were considered the easiest to use, probably because both protocols contained brief, self-contained guidance within each question, and the questions were based on hierarchical statements clearly specifying the context and reference situation.

Experts use various heuristics when assessing uncertain information, which may introduce bias in the outcomes (Morgan et al., 1992). We found, as has been shown in other fields (Burgman et al., 2011; Morgan, 2014), that experts may have placed greater confidence in their judgement than fully warranted. They generally had high confidence in the impact estimates they provided, both for the relatively data-poor MP and for the relatively better-studied RNP. Although both parakeet species have been observed in Europe since the seventies, doubts remain about the magnitude of their impacts. While some studies evidence impacts on native fauna due to competition for nesting cavities or food for RNP and, for MP, predominantly anecdotal evidence about agricultural and infrastructure damage (Menchetti & Mori, 2014; Senar et al., 2016), it is unclear whether and how such locally measured impacts translate into significant damages or native species population declines at regional to biogeographical scales. Interestingly, while the magnitude of impacts and the confidence estimates assigned by the experts remained largely unchanged in the consensus assessment, significantly lower confidence estimates were obtained for Harmonia+ than in the independent assessments. This could be a moderating effect of the discussion, or more likely due to a better understanding of the protocol, because the facilitator helped minimize linguistic uncertainty.

The consensus assessment considerably reduced disagreement among experts. The expert panel represented a breadth of perspectives and when given the opportunity to listen to one another and cross-examine reasoning and data in a structured process, experts converged to similar judgements, as found in other studies (Burgman et al., 2011). It is recognized that misunderstandings can occur and result in arbitrary disagreements when the language used is ambiguous or vague, or insufficient baseline reference or context is provided (Carey & Burgman, 2008). In the consensus assessment, the facilitator helped minimize such linguistic uncertainty. Accordingly, while discussions allowed to calibrate expert responses according to the protocol guidelines, they also provided an opportunity to share any evidence that might not have been known by everyone and to discuss the quality of the evidence base. Experts were alerted to the presence of potential sources of cognitive biases, such as anchoring and availability biases that may lead assessors to rely too heavily on the first piece of information offered, or to base their judgements on immediate examples that come to mind (Knol et al., 2010). Along with the anonymity of the voting process, this should have minimized such influences, and remaining differences most likely reflect valuable differences in opinions (Knol et al., 2010; Morgan, 2014). Consensus assessments are thus useful to gain higher expert endorsement for uptake into policy decisions. The exact source of remaining disagreements is also important to flag for decision-makers, so that they can factor in risk elements or design alternatives to address these dissents.

The choice of aggregation method has strong implications on the magnitude of impact assigned to IAS. Lower total impact scores are obtained when impacts are considered additive, whereas higher impact estimates result from ranking species on the basis of their most severe (maximum) impact. While the maximum impact approach may be a justifiable application of the precautionary principle, it lowers the discriminative power by causing the impact scores to be highly skewed towards the maximum score attainable (D’hondt et al., 2015). Additive impact scoring on the other hand has an inherent moderating effect (Holt, 2006; D’hondt et al., 2015). This suggests that both scoring approaches convey somewhat different and complementary information. Maximum impact scoring may be especially well suited to horizon scanning exercises as the cost of erroneously allowing the introduction or spread of a high-impact invader is likely to be substantial. Maximum approaches could thus be used to identify and prevent the introduction of IAS with potentially high impacts, or for the rapid management of recently established and geographically restricted IAS. Given the large costs typically associated with managing or eradicating widespread IAS, the more discerning additive scoring approaches may be most suited to guide management decisions in these cases.

**CONCLUSIONS AND RECOMMENDATIONS FOR IAS POLICY AND MANAGEMENT**

We show that when IAS are evaluated based on their most severe documented impacts (as recommended by Blackburn et al., 2014), the four generic protocols applied in this study yield consistent results, irrespective of the assessment method (individual vs. group consensus). This suggests that while aggregating scores on the basis of maximum impacts could err on the side of caution, such a scoring does allow policy and decision-makers to use the protocol that best fits their
means and needs. Additive scoring approaches are complementary and may be particularly relevant for deciding how to prioritize actions on already established IAS. When screening for potential high-impact IAS, published impact assessment protocols are seldom used and evaluators often rely on a limited set of *ad hoc* questions (Roy et al., 2014, 2015). A more robust approach would consist in using the easiest protocols and accounting for confidence scores. Low expert confidence can, under the precautionary principle, be regarded as a reason to argue for action against the species under consideration. While our in-depth focus on two avian invaders allows us to uncover how assessment procedures can influence protocol outcomes, we should acknowledge that further testing of our conclusions on a larger number of species and taxonomic groups is needed. Many invasive species are even less studied than RNP and MP, and only a larger scale assessment can reveal how different protocols handle more severe data uncertainty. However, in any case, IAS assessments should take advantage of the capacity of consensus assessments to consolidate discussion and agreement between experts, and perform a critical appraisal of the (evidence for) mechanisms underlying invasive species impacts.

**ACKNOWLEDGEMENTS**

This study is a joint effort from a workshop organized under COST European Cooperation in Science and Technology Actions Parrotnet (ES1304) and Alien Challenge (TD1209). We would also like to thank two anonymous referees for helpful comments on an earlier version of this manuscript. D.S. is currently funded by a Marie Skłodowska-Curie Action under the Horizon 2020 call (H2020-MSCA-IF-2015, grant number 706318) and acknowledges the Danish National Research Foundation for support to the Center for Macroeology, Evolution and Climate (grant number DNRFP96). J.C.S. was supported by funds from the Ministry of Ecology, Evolution and Climate (grant number DNRF96). Diversity and Distributions, 1–11, 2017 John Wiley & Sons Ltd

**REFERENCES**


**SUPPORTING INFORMATION**

Additional Supporting Information may be found in the online version of this article:

*Table S1* Comparison of published national and regional impact assessments of the Ring-Necked Parakeet (RNP) and the Monk Parakeet (MP) with different impact protocols

*Appendix S1* Full statistical results of all analyses conducted (excel file)

---

**BIOSKETCH**

This research was conducted as part of the Parrotnet COST action https://www.kent.ac.uk/parrotnet and a contribution to Working Group 1 activities led by A.S., looking at parakeet impacts on society/economy/environment, and public perception.

Author contributions: A.T., D.S. and A.S. conceived the ideas; A.T., D.S., E.M. and A.S. prepared the data, organized and ran the workshop; P.G.M shared the datasheets prepared for Alien Challenge; all co-authors but P.G.M., W.N. and J.C.S., conducted impact assessments during the workshop; D.S. and A.T. analysed the data; and A.T., D.S. and A.S. led the writing.

Editor: Franz Essl