NeoBiota 76: 163–190 (2022)
doi: 10.3897/neobiota.76.83028
https://neobiota.pensoft.net

RESEARCH ARTICLE

A peer-reviewed open-access journal
NeoBiota
Advancing research on alien species and biological invasions

Consistency in impact assessments of invasive species is generally high and depends on protocols and impact types

Rubén Bernardo-Madrid[1], Pablo González-Moreno[2,3], Belinda Gallardo[4,5], Sven Bacher[6], Montserrat Vilà[1,7]

1 *Estación Biológica de Doñana (EBD), CSIC, Avda. Américo Vespucio 26, 41092 Seville, Spain* **2** *CABI, Bakeham Lane, Egham, TW20 9TY, UK* **3** *Evaluación y Restauración de Sistemas Agrícolas y Forestales RNM360, Department of Forestry Engineering, University of Córdoba, Córdoba, Spain* **4** *Instituto Pirenaico de Ecología (IPE), CSIC, Avda. Montañana, 1005, 50059 Zaragoza, Spain* **5** *BioRISC (Biosecurity Research Initiative at St Catharine's), St Catharine's College, Cambridge CB2 1RL, UK* **6** *University of Fribourg, Department of Biology, Unit of Ecology and Evolution, Fribourg, Switzerland* **7** *Department of Plant Biology and Ecology, University of Seville, 41012 Seville, Spain*

Corresponding author: Rubén Bernardo-Madrid (r.bernardo.madrid@gmail.com)

Academic editor: Marina Piria | Received 3 March 2022 | Accepted 21 July 2022 | Published 3 October 2022

Citation: Bernardo-Madrid R, González-Moreno P, Gallardo B, Bacher S, Vilà M (2022) Consistency in impact assessments of invasive species is generally high and depends on protocols and impact types. In: Giannetto D, Piria M, Tarkan AS, Zięba G (Eds) Recent advancements in the risk screening of freshwater and terrestrial non-native species. NeoBiota 76: 163–190. https://doi.org/10.3897/neobiota.76.83028

Abstract

Impact assessments can help prioritising limited resources for invasive species management. However, their usefulness to provide information for decision-making depends on their repeatability, i.e. the consistency of the estimated impact. Previous studies have provided important insights into the consistency of final scores and rankings. However, due to the criteria to summarise protocol responses into one value (e.g. maximum score observed) or to categorise those final scores into prioritisation levels, the real consistency at the answer level remains poorly understood. Here, we fill this gap by quantifying and comparing the consistency in the scores of protocol questions with inter-rater reliability metrics. We provide an overview of impact assessment consistency and the factors altering it, by evaluating 1,742 impact assessments of 60 terrestrial, freshwater and marine vertebrates, invertebrates and plants conducted with seven protocols applied in Europe (EICAT; EPPO; EPPO prioritisation; GABLIS; GB; GISS; and Harmonia+). Assessments include questions about diverse impact types: environment, biodiversity, native species interactions, hybridisation, economic losses and human health. Overall, the great majority of assessments (67%) showed high consistency; only a small minority (13%) presented low consistency. Consistency of responses did

not depend on species identity or the amount of information on their impacts, but partly depended on the impact type evaluated and the protocol used, probably due to linguistic uncertainties (pseudo-R^2 = 0.11 and 0.10, respectively). Consistency of responses was highest for questions on ecosystem and human health impacts and lowest for questions regarding biological interactions amongst alien and native species. Regarding protocols, consistency was highest with Harmonia⁺ and GISS and lowest with EPPO. The presence of few, but very low, consistent assessments indicates that there is room for improvement in the repeatability of assessments. As no single factor explained largely the variance in consistency, low values can rely on multiple factors. We thus endorse previous studies calling for diverse and complementary actions, such as improving protocols and guidelines or consensus assessment to increase impact assessment repeatability. Nevertheless, we conclude that impact assessments were generally highly consistent and, therefore, useful in helping to prioritise resources against the continued relentless rise of invasive species.

Keywords

Alien species policy, biological invasions, ecological impact, epistemic uncertainty, inter-rater reliability, linguistic uncertainty, repeatability, socio-economic impact

Introduction

Invasive alien species are one of the greatest threats to biodiversity, economy and public health (Bellard et al. 2016; Mazza and Tricarico 2018; Diagne et al. 2020; Pyšek et al. 2020; Smith 2020). Concern about invasive species is growing due to the relentless increase in introductions and their spread, mostly associated with environmental change and increasing trade (Seebens et al. 2015, 2017; Chapman et al. 2017; Sardain et al. 2019). Although there are significant national and international efforts to reduce introductions, spreads and their impacts (Keller and Perrings 2011; Turbelin et al. 2017), human operational capacity to avert new invasions is limited (Genovesi and Shine 2004; Keller et al. 2007; Early et al. 2016). Thus, reliable tools to prioritise and underpin invasive species research, management and policy are required (Roberts et al. 2018; Booy et al. 2020). Under this urgent need, systematic semi-quantitative impact assessment protocols, based on available scientific evidence to rank and prioritise management of alien species are of paramount usefulness (e.g. Genovesi and Shine 2004; McGeoch et al. 2016; Vilà et al. 2019; Vilizzi et al. 2021).

The large number of protocols developed with similar objectives, as well as the substantial body of research comparing their outputs, shows the pivotal role of protocol choice in assessments (Glamuzina et al. 2017; Turbé et al. 2017; Vilà et al. 2019; Sohrabi et al. 2021). While this is important, there are also other crucial and more undervalued aspects in impact assessments. Previous studies have frequently illustrated the varying consistency of results when evaluating the same species with the same protocols (McGeoch et al. 2012; Almeida et al. 2013; Lawson et al. 2015; Turbé et al. 2017; González-Moreno et al. 2019; Vilizzi et al. 2019; Clarke et al. 2021; but see Volery et al. 2021). This finding raises doubts as to whether the choice of the evaluator can affect management prioritisations and, thus, whether risk assessments are reliable for providing information for decision-making. The fluctuating consistency is partly to be

expected as assessments in the end rely on the judgement of experts (Regan et al. 2002; Burgman et al. 2011; McGeoch et al. 2012; Vanderhoeven et al. 2017) which depend on their experience, i.e. amount and bias of knowledge and their subjective interpretation of evidence (Kumschick et al. 2017; Dorrough et al. 2018; Bindewald et al. 2020; Clarke et al. 2021). Certainly, there have been advances to control this subjectiveness (e.g. refinement of guidelines and protocol questions, as well as peer review and consensus process; Hawkins et al. 2015; Matthews et al. 2017; Vanderhoeven et al. 2017; Dorrough et al. 2018; Volery et al. 2020). However, information on the overall severity and extent of consistency in responses is still missing. For instance, information on the factors underlying different degrees of consistency is mostly theoretical (Regan et al. 2002; Vanderhoeven et al. 2017; Latombe et al. 2019; Probert et al. 2020), while the limited empirical information focuses mainly on consistency in final scores and rankings (e.g. Perdikaris et al. 2016; González-Moreno et al. 2019). However, the protocol's criteria for synthesising scores (e.g. the mean or maximum value for choosing a final score) and the subjective threshold value for ranking species into different categories (Almeida et al. 2013; D'hondt et al. 2015; González-Moreno et al. 2019; Vilà et al. 2019) add unintuitive noise to the real consistency in answers. To date, studies focused on protocol questions are limited to a single taxon and a single protocol (e.g. Clarke et al. 2021; Volery et al. 2021). Thus, empirical information on the factors influencing the consistency across assessors remains poorly understood.

To fill both knowledge gaps, we addressed two objectives. Objective 1: To provide generalisable results on consistency in individual protocol questions, we evaluated consistency when assessing a wide range of taxa (invasive plants, vertebrates and invertebrates), as well as when using multiple protocols. We measured consistency in scores of protocol questions using inter-rater reliability metrics (Hallgren 2012; Gwet 2014) benefiting from one of the most comprehensive datasets on impact assessment of invasive species in Europe (described in González-Moreno et al. 2019). By exploring a wide range of taxa and protocols, our results will provide information for the overall reliability of impact assessments that support decision-making. Objective 2: To evaluate which factors may influence the consistency of responses, we evaluated the relationship between the consistency and the protocol choice, impact type (e.g. environmental, socio-economic), taxonomic group, species identity and the amount of scientific literature available about species impacts. The evaluation of these factors, except for protocols, aims to answer if consistency varies due to epistemic uncertainties, such as if assessors had different knowledge about impacts or responded with greater subjectivity (e.g. due to bias, limited or inconsistent knowledge; McGeoch et al. 2012, 2016; Kumschick et al. 2017). The evaluation of protocol choice aims to detect if consistency is associated with protocol properties (e.g. number of questions per protocol and of responses per question) or with linguistic uncertainties (e.g. clarity or vagueness of the questions). For details on epistemic and linguistic uncertainties, see Regan et al. (2002), Leung et al. (2012), Latombe et al. (2019) and Probert et al. (2020). Altogether, these results can form the basis of future studies to provide information for the design or update of impact assessment protocols for invasive species.

Materials and methods

Assessors, species and impact assessment protocols

Within the Alien Challenge COST Action, 78 assessors with variable experience in biological invasions (PhD or PhD candidates; hereafter assessors) evaluated 60 invasive species with seven different risk assessment protocols (hereafter protocols) to provide information about the agreement of scores in protocols (González-Moreno et al. 2019). In total, we used 1,742 of those impact assessments.

Assessors were grouped according to their taxonomic expertise, under the coordination of a taxonomic leader. Assessors selected by consensus a list of 60 invasive species that covered a wide range of habitat types and biological characteristics: terrestrial plants (n = 10), freshwater plants (5), terrestrial vertebrates (10), terrestrial insects (13), other terrestrial invertebrates (4), freshwater invertebrates (6), freshwater fish (3), marine invertebrates (6) and marine vertebrates (3). See details in Suppl. material 1: Table S1. In our analyses, we focused on the level of species and the three higher taxonomic groups: vertebrates (n = 29 species), invertebrates (16) and plants (15).

Each assessor scored a minimum of three and a maximum of nine species (median = 3) and each species was assessed by a minimum of three and a maximum of eight evaluators (median = 5). Not all assessors evaluated all species of their expertise group; thus, the study design was neither crossed nor nested, an important point in understanding how to measure consistency (see below).

The seven protocols used were developed or applied in Europe: European Plant Protection Organisation-Environmental Impact Assessment for plants (EPPO Brunel et al. 2010); EPPO-Prioritisation scheme (EPPO-Prioritisation; Kenis et al. 2012); German-Austrian Black List Information System (GABLIS; Essl et al. 2011); Great Britain Non-native Species Risk Assessment (GB-NNRA; Baker et al. 2008; Mumford et al. 2010); Generic Impact Scoring System (GISS; Nentwig et al. 2010, 2016); Belgian risk screening tools for potentially invasive plants and animals (Harmonia^{+}; D'hondt et al. 2015) and Environmental Impact Classification of Alien Taxa (called at that time GISS IUCN and now EICAT; Blackburn et al. 2014). The selection of protocols does not consider updates that have become available after 2015 (e.g. for EICAT Volery et al. 2020). For details on protocols and the template used, see González-Moreno et al. (2019).

Before filling the spreadsheets, the assessors read the protocol guidelines and asked questions directly to the protocol developers, if needed. To conduct the assessments, experts decided on their own sources of information (i.e. scientific literature, own expertise or alternative sources). The assessors considered Europe as the risk assessment area. We provided the scores provided by each assessor in each impact assessments, i.e. combination of protocol and species, in Suppl. material 2 as an R list object called "list_impact_assessments.RData".

Table 1. Number of questions regarding different types of impacts of invasive species considered by the seven impact assessment protocols considered. Range of levels indicates the minimum and maximum number of available responses for each question of a given protocol. P-V-I = number of plant, vertebrate and invertebrate species evaluated with each protocol. See the questions and their classification in Suppl. material 1: Table S2. For details on protocols, see González-Moreno et al. (2019).

Protocol	Ecosystem	Biodiversity	Species interaction	Hybridi-sation	Economic losses	Human health	Range of levels	P-V-I
EICAT	2	3	4	1	1	0	5-5	15-16-29
EPPO	4	2	1	1	0	0	3-3	15-0-0
EPPO-Prioritisation	1	1	1	0	2	1	3-3	15-0-0
GABLIS	1	2	2	1	1	1	3-4	15-16-29
GB-NNRA	3	2	2	2	1	1	5-5	15-16-29
GISS	1	2	3	1	5	1	6-6	15-16-29
Harmonia+	1	4	6	2	5	3	3-6	15-16-29

Classification of impact types

Even if some protocols assessed all four components of the invasion process: introduction, establishment, spread and impacts, we only evaluated the latter. To evaluate whether consistency in responses systematically varies across impact types, we grouped the questions into six categories: ecosystem processes, biodiversity, species interactions, hybridisation with native species, economic losses and human health (Table 1 and Suppl. material 1: Table S2). These impacts were further grouped into two coarse impact types: environmental (i.e. biodiversity, species interaction, hybridisation, ecosystems) and socio-economic (i.e. economic losses and human heath).

Quantifying consistency

We measured the consistency of responses across assessors with inter-rater reliability metrics, which quantify the proportion of the variance in the scores associated with assessors (Furr 2021). The values range from 0 to 1 intuitively indicating a low or high consistency in the responses, respectively. For instance, a value of 0.8 would indicate that 20% of the variance observed is due to assessor choice (Hallgren 2012). See an overview on inter-rater reliability metrics provided by Hallgren (2012) and Gwet (2014).

Estimation of inter-rater reliability metrics is influenced by the structure of the data (i.e. which assessors evaluated which species; Putka et al. 2008; Koo and Li 2016). As our study design was neither crossed nor nested, we used the coefficient G (Putka et al. 2008). This coefficient G is based on generalisability theory (G-theory; Brennan 2001; Putka et al. 2008), which is focused on disentangling the sources of error using analyses of variance methods (Brennan 2001). To calculate the coefficient G, we first require estimating the variance associated with raters (e.g. assessors) and ratees (e.g. protocol questions). We did it with a mixed model using the identities of the raters and ratees as random variables (Putka et al. 2008). To address our objectives, we calculated

two types of coefficient G: one for the consistency of assessors scoring each question of a protocol for a given species, i.e. the overall consistency in an impact assessment (hereafter $G_{Prot\text{-}Spp}$) and a second coefficient G for the consistency of assessors scoring a given impact (i.e. protocol question) across all species of a given taxonomic group (hereafter $G_{Quest\text{-}Taxon}$). We differentiated between taxonomic groups, because impact knowledge may vary across them. We used the coefficient $G_{Prot\text{-}Spp}$ to provide information on the general consistency in a particular impact assessment (Objective 1), as well as to disentangle the effect of species identity, taxonomic groups and amount of published scientific articles on species impacts in consistency (Objective 2). We used $G_{Quest\text{-}Taxon}$ to disentangle the effect of impact types (Objective 2). In addition, we used both $G_{Prot\text{-}Spp}$ and $G_{Quest\text{-}Taxon}$ in complementary analyses to unravel whether the influence of protocols relies on methodological aspects, such as the number of questions per protocol and of available answers per question or whether the variability could be potentially more associated with linguistic uncertainties (Objective 2). See Table 2 for details.

In the following sections, we explain the calculations of the coefficient $G_{Prot\text{-}Spp}$ and $G_{Quest\text{-}Taxon}$. We advance that some mixed models to estimate the variance associated with raters and ratees had convergence issues (e.g. identifiability and singularity) and failed to calculate some coefficients G. We also explain in different sections the methodological approximations to disentangle the influence of each factor on consistency of scores.

Calculation of coefficient $G_{Prot\text{-}Spp}$

We calculated a $G_{Prot\text{-}Spp}$ for each combination of protocol and species (i.e. an impact assessment). A way to visualise the data required is a two-dimension array, where the columns are the assessors evaluating a given species, the rows the impact questions of a given protocol and the values within the matrix the scores estimated. For each array, we performed a mixed model to extract the variance associated with the assessors and the protocol questions (Putka et al. 2008; see Table 2). Second, following Putka et al. (2008), we used the estimated variances to calculate the coefficient G. See mathematical details of the coefficient G in Putka et al. 2008 and our R code (Suppl. material 2).

Table 2. Interpretation and use of $G_{Prot\text{-}Spp}$ and $G_{Quest\text{-}Taxon}$. Linear mixed models = formulation used to estimate the variances required for the calculation of the coefficients G. The formulation is the one used to run the models with the R function *lmer* of the R package *lme4*.

Metric	Interpretation	Linear mixed models	Use
$G_{Prot\text{-}Spp}$	Level of agreement in each impact assessment. (Protocol-Species combination).	*Scores ~ (1|ID Question) + (1|ID assessor)*	Objective 1: To quantify the general consistency of assessors in impact assessments. Objective 2: To evaluate if the consistency varies with the taxonomic group or species evaluated, the amount of published information on species impacts and the protocol choice or the number of questions per protocol.
$G_{Quest\text{-}Taxon}$	Level of agreement in each question of a given protocol. (Question-Taxonomic group combination)	*Scores ~ (1|ID Species) + (1|ID assessor)*	Objective 2: To evaluate if the consistency varies with the impact types and the number of available responses per protocol question.

In calculating the $G_{Prot\text{-}Spp}$ values of 330 combinations of species and protocols, we found convergence issues in the mixed models for 66 cases, reflecting in 65 cases of singular models. These issues were not systematically related to species (Chi-squared = 58.69, p-value = 0.52; Chi-squared test with Monte Carlo simulations), but were related to specific protocols (Chi-squared = 53.51, p-value < 0.001; specifically, to EPPO Priorisation and GABLIS protocols). We performed our subsequent analyses with the remaining 264 $G_{Prot\text{-}Spp}$ values. However, to ensure that excluding values from models with singularity issues had no effects on our inferences, we also evaluated differences in $G_{Prot\text{-}Spp}$ between taxonomic groups and protocols without removing the 65 values of the singular models (i.e. sensitivity analysis), which showed similar results.

Calculation of coefficient $G_{Quest\text{-}Taxon}$

We calculated $G_{Quest\text{-}Taxon}$ to evaluate the association between different impact types and levels of consistency. As consistency in answering the diverse impact types can vary across taxonomic groups, we calculated a $G_{Quest\text{-}Taxon}$ for each combination of taxonomic group, protocol and question of each protocol. A way to visualise the data required is a two-dimension array, where the columns are the assessors evaluating a given impact question for any species of a given taxonomic group, the rows, the species of a given taxonomic group and the values within the matrix, the scores estimated. Thus, for the same impact question, we have one to three databases depending on whether the impact can be applied to some or all taxonomic groups (i.e. plants, invertebrates and vertebrates; Table 1). For each array, we performed a linear mixed model to extract the variance associated with the assessors and species identity. Later, we used those variances to calculate $G_{Quest\text{-}Taxon}$ (Putka et al. 2008).

In calculating the $G_{Quest\text{-}Taxon}$ values of the 188 combinations of taxonomic groups, protocols and questions, we found convergence issues in the mixed models for 22 cases. These issues were not systematically associated with protocols (Chi-squared = 5.78, p-value = 0.45), neither impact types (six impact types: Chi-squared = 3.21, p-value = 0.65; two higher impact types: Chi-squared = 0.25, p-value = 0.70). As there was no systematic removal of protocols or impact types, unlike $G_{Prot\text{-}Spp}$, we did not perform sensitivity analyses including the values with warnings about singularity. We performed our subsequent analyses with the remaining 166 $G_{Prot\text{-}Quest}$ values: 64 on plant impacts, 59 on invertebrate impacts and 43 on vertebrate impacts.

Generality and extent of consistency in impact assessments

To interpret $G_{Prot\text{-}Spp}$ values, we classified them into three decision-meaningful categories: low, medium and high consistency in impact assessments. We followed Krippendorff (1980), who considered that impact assessments should be discarded for decision-making if G values were lower than 0.67, impact assessments can tentatively be used for decision-making if G values were between 0.67 and 0.80 and impact assessments can definitively be used for decision-making if G values were above 0.80 (low, medium and high, respectively). To provide information on the general consistency in impact assessments, we discussed the relative frequency of these three categories.

Species

Testing for differences in the consistency of scores between species is challenging due to the relative low amount of protocols and, thus, of $G_{Prot\text{-}Spp}$ values per species. The number of available protocols for each vertebrate and invertebrate species is five and seven for plant species (Table 1). Moreover, for some species, the number of $G_{Prot\text{-}Spp}$ values was lower due to convergence issues (see Table S3 for $G_{Prot\text{-}Spp}$ values estimated). Therefore, we conducted two complementary approximations to test expectations of the influence of species from different perspectives. We called these analyses: permutation test and descriptive analysis. The permutation test is a statistical analyses focused on the proportion of low consistent assessments, while the descriptive analyses is focused on the distribution of raw values.

In the permutation test, we statistically tested if low consistent assessments were associated with few specific species. If true, the number of observed species with a large proportion of low consistent assessments ($G_{Prot\text{-}Spp} < 0.67$) should be lower than those expected by chance. We focused on the proportion of low consistent assessments, instead of using the correlation with all $G_{Prot\text{-}Spp}$ values, since that is the subset challenging the reliability and usefulness of impact assessments. To test it, we performed 1,000 permutations swapping the $G_{Prot\text{-}Spp}$ between species and protocols at random but maintaining the number of $G_{Prot\text{-}Spp}$ values per species and protocol. We later compared, between the observed data and permuted data, the frequency of species with a proportion above 50% of low consistent assessments ($G_{Prot\text{-}Spp} < 0.67$). We looked for statistical differences using the unconditional Boschloo's test with the function *exact.test* of the R package *Exact* (Calhoun 2021). We performed inferences, based on the distribution of the 1,000 p-values. To ensure that our results did not depend on thresholds when calculating the frequency of species with low consistent assessments, we also used the thresholds 30 and 40% to calculate the proportions of low consistent assessments. When sample size is reduced, small variations in the frequency of events have important effects on proportions. We, therefore, conducted the permutation tests with those species with four or more assessments.

In the descriptive analysis, we visually assessed the mean and standard deviations of $G_{Prot\text{-}Spp}$ across species. If consistency depends on species identity, we expect to observe species with different means and non-overlapping standard deviations. Complementarily, large standard deviations (> 0.20), reflecting that the consistency in impact assessments for a same species are in different categories (low, medium and high), support the influence of factors associated with the protocols (e.g. linguistic differences or impact types asked). See Suppl. material 1: Table S4 for a summary of the goals and expectations of all analysis (Permutation test = Target 1; Descriptive analyses = Target 2 in Suppl. material 1: Table S4).

Amount of information available on species impacts

We examined the relationship between the proportions of assessments with low consistency per species ($G_{Prot\text{-}Spp} < 0.67$) with the number of scientific articles on impacts per species recorded in the Web of Science (hereafter correlation test). We expected that the number of articles per species reflects the amount and diversity of knowledge on species impacts and should, therefore, correlate negatively with the proportion of

assessments with low consistency (Target 3 in Suppl. material 1: Table S4). We used a generalised linear model using the Poisson family with the R package (Bates et al. 2015; Wickham et al. 2019; R Core Team 2021). To search the scientific articles, we used the advanced search of ISI Web of Science (11 July 2020). We used a query with three complementary sections. Two sections were fixed and indicated terms for searching (TS) papers about invasive species and their impacts, while the other section indicated synonyms of a species. See the following example: TS = ("*Cameraria oridella*" OR "*Cameraria ohridella*") AND TS = ("Alien" OR "Invasive" OR "Non-native" OR "Non native" OR "Invasion") AND TS = ("Impact" OR "Damage" OR "Harm"). See Suppl. material 1: Table S5 for details on the searches of each species.

Taxonomic groups and protocols

To statistically test whether consistency in assessments varied across taxonomic groups and protocols, we modelled $G_{Prot\text{-}Spp}$ with beta regression models using the R package *glmmTMB* (Brooks et al. 2017; Targets 4 and 5 in Suppl. material 1: Table S4). We modelled both the mean and the precision in the models. While the mean refers to the effect we are interested in, the precision considers a variable dispersion along the explanatory variables (see details in Cribari-Neto and Zeileis 2009; Ferrari et al. 2011; Zhao et al. 2014). When modelling, we also considered that $G_{Prot\text{-}Spp}$ may be influenced by other factors beyond our interest, such as the number of assessors considered in the calculation of the inter-rater metric (Hallgren 2012). Although we did not include the number of protocol questions due to convergence issues, we performed additional analyses to explore the relationship between the variables protocol and number of questions per protocol (see below; Target 6 in Suppl. material 1: Table S4). To model the mean, we used models representing all combinations of three explanatory variables: the taxonomic group to which the species belongs, the protocol used in the impact assessment and the number of assessors who evaluated each species with each protocol. To model the precision, we included all combinations of two variables: the taxonomic group and protocol identity. Additionally, we controlled the non-independency of data by including in all models the species identity as a random intercept. In total, we performed 28 models. See all models in Suppl. material 1: Table S6. For our inferences, we considered the models with $\Delta AICc \leq 4$ (models with an AICc equal or lower than the minimum observed AICc plus 4). See Targets 4 and 5 in Suppl. material 1: Table S4 for a summary of the main and sensitivity analyses.

We interpreted that statistical differences between taxonomic groups reflect diverse epistemic uncertainties across taxa. In contrast, statistical differences between protocols may reflect linguistic uncertainties, but also three other factors: the number of questions per protocol, the number of responses per question or the impact types evaluated in each protocol. To discuss the origin of protocol variability, we jointly interpreted the results of these beta regression models with three complementary analyses: one focused on $G_{Prot\text{-}Spp}$ (number of questions in a protocol) and two focused on $G_{Quest\text{-}Taxon}$ (the number of responses in the questions and the impact type evaluated; see following sections; Targets 6, 8 and 9 in Suppl. material 1: Table S4). We considered that differences in consistency when using different protocols that are not explained by the number of

questions, number of responses per question or the impact types, might support the influence of linguistic uncertainties. In the complementary analyses to quantify the influence of the number of questions per protocol, we repeated the previous 28 beta regression models (Suppl. material 1: Table S6), but exchanging the variable protocol for the variable number of impact questions per protocol. We later compared the marginal pseudo-R^2 associated with the variable protocol and number of questions per protocol (Target 6 in Suppl. material 1: Table S4).

Impact types

To evaluate the influence of impact type, we used $G_{Quest\text{-}Taxon}$, i.e. the metric providing information on the consistency when scoring a given protocol question across the species of a particular taxonomic group (Table 2). For each combination of protocol question and taxonomic taxon, we have its $G_{Quest\text{-}Taxon}$ value and its association with a detailed or coarse impact (see above; Suppl. material 1: Table S2). As some questions fell into several categories of impact types, we controlled this pseudo-replication in subsequent analyses (see below). In total, we analysed 76 $G_{Quest\text{-}Taxon}$ values for plants, 71 $G_{Quest\text{-}Taxon}$ values for invertebrates and 51 $G_{Prot\text{-}Quest}$ values for vertebrates.

We modelled $G_{Quest\text{-}Taxon}$ in relation to impact types and taxonomic groups to consider differences in the knowledge of impact types across taxonomic groups. In the analyses, we controlled four co-variables that can also affect $G_{Quest\text{-}Taxon}$ values: the number of species used to calculate $G_{Quest\text{-}Taxon}$, the number of assessors used to calculate $G_{Quest\text{-}Taxon}$, the protocol to which each question belongs and the specificity of the question (if it asked about one or more types of impact; binomial). In total, we used six variables to study variability in $G_{Quest\text{-}Taxon}$. The number of combinations of our four categorical variables were relatively large for our amount of data (166 $G_{Quest\text{-}Taxon}$ values for 252 combinations of levels; impact type = 6 levels; taxonomic group = 3; protocols = 7 and specificity = 2). To reduce overparametrisation, we conducted two nested models. First, we modelled the variance associated with the four co-variables (two categorical and two continuous variables; hereafter, first nested model). Later, we modelled its residuals with the impact type and taxonomic group (hereafter, second nested model). We avoided overparametrisation, but assigned to the co-variables any potential variance shared with our variables of interest. Therefore, the detected effect of the taxonomic group and impact types may be conservative.

These first nested models were beta regressions since $G_{Quest\text{-}Taxon}$ values ranges from 0 to 1. We modelled $G_{Quest\text{-}Taxon}$ with all combinations of the four co-variables, in the mean and precision parameter. We chose the best model, based on the corrected Akaike's Information Criterion approach (AICc; Target 10 in Suppl. material 1: Table S4). We then extracted its residuals and modelled them with the taxonomic group and impact types by using a linear mixed model. We explored five models: a) interactive effects of the impact types and taxonomic groups; b) additive effects of the impact types and taxonomic groups; c) single effect of impact types; d) single effect of taxonomic groups; and e) null model with just the intercept (Target 11 in Suppl. material 1: Table S4). Since the same question can be answered for multiple taxonomic groups (Table 1), we also included the identity of the question as a random effect. For our inferences,

we considered the models within the ΔAICc ≤ 4. We obtained the proportion of the variance in $G_{Quest\text{-}Taxon}$ explained by the explanatory variables by applying the function *summary* of the R package *base* to the output of the models (R Core Team 2021). See Suppl. material 1: Table S7 for details on models.

To account for pseudo-replication due to the classification of some questions into multiple impact types (Suppl. material 1: Table S2), we repeated the previous steps 1,000 times by choosing in each one a single impact type per question at random. We called these tests randomisation tests. Note the difference with the permutation tests where we swapped $G_{Prot\text{-}Spp}$ (see sensitivity analyses in Targets 10 and 11 in Suppl. material 1: Table S4). To consider the uncertainty in the results, we calculated the proportion of times each of the five models from the second nested model were selected: (i) interactive effect; (ii) additive effect; (iii) single effect of impact type; (iv) single effect of taxonomic group; or (v) just the intercept (Suppl. material 1: Table S7). Later, we calculated the averaged estimated marginal means of the models included in each of the five sets. We conducted these analyses twice, once considering the detailed impact types and another considering the coarse impact types.

Complementarily, we considered that evaluating questions that are not common across the three taxonomic groups limits our ability to quantify the influence of the impact type and taxonomic group. Thus, we also repeated all the previous steps, but using only the common questions across the taxonomic groups (see sensitivity analyses in Targets 10 and 11 in Suppl. material 1: Table S4).

We ran the beta regression models with the R package *glmmTMB* to include random effects (Brooks et al. 2017). We extracted the residuals with the R package *stats* (R Core Team 2021). We performed the linear mixed models with the R package *lme4* (Bates et al. 2015). We evaluated the performance of models by evaluating their residuals with the function *simulateResiduals* of the R package *DHARMa* (Hartig 2020; see its vignette for details). We evaluated the differences between the different variables by studying the estimated marginal means. We calculated the estimated marginal means of its factor levels in all iterations by using the functions *emmeans* and *immeans* of the R package *emmeans*, depending on whether the model has single effects or interactive effects. We performed the Tukey post-hoc test with the function *pairs* of the R package *emmeans* (Lenth 2021). We quantified the proportion of the variance explained by the model as the pseudo-R provided by Johnson (2014).

Factors associated with protocols

We also used $G_{Quest\text{-}Taxon}$ to complement the main analyses on the protocol variable (Target 5 in Suppl. material 1: Table S4). We explored whether the potential signal in the variable protocol can reflect the different impact types evaluated or number of responses in the questions in each protocol (Targets 8 and 9 in Suppl. material 1: Table S4). We calculated the variance partitioning of two sets of beta regressions modelling $G_{Quest\text{-}Taxon}$. In one set of beta regressions, we modelled the mean and the dispersion with the variables protocol and number of responses. In the second set of beta regressions, we modelled the mean and the dispersion with the variables protocol and impact types. In both models, we calculated the pseudo-R^2 of the saturated model and

compared it with the pseudo-R^2 of the models containing only one of the variables. We considered that no shared variance supports the influence of linguistic uncertainties in explaining the consistency in responses between protocols.

Results

Species

The mean $G_{Prot\text{-}Spp}$ was high for 40 out of 60 species ($G_{Prot\text{-}Spp} \geq 0.8$; 19 invertebrates, 12 plants and nine vertebrates), medium for 13 species ($G_{Prot\text{-}Spp} \geq 0.67$ and < 0.8); seven invertebrates, five vertebrates and one plant) and low for seven species ($G_{Prot\text{-}Spp} < 0.67$; three invertebrates, two plants and two vertebrates; Fig. 1). Only in five assessments, assessors scored impacts with a very low consistency ($G_{Prot\text{-}Spp} < 0.3$; *Hydrocotyle verticillata* and *Percnon gibbesi*, both evaluated with GABLIS; *Craspedacusta sowerbii* and *Phasianus colchicus* with GB; and *Solanum elaeagnifolium* with EPPO). See all $G_{Prot\text{-}Spp}$ values in Suppl. material: Table S3. In some cases, $G_{Prot\text{-}Spp}$ varied largely (standard deviations > 0.2). Species with low mean $G_{Prot\text{-}Spp}$ values tended to have larger standard deviations (Spearman correlation between the mean and the standard deviation = -0.82; Fig. 1). However, in general, the standard deviations of the different species overlapped. See Target 2 in Table 3.

The permutation tests showed that the concentration of low consistent assessments ($G_{Prot\text{-}Spp} < 0.67$) could be observed by chance, indicating that assessments with low consistency were not associated with few specific species (Target 1 in Table 3). The p-value of unconditional Boschloo's test was below 0.05 in 0 cases of the 1,000 randomisations, independently of the threshold used to calculate the proportions (30%, 40% and 50%).

Amount of information available on species impacts

The correlation test showed a negative relationship between the proportion of low consistent assessments and the number of published articles on species impact (Estimate = -1.85; Z-value = -14.49; p-value < 0.001). However, the variance explained was low (pseudo-$R^2 \approx 0.05$).

Taxonomic groups and protocols

From the 28 beta regression models used to evaluate the influence of the taxonomic group or the protocols, we identified three best models (Suppl. material 1: Table S8). We focused our results on the model with the variables protocol and taxonomic group because it is the simpler and included our two variables of interest. Nevertheless, the results of the common variable protocol were similar to those of the best models (Suppl. material 1: Tables S9 and S10).

The analyses of the residuals showed no significant deviations from uniformity and homogeneity assumptions for the variable taxonomic group (Kolmogorov-Smirnov test: D = 0.10, p-value = 0.30; uniformity test of each level had a p-value > 0.08; Levene's

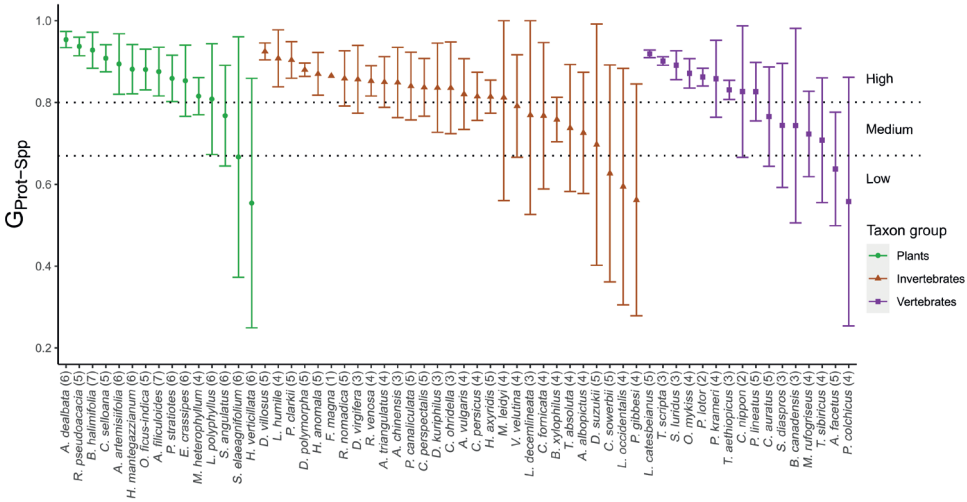

Figure 1. Mean ± standard deviations of the degree of assessor consistency when scoring the impacts of the same species across different protocols ($G_{Prot\text{-}Spp}$). The colours represent different taxonomic groups (green = plants, brown = invertebrates, purple = vertebrates). The number of protocols used to assess each species is indicated between brackets. See complete names of species in Suppl. material 1: Table S3.

test for homogeneity of variance: F value = 0.14, p-value = 0.87) or the variable protocol (Kolmogorov-Smirnov test: D = 0.14, p-value = 0.30; uniformity test of each level had a p-value > 0.20; Levene's test for homogeneity of variance: F value = 0.85, p-value = 0.54). The variable protocol explained greater variance in $G_{Prot\text{-}Spp}$ than the taxonomic group (marginal pseudo-$R^2 \approx 0.10$ and ≈ 0.03, respectively). See Targets 4 and 5 in Table 3.

Assessors tended to score plant impacts with high consistency, while invertebrate and vertebrate impacts were moderately consistent, although confidence intervals overlap with G = 0.80 (Fig. 2A). There were statistical differences between plants and vertebrates (Estimate = -0.551, SE = 0.174, p-value = 0.005) and plants and invertebrates (Estimate = -0.422, SE = 0.155, p-value = 0.019), but not between vertebrates and invertebrates (Estimate = -0.129, SE = 0.164, p-value = 0.711). For the protocols, assessors tended to score impacts highly and consistently when using Harmonia⁺, GISS and EICAT protocols, moderately with GB, moderately-low with GABLIS and low consistently with EPPO. Consistency when using EPPO prioritisation, a protocol that only considered three questions on impacts and with many singularities issues when estimating $G_{Prot\text{-}Spp}$, was highly variable (Fig. 2B; see statistical differences between pairs of protocols in Suppl. material 1: Table S9; Tukey post-hoc test).

The sensitivity analysis, i.e. a repetition of the beta regressions, but also including the $G_{Prot\text{-}Spp}$ values from the mixed models with a warning about singularity, showed greater differences between the levels of the variables protocol and taxonomic group (Suppl. material 3: Fig. S1). However, uniformity and homoscedasticity assumptions were violated.

On the other hand, our complementary analysis to evaluate whether the variable protocol reflected variations in the number of questions per protocol (Target 6 in Suppl. Material 1: Table S4), showed that a model including the variable number of

Table 3. Summary of the main results. Target = Factor evaluated. See details on hypotheses and expectations in Suppl. material 1: Table S4.

Target	Analyses	Result	Interpretation
1) Species	Permutation test	The frequency of species with large proportions of low-consistent assessments can be obtained by chance.	There is no evidence that low-consistent assessments are associated with particular species and, thus, no evidence of clear epistemic uncertainty on species.
2) Species	Descriptive analyses	Visually, the standard deviations overlap across species.	There are no differences in the consistency of responses when assessing different species.
3) Species	Correlation test	Negative correlation between the number of published articles and the proportion of low-consistent assessments. The pseudo-R^2 was low (pseudo-$R^2 \approx 0.05$).	The number of published articles is of little relevance for explaining differences observed.
4) Taxon group	Beta regression	Consistency evaluating plants tended to be larger than when evaluating vertebrates and invertebrates. However, variance explained is small (pseudo-$R^2 \approx 0.03$).	Factors associated with taxonomic groups (e.g. epistemic uncertainties) are not relevant to explain the consistency in assessments.
5) Protocol	Beta regression	Consistency in assessments varied when using different protocols. The protocol explained a low, but relevant 10% of the variance.	Factors associated with protocols are partly relevant to explain the consistency in assessments.
6) Protocol (number of questions per protocol)	Beta regression	The number of protocol questions explains half as much variance as the protocol variable.	Factors associated with protocols are important to some extent. However, some relevance of the protocols is unrelated to the number of questions per protocol (e.g. linguistic uncertainties; see complementary analyses in Targets 8 and 9).
7) Protocol	Descriptive analyses	Some species showed large standard deviations	Factors associated with protocols are important for the impact assessments of some species.
8) Protocol (number of responses per question)	Beta regression	Small variance shared between the number of response questions and the protocol.	The signal observed in protocol (target 5) is not due to number of responses per question and could be caused by linguistic uncertainties.
9) Protocol (Impact type)	Beta regression	Small variance shared between the impact types and the protocol.	The signal observed in protocol (target 5) is not due to the impact types asked in each protocol and could be caused by linguistic uncertainties.
10) Impact types	Beta regression (Nested 1)	Not interesting result. Analysis to avoid overparameterisation. See results on nested linear models 2 (Target 11).	
11) Impact types	Linear model (Nested 2)	As for the coarse impacts, the 1,000 iterations selected as the best model is the one including just the intercept. As for the detailed impacts, only 12.7% of the models showed a statistical signal on impact types. In those cases, impact type explained ≈ 10% of the variance. Sensitivity analyses When using only the common questions for the three taxonomic groups, there is no signal on impact types.	Impact type partly explains the variance in consistency. However, the disappearance of the signal when using the common questions to the three taxonomic groups, suggests the importance of questions specific for each taxon.

questions was worse ($AICc_{questions}$ -387.21 Vs $AICc_{Protocol}$ -416.53). In addition, the marginal pseudo-R^2 of the model including the number of questions was approximately half of the model including the protocol.

Impact types

Our analyses found no statistical differences in $G_{Quest\text{-}Taxon}$ between questions on the coarser impacts (i.e. environmental vs. socio-economic). However, when focusing

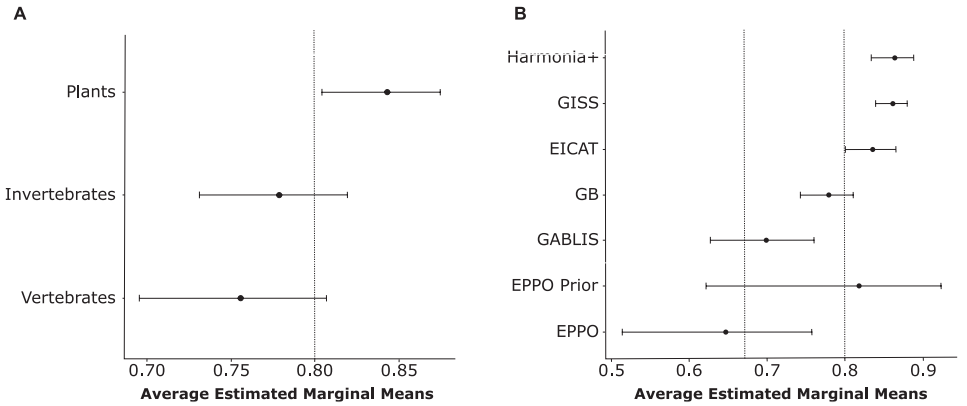

Figure 2. Estimated inter-rater reliability ($G_{Prot\text{-}Spp}$ values) when scoring species belonging to different taxonomic groups (**A**) or using different protocols (**B**). Values averaged over the levels of the variable taxonomic group and protocol, **A** and **B**, respectively, included in the beta regression model (i.e. average estimated marginal means). The dot depicts the mean and the brackets the confidence level at 95%. X-axis values apply the R function *emmeans* with type 'response'. The vertical dotted lines represent the thresholds used to categorise the coefficients G as low, medium and high consistent.

on the detailed impacts, there were no statistical differences in 87.3% of the 1,000 randomisations, i.e. the best model included just the intercept, but there were some differences in the remaining 12.7%. In this reduced subset of models, the consensus of average estimated marginal means showed that assessors most consistently scored questions about impacts on ecosystems and human health and least consistently scored questions about hybridisation and biological interaction amongst species (Fig. 3; see consensus Tukey posthoc-test in Suppl. material 1: Table S11). The single effect of the impact types explained on average 11.4% of the variance in $G_{Prot\text{-}Quest}$. In our sensitivity analyses using only the common questions amongst the three taxonomic groups, there were neither statistical differences between taxonomic groups nor impact types at the coarse and detailed levels. See Targets 10 and 11 in Table 3.

Our complementary analyses to unravel if the signal about the protocol reflected differences in the number of responses per question or the impact types asked in each protocol, showed that the variable protocol shared an irrelevant variance with the variables number of responses per protocol question or the impact types asked (see variance partitioning in Suppl. material 1: Table S12; see Targets 8 and 9 in Table 3).

For similarity with results on $G_{Prot\text{-}Spp}$, we indicated which questions had the highest and lowest consistency ($G_{Quest\text{-}taxon}$). The questions with the highest consistency ($G_{Quest\text{-}taxon} > 0.80$) belonged to protocols Harmonia⁺ (20 combinations of questions and taxonomic group), GB (20), GISS (20), EICAT (10), GABLIS (4) and EPPO (1); while those with the lowest consistency ($G_{Quest\text{-}taxon} < 0.30$) belonged to protocols Harmonia⁺ (8), EICAT (2) and GABLIS (2). See the complete list of $G_{Prot\text{-}Spp}$ and $G_{Quest\text{-}taxon}$ values in Suppl. material 1: Tables S3 and S13.

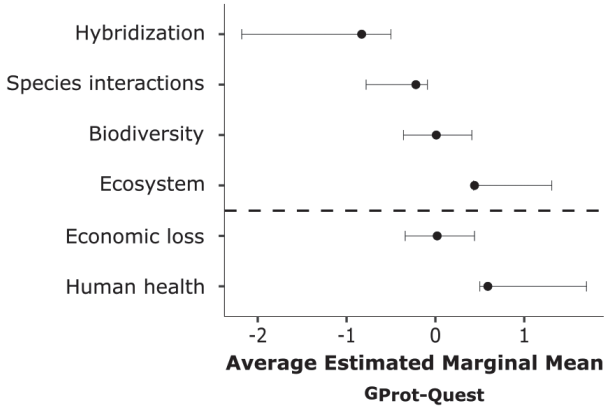

Figure 3. Assessor consistency when scoring different impact types. Results from the 12.7% of the 1,000 randomisations, i.e. models including only the single effect of the detailed impact types as explanatory variable, when using the dataset including all protocol questions on impact ($G_{Quest\text{-}Taxon}$). The unit of the x-axis is residuals; note that these estimates are from a model using the residuals of a previous model as dependent variable. The dot depicts the mean and the brackets the confidence level at 95%. See consensus Tukey adhoc-test in Suppl. material 1: Table S11.

Discussion

We provide the first empirical overview of the consistency amongst assessors in scoring particular questions of invasive species impacts in risk assessment. The broad coverage of this study (60 species from three major taxonomic groups and seven protocols) makes our results highly generalisable, while the focus on particular questions, beyond final scores and rankings, provided accurate estimates of the importance of the assessor in risk assessment, as well as evidence on the importance of the drivers, such as the impact types evaluated. In summary, this study provides new and essential information on one of the many sides of the complex prism that is repeatability in impact assessments.

Our most important finding is that assessor consistency was generally high, with up to 67% of the species studied showing high consistency. Thus, it is reasonable to conclude that impact assessments are largely reproducible and reliable. Our results both support and contrast with those of the limited number of existing studies on the consistency of assessments protocols at the answer level (Volery et al. 2021 and Clarke et al. 2021, respectively). However, comparisons are difficult because of the focus of previous studies on a single protocol (EICAT) and taxonomic group (Volery et al. 2021 = alien ungulates; Clarke et al. 2021 = insects), as well as because of the differences in the number of assessors involved or in the guidelines used (Volery et al. 2021 = similar number; Clarke et al. 2021 = two assessors). Another important point is that the methods for calculating consistency vary and the criteria for considering responses as high or low consistent were not explicit as here. Therefore, to move forward with confidence in this field of knowledge, we call for an intuitive and general criterion for measuring the consistency

of impact assessments, such as the inter-rater reliability metric, as well as to set standards for the values at which consistency is considered high enough to underpin management.

No species had all its assessments with low consistency and the number of species with a large proportion of low-consistent assessments could have been caused by chance (Targets 1 and 2 in Table 3). This lack of support for the importance of epistemic uncertainties may contrast a priori with the observed negative correlation between the number of published articles on species impacts and the proportion of low-consistent assessments in those species or by the different consistency of assessors scoring impacts of the diverse taxonomic groups (Targets 3 and 4 in Table 3). However, the variance explained by both was very low. Thus, although the invasive species analysed here are not a random subset of all alien species, but arbitrarily selected, epistemic factors associated with particular species and taxonomic groups may be less relevant than expected (Leung et al. 2012; McGeoch et al. 2012).

As for impact types, a small fraction of our nested randomised models (12.7%) suggested that assessors scored questions on ecosystem and human health impacts more consistently than questions on hybridisation and biological interactions with native species (Target 11 in Table 3). These results may be surprising as previous studies have shown how scientific evidence for plant impacts on species is greater and more consistent than for ecosystems (Vilà et al. 2011). Our results also support the fact that, although information on economic impacts is sometimes relatively detailed or more readily available than on ecological ones (e.g. Pimentel et al. 2005; Vilà et al. 2010; Roberts et al. 2018; Diagne et al. 2020), the consistency when answering impacts may not be one of the highest due to the also frequent knowledge gaps (McLaughlan et al. 2014; Renault et al. 2021) and context dependency (Haubrock et al. 2021). Human health impact questions showed the highest consistency, which might be related to the well-known health impact of certain species (e.g. hay fever and disease transmission; Mazza and Tricarico 2018). However, these inferences must be taken with care as most of the nested randomised models (87.3%) did not show statistical differences amongst impact types (i.e. the best second nested model included just the intercept). Moreover, the complete disappearance of the signal in the impact types when considering only the common questions across the three taxonomic groups (sensitivity analyses) can also support that variability in consistency can depend on impacts associated with particular taxa. Therefore, these results can highlight the need for quantitative species-specific evidence (Hulme et al. 2013) and for evaluating the degree of confidence on taxon-specific tools (Glamuzina et al. 2017).

As for protocols, our results support previous studies observing high consistency in assessments using the Harmonia^{+}, GISS and EICAT protocols (Essl et al. 2011; Kenis et al. 2012; Turbé et al. 2017; Volery et al. 2021), while EPPO and GABLIS protocols showed less consistency (Target 5 in Table 3). Our complementary analyses to discern the source of the variability associated with the protocols showed that a relative important part of the variance associated with protocols was not explained by the number of questions per protocol, the number of responses per question or the impact types asked in each protocol (Targets 6, 8 and 9 in Table 3). Potentially, the ability of some protocols to consider knowledge gaps in their responses can partly explain differences in

consistency when using alternative protocols (a hypothesis that we did not explore statistically). However, if that is the case, the protocols GABLIS and GISS should have the highest consistency, as they are the only ones considering the response "unknown impact". While this is true for GISS, we, however, observed the contrary result for GABLIS. Thus, our results open the door to the possibility that some variability associated with protocols may be due to linguistic factors, such as the form of the question and language clarity (Turbé et al. 2017; White et al. 2019; Clarke et al. 2021). Although our analyses provide some insights into the role of linguistic uncertainties for consistency, their unravelling would require multidisciplinary collaboration (between ecologists and sociologists). In the meantime, our results call into question whether uncertainty in the alien species lists is almost exclusively epistemic (McGeoch et al. 2012) and support the view that there is still room for improvement of protocols and guidelines (Hawkins et al. 2015; Kumschick et al. 2017; Sandvik et al. 2019; Volery et al. 2020).

Despite the commented differences when scoring different impact types or when using diverse protocols, we note that most impact assessments were highly consistent and that no single factor explained variance to a large extent, important points to prioritise efforts against invasive species. The lack of a clear major factor may suggest that the variability in consistency may be due to different causes and that increasing consistency requires multiple and complementary approaches. To explore this possibility, we conducted additional visual and non-statistical inspections of the nature of the disagreements amongst assessors of the raw data. We observed that the reason of inconsistencies in $G_{Prot\text{-}Spp}$ were diverse, such as the awareness of impacts (e.g. unknown vs. known impacts; GABLIS protocol) or the severity (e.g. low vs. medium in EPPO and GB protocols). Similarly, we observed that low consistencies in $G_{Quest\text{-}taxon}$ were due to assessors disagreeing on the impact severity (e.g. EICAT), the strength of evidence (e.g. "yes" vs. "evidence-based assumption"; GABLIS), or applying the guidelines wrongly (e.g. inapplicable vs. low; Harmonia+). These observations, not shown here, support that the lack of consistency can be due to multiple factors already commented upon in literature (McGeoch et al. 2012; Turbé et al. 2017; White et al. 2019; Probert et al. 2020; Clarke et al. 2021).

Although addressing this question adequately requires analyses beyond the goal of our study, the consistency in scores may be increased by following recommendations from literature. At the assessors group level, it may be promoted by the organisation of iteration-consensus meetings amongst assessors within taxa and across taxa (e.g. horizon scanning; Roy et al. 2014; Gallardo et al. 2016), the use of the same information (Volery et al. 2020), the use of working groups and of peer review panels with clear feedback between assessors and reviewers (Burgman et al. 2011; D'hondt et al. 2015; Vanderhoeven et al. 2017; Volery et al. 2021). At the assessor level, information gathered from scientific literature can be requested to support scores (Vanderhoeven et al. 2017; Vilà et al. 2019) or promote the training of assessors (González-Moreno et al. 2019), an aspect not considered in our dataset, but currently done in some assessments (e.g. EICAT). At the protocol level, it would be desirable to provide clear explanations and guidelines on the information requested for scoring impacts (D'hondt et al. 2015; Turbé et al. 2017; Vilà et al. 2019; Volery et al. 2020), to foment closed-ended questions and improve their wording to avoid

ambiguity (Turbé et al. 2017; Vilà et al. 2019) and, at the level of the information used, to foment studies without the presence of confounding factors and with details on data quality and type of the impact observation (see more details in Volery et al. 2020).

In summary, there is still room for improvement in impact assessments and may require multiple and complementary approaches, such as those described above. However, impact assessments are highly consistent and, therefore, reliable to underpin decision-making. This is a positive and hopeful message, since in view of the expected increase in non-native species introductions (Seebens et al. 2021), we will have to prioritise management and tools, such as impact assessments, will play a key role.

Acknowledgements

We appreciate the past collaboration of all participants and funding of the Alien Challenge COST Action. We also acknowledge the constructive comments of the three reviewers that have made our study more robust and easier to interpret. This research was funded through the 2017–2018 Belmont Forum and BIODIVERSA joint call for research proposals, under the BiodivScen ERANet COFUND programme, under the InvasiBES project (biodiversa.org/1423), with the funding organisations Spanish State Research Agency (MCI/AEI/FEDER, UE, PCI2018-092939 to MV and RBM and PCI2018-092986 to BG) and the Swiss National Science Foundation (SNSF grant number 31BD30_184114 to SB). RBM was supported by MICINN through the European Regional Development Fund (SUMHAL, LIFEWATCH-2019-09-CSIC-13, POPE 2014-2020). PGM was supported by a "Juan de la Cierva-Incorporación" contract (MINECO, IJCI-2017-31733) and Plan Propio Universidad de Córdoba 2020. Publication fee was supported by the CSIC Open Access Publication Support Initiative through its Unit of Information Resources for Research (URICI).

References

Almeida D, Ribeiro F, Leunda PM, Vilizzi L, Copp GH (2013) Effectiveness of FISK, an invasiveness screening tool for non-native freshwater fishes, to perform risk identification assessments in the Iberian Peninsula. Risk Analysis 33(8): 1404–1413. https://doi.org/10.1111/risa.12050

Baker R, Black R, Copp GH, Haysom K, Hulme PE, Thomas M, Ellis M (2008) The UK risk assessment scheme for all non-native species. In: Rabitsch W, Essl F, Klingenstein F (Eds) Biological invasions: from ecology to conservation. Institute of Ecology of the TU Berlin, Berlin, 46–57.

Bates D, Mächler M, Bolker B, Walker S (2015) Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software 67(1): 1–48. https://doi.org/10.18637/jss.v067.i01

Bellard C, Cassey P, Blackburn TM (2016) Alien species as a driver of recent extinctions. Biology Letters 12(2): 20150623. https://doi.org/10.1098/rsbl.2015.0623

Bindewald A, Michiels HG, Bauhus J (2020) Risk is in the eye of the assessor: Comparing risk assessments of four non-native tree species in Germany. Forestry. International Journal of Forestry Research 93(4): 519–534. https://doi.org/10.1093/forestry/cpz052

Blackburn TM, Essl F, Evans T, Hulme PE, Jeschke JM, Kühn I, Kumschick S, Marková Z, Mrugała A, Nentwig W, Pergl J, Pyšek P, Rabitsch W, Ricciardi A, Richardson DM, Sendek A, Vilà M, Wilson JRU, Winter M, Genovesi P, Bacher S (2014) A unified classification of alien species based on the magnitude of their environmental impacts. PLoS Biology 12(5): e1001850. https://doi.org/10.1371/journal.pbio.1001850

Booy O, Robertson PA, Moore N, Ward J, Roy HE, Adriaens T, Shaw R, Van Valkenburg J, Wyn G, Bertolino S, Blight O, Branquart E, Brundu G, Caffrey J, Capizzi D, Casaer J, De Clerck O, Coughlan NE, Davis E, Dick JTA, Essl F, Fried G, Genovesi P, González-Moreno P, Huysentruyt F, Jenkins SR, Kerckhof F, Lucy FE, Nentwig W, Newman J, Rabitsch W, Roy S, Starfinger U, Stebbing PD, Stuyck J, Sutton-Croft M, Tricarico E, Vanderhoeven S, Verreycken H, Mill AC (2020) Using structured eradication feasibility assessment to prioritize the management of new and emerging invasive alien species in Europe. Global Change Biology 26(11): 6235–6250. https://doi.org/10.1111/gcb.15280

Brennan RL (2001) Generalizability Theory. Springer New York, NY. https://doi.org/10.1007/978-1-4757-3456-0

Brooks ME, Kristensen K, Van Benthem KJ, Magnusson A, Berg CW, Nielsen A, Skaug HJ, Machler M, Bolker BM (2017) glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. The R Journal 9(2): 378–400. https://doi.org/10.32614/RJ-2017-066

Brunel S, Branquart E, Fried G, Van Valkenburg J, Brundu G, Starfinger U, Buholzer S, Uludag A, Joseffson M, Baker R (2010) The EPPO prioritization process for invasive alien plants. Bulletin OEPP. EPPO Bulletin. European and Mediterranean Plant Protection Organisation 40(3): 407–422. https://doi.org/10.1111/j.1365-2338.2010.02423.x

Burgman MA, McBride M, Ashton R, Speirs-Bridge A, Flander L, Wintle B, Fidler F, Rumpff L, Twardy C (2011) Expert status and performance. PLoS ONE 6(7): e22998. https://doi.org/10.1371/journal.pone.0022998

Calhoun P (2021). Exact: Unconditional Exact Test. R package version 3.1. https://CRAN.R-project.org/package=Exact

Chapman D, Purse BV, Roy HE, Bullock JM (2017) Global trade networks determine the distribution of invasive non-native species. Global Ecology and Biogeography 26(8): 907–917. https://doi.org/10.1111/geb.12599

Clarke DA, Palmer DJ, McGrannachan C, Burgess TI, Chown SL, Clarke RH, Kumschick S, Lach L, Liebhold AM, Roy HE, Saunders ME, Yeates DK, Zalucki MP, McGeoch MA (2021) Options for reducing uncertainty in impact classification for alien species. Ecosphere 12(4): e03461. https://doi.org/10.1002/ecs2.3461

Cribari-Neto F, Zeileis A (2009) Beta Regression in R. Research Report Series. Department of Statistics and Mathematics, 98. https://doi.org/10.18637/jss.v034.i02

D'hondt B, Vanderhoeven S, Roelandt S, Mayer F, Versteirt V, Adriaens T, Ducheyne E, San Martin G, Grégoire J-C, Stiers I, Quoilin S, Cigar J, Heughebaert A, Branquart E (2015) Harmonia+ and Pandora+: Risk screening tools for potentially invasive plants, animals and their pathogens. Biological Invasions 17(6): 1869–1883. https://doi.org/10.1007/s10530-015-0843-1

Diagne C, Catford JA, Essl F, Nuñez MA, Courchamp F (2020) What are the economic costs of biological invasions? A complex topic requiring international and interdisciplinary expertise. NeoBiota 63: 25–37. https://doi.org/10.3897/neobiota.63.55260

Dorrough J, Oliver I, Wall J (2018) Consensus when experts disagree: A priority list of invasive alien plant species that reduce ecological restoration success. Management of Biological Invasions 9(3): 329–341. https://doi.org/10.3391/mbi.2018.9.3.15

Early R, Bradley BA, Dukes JS, Lawler JJ, Olden JD, Blumenthal DM, Gonzalez P, Grosholz ED, Ibañez I, Miller LP, Sorte CJB, Tatem AJ (2016) Global threats from invasive alien species in the twenty-first century and national response capacities. Nature Communications 7(1): 1–9. https://doi.org/10.1038/ncomms12485

Essl F, Nehring S, Klingenstein F, Milasowszky N, Nowack C, Rabitsch W (2011) Review of risk assessment systems of IAS in Europe and introducing the German–Austrian Black List Information System (GABLIS). Journal for Nature Conservation 19(6): 339–350. https://doi.org/10.1016/j.jnc.2011.08.005

Ferrari SL, Espinheira PL, Cribari-Neto F (2011) Diagnostic tools in beta regression with varying dispersion. Statistica Neerlandica 65(3): 337–351. https://doi.org/10.1111/j.1467-9574.2011.00488.x

Furr RM (2021) Psychometrics: an introduction. SAGE publications, Thousand Oaks, California.

Gallardo B, Zieritz A, Adriaens T, Bellard C, Boets P, Britton JR, Newman JR, van Valkenburg JLCH, Aldridge DC (2016) Trans-national horizon scanning for invasive non-native species: A case study in western Europe. Biological Invasions 18(1): 17–30. https://doi.org/10.1007/s10530-015-0986-0

Genovesi P, Shine C (2004) European strategy on invasive alien species: Convention on the Conservation of European Wildlife and Habitats (Bern Convention). Council of Europe.

Glamuzina B, Tutman P, Nikolić V, Vidović Z, Pavličević J, Vilizzi L, Copp GH, Simonović P (2017) Comparison of taxon-specific and taxon-generic risk screening tools to identify potentially invasive non-native fishes in the River Neretva Catchment (Bosnia and Herzegovina and Croatia). River Research and Applications 33(5): 670–679. https://doi.org/10.1002/rra.3124

González-Moreno P, Lazzaro L, Vilà M, Preda C, Adriaens T, Bacher S, Brundu G, Copp GH, Essl F, García-Berthou E, Katsanevakis S, Moen TL, Lucy FE, Nentwig W, Roy HE, Srėbalienė G, Talgø V, Vanderhoeven S, Andjelković A, Arbačiauskas K, Auger-Rozenberg M-A, Bae M-J, Bariche M, Boets P, Boieiro M, Borges PA, Canning-Clode J, Cardigos F, Chartosia N, Cottier-Cook EJ, Crocetta F, D'hondt B, Foggi B, Follak S, Gallardo B, Gammelmo Ø, Giakoumi S, Giuliani C, Guillaume F, Jelaska LŠ, Jeschke JM, Jover M, Juárez-Escario A, Kalogirou S, Kočić A, Kytinou E, Laverty C, Lozano V, Maceda-Veiga A, Marchante E, Marchante H, Martinou AF, Meyer S, Minchin D, Montero-Castaño A, Morais MC, Morales-Rodriguez C, Muhthassim N, Nagy ZÁ, Ogris N, Onen H, Pergl J, Puntila R, Rabitsch W, Ramburn TT, Rego C, Reichenbach F, Romeralo C, Saul W-C, Schrader G, Sheehan R, Simonović P, Skolka M, Soares AO, Sundheim L, Tarkan AS, Tomov R, Tricarico E, Tsiamis K, Uludağ A, van Valkenburg J, Verreycken H, Vettraino AM, Vilar L, Wiig Ø, Witzell J, Zanetta A, Kenis M (2019) Consistency of impact assessment protocols for non-native species. NeoBiota 44: 1–25. https://doi.org/10.3897/neobiota.44.31650

Gwet KL (2014) Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters. Advanced Analytics, LLC.

Hallgren KA (2012) Computing inter-rater reliability for observational data: An overview and tutorial. Tutorials in Quantitative Methods for Psychology 8(1): 23–34. https://doi.org/10.20982/tqmp.08.1.p023

Hartig F (2020) DHARMa: residual diagnostics for hierarchical regression models. Compr. R Arch. Netw.

Haubrock PJ, Turbelin AJ, Cuthbert RN, Novoa A, Taylor NG, Angulo E, Ballesteros-Mejia L, Bodey TW, Capinha C, Diagne C, Essl F, Golivets M, Kirichenko N, Kourantidou M, Leroy B, Renault D, Verbrugge L, Courchamp F (2021) Economic costs of invasive alien species across Europe. NeoBiota 67: 153–190. https://doi.org/10.3897/neobiota.67.58196

Hawkins CL, Bacher S, Essl F, Hulme PE, Jeschke JM, Kühn I, Kumschick S, Nentwig W, Pergl J, Pyšek P, Rabitsch W, Richardson DM, Vilà M, Wilson JRU, Genovesi P, Blackburn TM (2015) Framework and guidelines for implementing the proposed IUCN Environmental Impact Classification for Alien Taxa (EICAT). Diversity & Distributions 21(11): 1360–1363. https://doi.org/10.1111/ddi.12379

Hulme PE, Pyšek P, Jarošík V, Pergl J, Schaffner U, Vila M (2013) Bias and error in understanding plant invasion impacts. Trends in Ecology & Evolution 28(4): 212–218. https://doi.org/10.1016/j.tree.2012.10.010

Johnson PC (2014) Extension of Nakagawa & Schielzeth's R2GLMM to random slopes models. Methods in Ecology and Evolution 5(9): 944–946. https://doi.org/10.1111/2041-210X.12225

Keller RP, Perrings C (2011) International policy options for reducing the environmental impacts of invasive species. Bioscience 61(12): 1005–1012. https://doi.org/10.1525/bio.2011.61.12.10

Keller RP, Lodge DM, Finnoff DC (2007) Risk assessment for invasive species produces net bioeconomic benefits. Proceedings of the National Academy of Sciences of the United States of America 104(1): 203–207. https://doi.org/10.1073/pnas.0605787104

Kenis M, Bacher S, Baker RHA, Branquart E, Brunel S, Holt J, Hulme PE, MacLeod A, Pergl J, Petter F, Pyšek P, Schrader G, Sissons A, Starfinger U, Schaffner U (2012) New protocols to assess the environmental impact of pests in the EPPO decision-support scheme for pest risk analysis. Bulletin OEPP. EPPO Bulletin. European and Mediterranean Plant Protection Organisation 42(1): 21–27. https://doi.org/10.1111/j.1365-2338.2012.02527.x

Koo TK, Li MY (2016) A guideline of selecting and reporting intraclass correlation coefficients for reliability research. Journal of Chiropractic Medicine 15(2): 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Krippendorff K (1980) Validity in content analysis. In: Mochmann E (Eds) Computerstrategien fur die kommunikationsanalyse, 69–112.

Kumschick S, Vimercati G, De Villiers FA, Mokhatla MM, Davies SJ, Thorp CJ, Rebelo AD, Measey GJ (2017) Impact assessment with different scoring tools: How well do alien amphibian assessments match? NeoBiota 33: 53–66. https://doi.org/10.3897/neobiota.33.10376

Latombe G, Canavan S, Hirsch H, Hui C, Kumschick S, Nsikani MM, Potgieter LJ, Robinson TB, Saul W-C, Turner SC, Wilson JRU, Yannelli FA, Richardson DM (2019) A four-component classification of uncertainties in biological invasions: Implications for management. Ecosphere 10(4): e02669. https://doi.org/10.1002/ecs2.2669

Lawson LL, Hill JE, Hardin S, Vilizzi L, Copp GH (2015) Evaluation of the fish invasiveness screening kit (FISK v2) for peninsular Florida. Management of Biological Invasions 6(4): 413–422. https://doi.org/10.3391/mbi.2015.6.4.09

Lenth RV (2021) Estimated marginal means, aka least-squares means [R Package Emmeans Version 1.6. 0]. Comprehensive R Archive Network (CRAN).

Leung B, Roura-Pascual N, Bacher S, Heikkilä J, Brotons L, Burgman MA, Dehnen-Schmutz K, Essl F, Hulme PE, Richardson DM, Sol D, Vilà M (2012) TEASIng apart alien species risk assessments: A framework for best practices. Ecology Letters 15(12): 1475–1493. https://doi.org/10.1111/ele.12003

Matthews J, van der Velde G, Collas FP, de Hoop L, Koopman KR, Hendriks AJ, Leuven RS (2017) Inconsistencies in the risk classification of alien species and implications for risk assessment in the European Union. Ecosphere 8(6): e01832. https://doi.org/10.1002/ecs2.1832

Mazza G, Tricarico E (2018) Invasive species and human health. CABI, 210 pp. https://doi.org/10.1079/9781786390981.0000

McGeoch MA, Spear D, Kleynhans EJ, Marais E (2012) Uncertainty in invasive alien species listing. Ecological Applications 22(3): 959–971. https://doi.org/10.1890/11-1252.1

McGeoch MA, Genovesi P, Bellingham PJ, Costello MJ, McGrannachan C, Sheppard A (2016) Prioritizing species, pathways, and sites to achieve conservation targets for biological invasion. Biological Invasions 18(2): 299–314. https://doi.org/10.1007/s10530-015-1013-1

McLaughlan C, Gallardo B, Aldridge DC (2014) How complete is our knowledge of the ecosystem services impacts of Europe's top 10 invasive species? Acta Oecologica 54: 119–130. https://doi.org/10.1016/j.actao.2013.03.005

Mumford JD, Booy O, Baker RHA, Rees M, Copp GH, Black K, Holt J, Leach AW, Hartley M (2010) Invasive non-native species risk assessment in Great Britain. Aspects of Applied Biology: 49–54.

Nentwig W, Kühnel E, Bacher S (2010) A generic impact-scoring system applied to alien mammals in Europe. Conservation Biology 24(1): 302–311. https://doi.org/10.1111/j.1523-1739.2009.01289.x

Nentwig W, Bacher S, Pyšek P, Vilà M, Kumschick S (2016) The generic impact scoring system (GISS): A standardized tool to quantify the impacts of alien species. Environmental Monitoring and Assessment 188(5): 315. https://doi.org/10.1007/s10661-016-5321-4

Perdikaris C, Koutsikos N, Vardakas L, Kommatas D, Simonović P, Paschos I, Detsis V, Vilizzi L, Copp GH (2016) Risk screening of non-native, translocated and traded aquarium freshwater fishes in Greece using Fish Invasiveness Screening Kit. Fisheries Management and Ecology 23(1): 32–43. https://doi.org/10.1111/fme.12149

Pimentel D, Zuniga R, Morrison D (2005) Update on the environmental and economic costs associated with alien-invasive species in the United States. Ecological Economics 52(3): 273–288. https://doi.org/10.1016/j.ecolecon.2004.10.002

Probert AF, Volery L, Kumschick S, Vimercati G, Bacher S (2020) Understanding uncertainty in the Impact Classification for Alien Taxa (ICAT) assessments. NeoBiota 62: 387–405. https://doi.org/10.3897/neobiota.62.52010

Putka DJ, Le H, McCloy RA, Diaz T (2008) Ill-structured measurement designs in organizational research: Implications for estimating interrater reliability. Journal of Applied Psychology 93(5): 959–981. https://doi.org/10.1037/0021-9010.93.5.959

Pyšek P, Hulme PE, Simberloff D, Bacher S, Blackburn TM, Carlton JT, Dawson W, Essl F, Foxcroft LC, Genovesi P, Jeschke JM, Kühn I, Liebhold AM, Mandrak NE, Meyerson LA, Pauchard A, Pergl J, Roy HE, Seebens H, Kleunen M, Vilà M, Wingfield MJ, Richardson DM (2020) Scientists' warning on invasive alien species. Biological Reviews of the Cambridge Philosophical Society 95(6): 1511–1534. https://doi.org/10.1111/brv.12627

R Core Team (2021) R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna. https://www.R-project.org/

Regan HM, Colyvan M, Burgman MA (2002) A taxonomy and treatment of uncertainty for ecology and conservation biology. Ecological Applications 12(2): 618–628. https://doi.org/10.1890/1051-0761(2002)012[0618:ATATOU]2.0.CO;2

Renault D, Manfrini E, Leroy B, Diagne C, Ballesteros-Mejia L, Angulo E, Courchamp F (2021) Biological invasions in France: Alarming costs and even more alarming knowledge gaps. NeoBiota 67: 191–224. https://doi.org/10.3897/neobiota.67.59134

Roberts M, Cresswell W, Hanley N (2018) Prioritising invasive species control actions: Evaluating effectiveness, costs, willingness to pay and social acceptance. Ecological Economics 152: 1–8. https://doi.org/10.1016/j.ecolecon.2018.05.027

Roy HE, Peyton J, Aldridge DC, Bantock T, Blackburn TM, Britton R, Clark P, Cook E, Dehnen-Schmutz K, Dines T, Dobson M, Edwards F, Harrower C, Harvey MC, Minchin D, Noble DG, Parrott D, Pocock MJO, Preston CD, Roy S, Salisbury A, Schönrogge K, Sewell J, Shaw RH, Stebbing P, Stewart AJA, Walker KJ (2014) Horizon scanning for invasive alien species with the potential to threaten biodiversity in Great Britain. Global Change Biology 20(12): 3859–3871. https://doi.org/10.1111/gcb.12603

Sandvik H, Hilmo O, Finstad AG, Hegre H, Moen TL, Rafoss T, Skarpaas O, Elven R, Sandmark H, Gederaas L (2019) Generic ecological impact assessment of alien species (GEIAA): The third generation of assessments in Norway. Biological Invasions 21(9): 2803–2810. https://doi.org/10.1007/s10530-019-02033-6

Sardain A, Sardain E, Leung B (2019) Global forecasts of shipping traffic and biological invasions to 2050. Nature Sustainability 2(4): 274–282. https://doi.org/10.1038/s41893-019-0245-y

Seebens H, Essl F, Dawson W, Fuentes N, Moser D, Pergl J, Pyšek P, van Kleunen M, Weber E, Winter M, Blasius B (2015) Global trade will accelerate plant invasions in emerging economies under climate change. Global Change Biology 21(11): 4128–4140. https://doi.org/10.1111/gcb.13021

Seebens H, Blackburn TM, Dyer EE, Genovesi P, Hulme PE, Jeschke JM, Pagad S, Pyšek P, Winter M, Arianoutsou M, Bacher S, Blasius B, Brundu G, Capinha C, Celesti-Grapow L, Dawson W, Dullinger S, Fuentes N, Jäger H, Kartesz J, Kenis M, Kreft H, Kühn I, Lenzner B, Liebhold A, Mosena A, Moser D, Nishino M, Pearman D, Pergl J, Rabitsch W, Rojas-Sandoval J, Roques A, Rorke S, Rossinelli S, Roy HE, Scalera R, Schindler S, Štajerová K,

Tokarska-Guzik B, van Kleunen M, Walker K, Weigelt P, Yamanaka T, Essl F (2017) No saturation in the accumulation of alien species worldwide. Nature Communications 8(1): 14435. https://doi.org/10.1038/ncomms14435

Seebens H, Bacher S, Blackburn TM, Capinha C, Dawson W, Dullinger S, Genovesi P, Hulme PE, Kleunen M, Kühn I, Jeschke JM, Lenzner B, Liebhold AM, Pattison Z, Pergl J, Pyšek P, Winter M, Essl F (2021) Projecting the continental accumulation of alien species through to 2050. Global Change Biology 27(5): 970–982. https://doi.org/10.1111/gcb.15333

Smith K (2020) The IUCN Red List and invasive alien species: an analysis of impacts on threatened species and extinctions. IUCN.

Sohrabi S, Pergl J, Pyšek P, Foxcroft LC, Gherekhloo J (2021) Quantifying the potential impact of alien plants of Iran using the Generic Impact Scoring System (GISS) and Environmental Impact Classification for Alien Taxa (EICAT). Biological Invasions 23(8): 1–15. https://doi.org/10.1007/s10530-021-02515-6

Turbé A, Strubbe D, Mori E, Carrete M, Chiron F, Clergeau P, González-Moreno P, Le Louarn M, Luna A, Menchetti M, Nentwig W, Pârâu LG, Postigo J-L, Rabitsch W, Senar JC, Tollington S, Vanderhoeven S, Weiserbs A, Shwartz A (2017) Assessing the assessments: Evaluation of four impact assessment protocols for invasive alien species. Diversity & Distributions 23(3): 297–307. https://doi.org/10.1111/ddi.12528

Turbelin AJ, Malamud BD, Francis RA (2017) Mapping the global state of invasive alien species: Patterns of invasion and policy responses. Global Ecology and Biogeography 26(1): 78–92. https://doi.org/10.1111/geb.12517

Vanderhoeven S, Branquart E, Casaer J, D'hondt B, Hulme PE, Shwartz A, Strubbe D, Turbé A, Verreycken H, Adriaens T (2017) Beyond protocols: Improving the reliability of expert-based risk analysis underpinning invasive species policies. Biological Invasions 19(9): 2507–2517. https://doi.org/10.1007/s10530-017-1434-0

Vilà M, Basnou C, Pyšek P, Josefsson M, Genovesi P, Gollasch S, Nentwig W, Olenin S, Roques A, Roy D, Hulme PE (2010) How well do we understand the impacts of alien species on ecosystem services? A pan-European, cross-taxa assessment. Frontiers in Ecology and the Environment 8(3): 135–144. https://doi.org/10.1890/080083

Vilà M, Espinar JL, Hejda M, Hulme PE, Jarošík V, Maron JL, Pergl J, Schaffner U, Sun Y, Pyšek P (2011) Ecological impacts of invasive alien plants: A meta-analysis of their effects on species, communities and ecosystems. Ecology Letters 14(7): 702–708. https://doi.org/10.1111/j.1461-0248.2011.01628.x

Vilà M, Gallardo B, Preda C, García-Berthou E, Essl F, Kenis M, Roy HE, González-Moreno P (2019) A review of impact assessment protocols of non-native plants. Biological Invasions 21(3): 709–723. https://doi.org/10.1007/s10530-018-1872-3

Vilizzi L, Copp GH, Adamovich B, Almeida D, Chan J, Davison PI, Dembski S, Ekmekçi FG, Ferincz Á, Forneck SC, Hill JE, Kim J-E, Koutsikos N, Leuven RSEW, Luna SA, Magalhães F, Marr SM, Mendoza R, Mourão CF, Neal JW, Onikura N, Perdikaris C, Piria M, Poulet N, Puntila R, Range IL, Simonović P, Ribeiro F, Tarkan AS, Troca DFA, Vardakas L, Verreycken H, Vintsek L, Weyl OLF, Yeo DCJ, Zeng Y (2019) A global review and meta-analysis of applications of the freshwater Fish Invasiveness Screening Kit. Reviews in Fish Biology and Fisheries 29(3): 529–568. https://doi.org/10.1007/s11160-019-09562-2

Vilizzi L, Copp GH, Hill JE, Adamovich B, Aislabie L, Akin D, Al-Faisal AJ, Almeida D, Azmai MNA, Bakiu R, Bellati A, Bernier R, Bies JM, Bilge G, Branco P, Bui TD, Canning-Clode J, Cardoso Ramos HA, Castellanos-Galindo GA, Castro N, Chaichana R, Chainho P, Chan J, Cunico AM, Curd A, Dangchana P, Dashinov D, Davison PI, de Camargo MP, Dodd JA, Durland Donahou AL, Edsman L, Ekmekçi FG, Elphinstone-Davis J, Erős T, Evangelista C, Fenwick G, Ferincz Á, Ferreira T, Feunteun E, Filiz H, Forneck SC, Gajduchenko HS, Gama Monteiro J, Gestoso I, Giannetto D, Gilles Jr AS, Gizzi F, Glamuzina B, Glamuzina L, Goldsmit J, Gollasch S, Goulletquer P, Grabowska J, Harmer R, Haubrock PJ, He D, Hean JW, Herczeg G, Howland KL, İlhan A, Interesova E, Jakubčinová K, Jelmert A, Johnsen SI, Kakareko T, Kanongdate K, Killi N, Kim J-E, Kırankaya ŞG, Kňazovická D, Kopecký O, Kostov V, Koutsikos N, Kozic S, Kuljanishvili T, Kumar B, Kumar L, Kurita Y, Kurtul I, Lazzaro L, Lee L, Lehtiniemi M, Leonardi G, Leuven RSEW, Li S, Lipinskaya T, Liu F, Lloyd L, Lorenzoni M, Luna SA, Lyons TJ, Magellan K, Malmstrøm M, Marchini A, Marr SM, Masson G, Masson L, McKenzie CH, Memedemin D, Mendoza R, Minchin D, Miossec L, Moghaddas SD, Moshobane MC, Mumladze L, Naddafi R, Najafi-Majd E, Năstase A, Năvodaru I, Neal JW, Nienhuis S, Nimtim M, Nolan ET, Occhipinti-Ambrogi A, Ojaveer H, Olenin S, Olsson K, Onikura N, O'Shaughnessy K, Paganelli D, Parretti P, Patoka J, Pavia Jr RTB, Pellitteri-Rosa D, Pelletier-Rousseau M, Peralta EM, Perdikaris C, Pietraszewski D, Piria M, Pitois S, Pompei L, Poulet N, Preda C, Puntila-Dodd R, Qashqaei AT, Radočaj T, Rahmani H, Raj S, Reeves D, Ristovska M, Rizevsky V, Robertson DR, Robertson P, Ruykys L, Saba AO, Santos JM, Sarı HM, Segurado P, Semenchenko V, Senanan W, Simard N, Simonović P, Skóra ME, Slovák Švolíková K, Smeti E, Šmídová T, Špelić I, Srėbalienė G, Stasolla G, Stebbing P, Ševove B, Suresh VR, Szajbert B, Ta KAT, Tarkan AS, Tempesti J, Therriault TW, Tidbury HJ, Top-Karakuş N, Tricarico E, Troca DFA, Tsiamis K, Tuckett QM, Tutman P, Uyan U, Uzunova E, Vardakas L, Velle G, Verreycken H, Vintsek L, Wei H, Weiperth A, Weyl OLF, Winter ER, Włodarczyk R, Wood LE, Yang R, Yapıcı S, Yeo SSB, Yoğurtçuoğlu B, Yunnie ALE, Zhu Y, Zięba G, Žitňanová K, Clarke S (2021) A global-scale screening of non-native aquatic organisms to identify potentially invasive species under current and future climate conditions. Science of the Total Environment 788: 147868. https://doi.org/10.1016/j.scitotenv.2021.147868

Volery L, Blackburn TM, Bertolino S, Evans T, Genovesi P, Kumschick S, Roy HE, Smith KG, Bacher S (2020) Improving the Environmental Impact Classification for Alien Taxa (EICAT): A summary of revisions to the framework and guidelines. NeoBiota 62: 547–567. https://doi.org/10.3897/neobiota.62.52723

Volery L, Jatavallabhula D, Scillitani L, Bertolino S, Bacher S (2021) Ranking alien species based on their risks of causing environmental impacts: A global assessment of alien ungulates. Global Change Biology 27(5): 1003–1016. https://doi.org/10.1111/gcb.15467

White RL, Strubbe D, Dallimer M, Davies ZG, Davis AJS, Edelaar P, Groombridge J, Jackson HA, Menchetti M, Mori E, Nikolov BP, Pârâu LG, Pečnikar ŽF, Pett TJ, Reino L, Tollington S, Turbé A, Shwartz A (2019) Assessing the ecological and societal impacts of alien parrots in Europe using a transparent and inclusive evidence-mapping scheme. NeoBiota 48: 45–69. https://doi.org/10.3897/neobiota.48.34222

Wickham H, Averick M, Bryan J, Chang W, McGowan L, François R, Grolemund G, Hayes A, Henry L, Hester J, Kuhn M, Pedersen T, Miller E, Bache S, Müller K, Ooms J, Robinson D, Seidel D, Spinu V, Takahashi K, Vaughan D, Wilke C, Woo K, Yutani H (2019) Welcome to the Tidyverse. Journal of Open Source Software 4(43): 1686. https://doi.org/10.21105/joss.01686

Zhao W, Zhang R, Lv Y, Liu J (2014) Variable selection for varying dispersion beta regression model. Journal of Applied Statistics 41(1): 95–108. https://doi.org/10.1080/02664763.2013.830284

Supplementary material 1

Tables S1–S13

Author: Rubén Bernardo-Madrid

Data type: Tables.

Explanation note: **Table S1.** Species evaluated with impact assessments. **Table S2.** Classification of the impact questions into the different impact types. **Table S3.** GProt-Spp per impact assessment. Inter-rater reliability using all impact questions of the protocol. **Table S4.** Summary of the principal and sensitivity analyses performed to study the influence of different factors on the consistency of responses in protocol questions. **Table S5.** Queries used to search scientific articles in Web of Science. **Table S6.** Models used to evaluate the influence of the protocol and taxonomic group in assessor consistency. **Table S7.** Saturated models for the two nested model to unravel the influence of impact types and their potential interaction with the taxonomic groups. **Table S8.** The 10 regression models with the lowest AICc to evaluate the influence of the protocol and the taxonomic groups. **Table S9.** Tukey post-hoc for the variable protocol in the model including the variable taxonomic group. **Table S10.** Tukey post-hoc for the variable protocol in the model including the number of assessors. **Table S11.** Consensus Tukey post-hoc for the variable impact type. **Table S12.** Variance partitioning of the models to unravel the shared variance of the variable protocol with the number of responses per protocol question and impact types. **Table S13.** GProt-Quest per protocol question. Inter-rater reliability per question when considering the impact scores of all species of the same taxonomic group.

Link: https://doi.org/10.3897/neobiota.76.83028.suppl1

Supplementary material 2

Impact assessments and function to calculate G coefficient
Author: Rubén Bernardo-Madrid
Data type: R objects.
Explanation note: An R list object containing the used impact assessments in the study An R function to calculate the coefficient G (inter-rater reliability metric).

Link: https://doi.org/10.3897/neobiota.76.83028.suppl2

Supplementary material 3

Figure S1
Author: Rubén Bernardo-Madrid
Data type: Figure.
Explanation note: Consistency in impact assessments of invasive species is generally high and depends on protocols and impact types.

Link: https://doi.org/10.3897/neobiota.76.83028.suppl3