Lessons learned from testing the Australian weed risk assessment system: the devil is in the details

Daphne A. Onderdonk[A], **Doria R. Gordon**[B], **Alison M. Fox**[C] and **Randall K. Stocker**[C]

[A] Department of Botany, PO Box 118526, University of Florida, Gainesville, Florida 32611, USA.

[B] The Nature Conservancy, Department of Botany, PO Box 118526, University of Florida, Gainesville, Florida 32611, USA.

[C] Department of Agronomy and Centre for Aquatic and Invasive Plants, PO Box 110500, University of Florida, Gainesville, Florida 32611-0500, USA.

Summary

Our test of the Australian Weed Risk Assessment system (WRA) in Florida and comparison of these results with those from tests of the system in other geographies reveal a number of areas where methodological variation may influence results. We demonstrate differences among the tests, such as variability in the base rate of test species and in the evidence required to answer WRA questions. We use the Florida dataset to explore the effects of elements of this variation on accuracy of the WRA and make recommendations to increase consistency in future WRA application and reporting. While we find that the overall accuracy is relatively insensitive to the variation tested, the probability of accurate prediction and comparability of results from different geographies would increase if the system were more consistently and transparently applied.

Keywords: Consistency of WRA tests, invasive species, prediction, screening, variation among WRA tests.

Introduction

The Australian weed risk assessment (WRA) system was developed as a predictive screening tool for distinguishing between non-native species with high and low probability of becoming invasive in a new location (Pheloung *et al.* 1999). This system has now been tested for accuracy in several regions (Hawaii – Daehler and Carino 2000, Hawaii and other Pacific Islands – Daehler *et al.* 2004, Czech Republic – Křivánek and Pyšek 2006, Bonin Islands – Kato *et al.* 2006, Florida – Gordon *et al.* 2008b, Japan – Nishida *et al.* 2009). The existing tests span temperate and tropical and island and continental regions, allowing comparison of the accuracy of the WRA across several geographies. Results of such a comparison (Gordon *et al.* 2008a) suggest that the WRA is a broadly effective screen, with harmful invaders identified with 90% accuracy, on average, and non-invaders identified with 70% accuracy. However, this comparison of results, combined with our experience conducting the test in Florida (Gordon *et al.* 2008b), highlighted several potential inconsistencies in how the WRA system has been implemented and how results are reported. Here, we outline these issues so that future tests and applications of this system can be more consistently conducted. We present several analyses of the Florida data to illustrate many of these issues.

Variation introduced when testing the WRA system

A priori *classification of species*

Evaluation of the accuracy of the WRA requires a retroactive approach: a list of species of known invasiveness in the location of interest is developed and the species are classified into *a priori* categories of invasiveness (e.g. non-invader, minor invader and major invader). Species are then run through the WRA, omitting any data on the impacts of the species in the location of interest, to determine whether the WRA scores correctly predict that non-invaders are unlikely to invade ('accept' outcome), and that invaders are likely to invade ('reject' outcome). Accuracy is generally evaluated as the number of correct acceptances of non-invaders and correct rejections of invaders (Smith *et al.* 1999).

One factor complicating comparison of the various tests of the WRA is that each test has used somewhat different *a priori* categories for the species used to test the accuracy of the WRA. The number of *a priori* categories has varied from two (Daehler and Carino 2000) to four (Křivánek and Pyšek 2006). How species are classified *a priori* influences the evaluation of the success of the WRA test. The broader the definition of 'non-invader' or 'invader', the more false positives (incorrect rejections of non-invaders) or false negatives (incorrect acceptances of invaders), respectively, are likely to result. While definitions of categories at either end of the invasiveness spectrum may be reasonably consistent, the most variation likely lies in the middle category of 'minor invaders'. Not surprisingly, this category had the most variation in outcomes across tests: the percentage of minor invaders predicted not to be invasive ranged from 11% for the Bonin Islands to 45% for the Czech Republic; and that for those predicted to be invasive ranged from 22% for the Czech Republic to 80% for the Bonin Islands (Gordon *et al.* 2008a). It is unclear, however, whether this variation is due to differences in the *a priori* category definition, differences in implementation of the WRA, differences specific to the various geographies, the non-random method used to select species for assessment, or combinations of these sources.

In addition to definitional variation, the method by which species have been identified for the *a priori* categories has varied as well. Varying numbers of experts were asked to rank or categorize species in the tests for Australia (Pheloung *et al.* 1999), Hawaii and other Pacific Islands (Daehler *et al.* 2004), Bonin Islands (Kato *et al.* 2006), and Japan (Nishida *et al.* 2009). The Hawaii test (Daehler and Carino 2000) used invasive species lists to identify invaders, and used recommended lists of species for planting that land managers identified as non-invasive, to identify non-invaders. The Czech Republic test employed previously published information on species' invasive status to classify species, while the Florida test (Gordon *et al.* 2008b) used an existing status assessment tool, published weed lists, and floras of Florida for species categorization. The background of experts has been shown to influence their classification of species: agriculturalists in New Zealand were more likely than conservationists to rate species as harmless, particularly those with high economic value (Pheloung *et al.* 1999). Since determination of the accuracy of the WRA system depends on the accuracy of the *a priori* classification, this classification should be as objective as possible. However, issues regarding *a priori* classification of species are only relevant for tests of the WRA, since novel species screened when the WRA system is implemented require no classification.

Balance of families and life forms across a priori *categories*

Earlier research on patterns in invasive and non-invasive species has demonstrated that the probability of becoming invasive is not independent of plant family or life form (e.g. Daehler 1998). Further, some of these patterns are explicitly included in the WRA system (questions 4.11, 5.01–5.04). However, bias may be unintentionally incorporated into WRA tests if families and life forms are not evenly represented across *a priori* categories. In that case the WRA may merely distinguish between families or forms with invasive or non-invasive tendencies, rather than among species.

Several tests of the WRA system intentionally selected species to include a variety of plant families representing the taxa being introduced into their region (Daehler *et al.* 2004, Kato *et al.* 2006). To control for phylogenetic differences, some tests (e.g. Gordon *et al.* 2008b) explicitly attempted to balance plant families or orders by including as many species as possible of common phylogeny in both the invasive and non-invasive categories. For the Florida test, only three of the 22 families (14%) for which we had at least one *a priori* major invader and non-invader had all species within a family either rejected or accepted. Fifteen of those 22 families had at least one species with a definitive outcome (i.e. did not require further evaluation) in both *a priori* categories (57 species had definitive outcomes in these 15 families). For 12 of those families (80%), all major invaders were rejected, and all non-invaders were accepted. Thus, there was no evidence of a family bias in the WRA outcomes.

Most tests of the WRA included a diversity of plant life forms, but their distribution across *a priori* categories is not usually reported. To balance life forms in the Florida test, major invaders were paired with non-invaders or minor invaders with the same life form wherever possible (Gordon *et al.* 2008b). We used analysis of variance to examine the dependence of WRA score on life form (forb/herb, graminoid, shrub, tree, or vine) across the 158 Florida test species. While there was no overall difference between mean scores ($F_{4, 153}$ = 1.59, P = 0.18), trees and shrubs tended to have lower scores (mean 6.7, s.d. 7.8, and mean 5.7, s.d. 8.6, respectively) than forb/herbs (mean 8.9, s.d. 7.5), graminoids (mean 10.9, s.d. 9.6), and vines (mean 9.3, s.d. 8.0). Similarly, in a comparison of WRA scores across tests of the WRA system, scores from the Czech Republic test, which included only woody species, tended to be lower than scores from other geographies (Gordon *et al.* 2008a). Woody species are less likely to be agricultural weeds and tend to have longer generation times, which can reduce WRA scores (Gordon *et al.* 2008a). Plants with slower generation times may also take longer to become naturalized or invasive; evidence of invasiveness may not be apparent yet as many trees have been introduced in the last 100 years (Caley *et al.* 2008). Because the WRA system treats various life forms differently, the balance of life forms across *a priori* categories should be considered in tests of the WRA system.

Effect of base-rate on accuracy and reliability of results

If families and life forms are balanced across *a priori* categories, the result will be a roughly equal number of species in each category. This breakdown, however, does not reflect the actual relative occurrence of invaders and non-invaders, where an estimated 0.1 to 1% of non-native species introduced become invasive (Williamson and Fitter 1996, Groves *et al.* 2003, Mack 2005). If test species are selected to accurately represent the base-rate (*sensu* Smith *et al.* 1999) of invasive and non-invasive species, then families and life forms cannot be balanced. The Florida (Gordon *et al.* 2008b), Hawaii (Daehler and Carino 2000), and Bonin Islands (Kato *et al.* 2006) tests included roughly the same number of species in each *a priori* category, while Australia (Pheloung *et al.* 1999) and Japan (Nishida *et al.* 2009) used more invaders than non-invaders. The Czech Republic (Křivánek and Pyšek 2006) and the Hawaii and other Pacific Islands (Daehler *et al.* 2004) tests used a greater proportion of non-invaders than invaders, reflecting the pattern (but not magnitude) of the natural base-rate. Since base-rate affects the reliability of predictions (Smith *et al.* 1999), or the proportion of accept or reject decisions that are correct, base-rate should be taken into consideration when comparing tests (Gordon *et al.* 2008a).

We examined how the reliability of the Florida test results would respond to a hypothetical 10-fold increase in the number of non-invaders. When we held constant the accuracy of predictions, or the percentage of invaders or non-invaders correctly identified, the reliability of accept decisions changed from 66% to 95%, and the reliability of reject decisions changed from 96% to 68% (Table 1). Thus, further examination of WRA efficacy with more realistic representation of species across *a priori* categories would be valuable.

Reducing potential assessor bias

If the person conducting the WRA already has knowledge or opinions about the invasiveness of non-native species in their test region, answers to the WRA questions could be subconsciously biased. We attempted to reduce this potential bias in the Florida test by having an assessor who had no prior knowledge of either invasive plants in Florida, or the *a priori* categories in which the test species had been placed (Gordon *et al.* 2008b). The assessors for the Japanese test were similarly unaware of the *a priori* species categorization (Nishida *et al.* 2009). The screenings for the Bonin Islands (Kato *et al.* 2006) and Japanese (Nishida *et al.* 2009) tests were each conducted by at least two or five people, respectively, whose results were averaged to further reduce any bias. We are unaware of whether or how other tests addressed this issue. Because implementation of the WRA system involves screening of species new to a given region, this issue of bias is only relevant to retrospective tests of the WRA system. Reports on tests of the WRA system should ideally include information on the number of people conducting screenings, how bias was minimized, and the consistency of their results.

Geographic source of data

Another issue for tests of the WRA system is the geographic range of the data used to address those questions pertaining to whether a species is considered to

Table 1. Demonstration of the effect of base-rate on reliability with the Florida test data (Gordon *et al.* 2008b). Reliability is the likelihood that an accept or reject decision is correct, while accuracy is the percentage of species in each *a priori* category that is correctly identified. (a) The baseline data, showing number of non-invaders and invaders (minor and major invaders combined) in each of the three WRA outcomes, and (b) the same data with a hypothetical 10-fold increase in non-invaders, maintaining the same level of accuracy, but showing the change in reliability.

a)

		A priori category			
		Non-Invaders	Invaders	Total	Reliability
WRA outcome	Accept	35	18	53	35/53 = 66%
	Evaluate further	9	7	16	
	Reject	4	85	89	85/89 = 96%
	Total	48	110	158	
	Accuracy	35/48 = 73%	85/110 = 77%		

b)

		A priori category			
		Non-invaders	Invaders	Total	Reliability
WRA outcome	Accept	350	18	368	350/368 = 95%
	Evaluate further	90	7	97	
	Reject	40	85	125	85/125 = 68%
	Total	480	110	590	
	Accuracy	350/480 = 73%	85/110 = 77%		

be a weed elsewhere in its introduced range (questions 3.01–3.04). Clearly, all evidence used to address these questions must be from outside the region in which the system is being tested, since the intent of a test is to evaluate the ability of the WRA to predict species' invasiveness in the test region had they been novel introductions. For tests of the WRA conducted on islands, the question of what constitutes 'elsewhere' is straightforward. For non-island tests of the WRA, however, the issue is less clear. 'Elsewhere' could mean immediately outside the political boundaries of the test region (e.g. state or nation), but often these boundaries have little biogeographical meaning. Alternatively, 'elsewhere' could mean outside a biogeographical barrier surrounding the test region, or it could mean outside the continent in which the test region lies.

We examined this issue of the geographic source of data used for answering whether the species is a weed elsewhere using the data from Florida's test of the WRA (Gordon *et al.* 2008b). We recorded the geographic source of evidence at two different scales: outside Florida, and outside North America. We then compared the results to determine the impacts of restricting geography of the source data. While restricting the data sources to areas outside of North America resulted in fewer positive responses for the weed elsewhere questions (lowering the point totals), only 16 out of 158 scores were affected. These score differences changed the WRA outcomes for five species, with neither geographic source of evidence providing consistently more accurate outcomes. After the secondary screen (Daehler *et al.* 2004) was applied, outcomes for only three species differed. Thus, accuracy in WRA tests may be largely independent of how 'elsewhere' is defined when questions 3.01–3.04 are answered.

Variation introduced when testing or applying the WRA system

One of the greatest potential sources of inconsistency between tests or applications of the WRA is exactly how the WRA questions are answered (Gordon *et al.* 2008a, Nishida *et al.* 2009). While some guidance on how to address the questions exists (e.g. Biosecurity Australia http://www.daffa.gov.au/ba/reviews/weeds/system, and Hawaii and Pacific Islands http://www.botany.hawaii.edu/faculty/daehler/wra/screening_criteria.pdf), gaps and inconsistencies remain. We sought to understand how important these inconsistencies might be since even minor changes in how questions are answered have been shown to affect scores and outcomes (Barney and DiTomaso 2008). Much of the inconsistency is being resolved with the development of a consensus set of guidelines (Gordon *et al.* 2010); the discussion below identifies inconsistencies present prior to this more detailed guidance. However, differences in interpretation of questions will remain no matter how thorough the guidance.

Distinguishing between 'no' and 'don't know' responses

Existing guidance focused primarily on the information supporting a positive answer rather than the information supporting a negative answer. Further, the difference between a negative answer and no answer is generally not clarified where there is no supporting evidence for that answer (negative evidence is often not reported even when known). This ambiguity is most important for the 18 questions that receive different scores for 'no' answers than for 'don't know' answers (Pheloung *et al.* 1999).

A lack of evidence may be used to support a negative answer where positive evidence is likely to have been reported in the literature (e.g. whether a species is toxic or a produce contaminant). This approach was adopted in the Florida test (Gordon *et al.* 2008b). Similar guidance for the toxicity questions (questions 4.05 and 4.07) is provided for Hawaii and Pacific Islands tests (see above url), but the evidence required for a negative answer to the question regarding dispersal as a produce contaminant (question 7.03) is not included. For questions such as whether a species forms dense thickets (4.12), it is unclear from existing guidance whether a lack of evidence should result in a 'no' response. In contrast, questions such as whether a species has self-compatible fertilization (6.04) clearly require direct evidence for either a positive or negative answer. While the new guidance on how to address the WRA questions clarifies this issue for all questions (Gordon *et al.* 2010), little information is provided in earlier tests on any rules developed, likely reducing the consistency among tests (Gordon *et al.* 2008a).

Defining the evidence needed to answer WRA questions

Little guidance exists on whether 'yes' or 'no' answers need to be supported by specific evidence, or whether morphological or other features of the species may be used as indicators of those answers. Further, existing guidance generally does not clarify whether a statement not supported by data may be used as evidence for the answer. For example, the Australian guidance for the question of whether propagules are dispersed by birds (7.06) requires evidence of post-dispersal viability of ingested seeds or fruit for a positive answer. However, these data are often not available, so Hawaii's guidance allows for inference based on fruit morphology if no direct evidence on the dispersal of the species exists. Another approach would be to require some evidence that the fruit or seed is ingested by the bird without requiring evidence of post-dispersal viability. The existing guidance also does not specify the circumstances under which a 'no' response should be given to this question.

The existing guidance from both Australia and Hawaii for question 1.01, whether a species is highly domesticated, describes domestication as substantial human selection for at least 20 generations. This definition assumes that human selection would result in genotypes with reduced weediness. However, selection in the horticulture and fruit growing industries can be for traits that increase weediness, such as reduced generation time or increased number of seeds. Horticultural cultivation of *Ardisia crenata* Sims, for example, has resulted in increased production of the attractive red fruit, thus increasing seed production (Kitajima *et al.* 2006). As a result, a two-step approach was adopted in the Florida test (Gordon *et al.* 2008b): if the answer to the domestication question was positive, we then asked whether selection has likely made the species less weedy. Both questions required a positive answer for a 'yes' response to question 1.01. Consistent responses to this question are particularly important since it receives –3 points for a 'yes' answer.

Australia's guidance for question 4.04, whether a species is unpalatable to grazing animals, includes a component regarding whether the plant can be controlled by herbivores. Regardless of whether different WRA tests included browsers in this category, evidence that a species can be controlled by herbivores is not commonly available. As a result, a more relevant definition might require evidence only that a plant species is highly preferred or readily eaten by herbivores. Similarly, existing definitions for question 8.01, whether a species is a prolific seed producer, involve quantitative cut-offs of the number of seeds per unit area that qualify as prolific. Often, however, only qualitative descriptions of high or low seed production are given. Different tests likely varied in whether qualitative information could be substituted for quantitative data.

To assess the impact of responding to the WRA questions with different levels of data required, the Florida test used two versions of data rigor for eight of the questions, one requiring explicit evidence and the other allowing some assumption (Table 2). Comparison of the results showed that WRA scores were generally higher when specific data were required (Table 3), likely because this approach resulted in fewer 'no' answers, which typically have negative scores. However, when species requiring further evaluation are run through the secondary screen (Daehler *et al.* 2004), differences in outcomes between the two versions are minor (Table 3). The

Table 2. Two versions of guidance used in the Florida WRA test (Gordon *et al.* 2008b) for eight of the 49 questions, one allowing for some assumptions to be made and the other requiring explicit evidence. Guidance in regular font is from the Biosecurity Australia WRA website (http://www.daffa.gov.au/ba/reviews/weeds/system); guidance in italics was added for the Florida test.

Question	Some assumptions allowed	Explicit evidence required
4.02 Allelopathic	The plant is documented as a potential suppressor of the growth of other species by chemical (e.g. hormonal) means. Such evidence is rare throughout the whole plant kingdom. *Accept all statements and evidence of allelopathy in the literature. A lack of positive evidence for this question results in a 'no' answer.*	The plant is well documented as a potential suppressor of the growth of other species by chemical (e.g. hormonal) means. Such evidence is rare throughout the whole plant kingdom. *Answer 'yes' only if experimental evidence involving the use of non-concentrated leaf or root leachates (or other natural plant parts or products) exists. Where there is no evidence for this question, or data rely on concentrated extracts, answer 'don't know'. Answer 'no' where literature states the species is not allelopathic.*
4.04 Unpalatable to grazing animals	Consider the plant with respect to where the plant has the potential to grow and if the herbivores present could keep it under control. This trait may be found at any stage during the lifecycle of the plant and/or over periods of the growing season. *Consider all vertebrates, grazing and browsing, wild and domestic. Evidence that the plant is highly palatable, readily eaten, or preferred is sufficient to answer 'no'; evidence that herbivores can keep it under control is not necessary.*	Consider the plant with respect to where the plant has the potential to grow and if the herbivores present could keep it under control. This trait may be found at any stage during the lifecycle of the plant and/or over periods of the growing season. *Consider all vertebrates, grazing and browsing, wild and domestic. A 'no' answer requires evidence that herbivores can control the plant.*
5.03 Nitrogen fixing woody plant	*Assume that all woody members of the family Fabaceae fix nitrogen (unless there is evidence that a particular species does not). Also answer 'yes' for any other woody species that are documented to fix nitrogen. Answer 'no' for all herbaceous species, for woody Fabaceae that are documented not to fix nitrogen, and for woody non-Fabaceae for which there is no evidence of nitrogen fixing (since it is likely that nitrogen fixing in non-legumes will be reported).*	*Include all woody plants that are documented to fix nitrogen (mostly Fabaceae, but include other families as well). Answer 'no' for all herbaceous species, for woody Fabaceae that are documented not to fix nitrogen, and for woody non-Fabaceae for which there is no evidence of nitrogen fixing (since it is likely that nitrogen fixing in non-Fabaceae will be reported). Answer 'don't know' for woody legumes with no evidence regarding nitrogen fixing.*
6.07 Minimum generative time (years)	This is the time from germination to production of viable seed, or the time taken for a vegetatively reproduced plant to duplicate itself. The shorter the time span, the more weedy a plant is likely to be. The score for this trait uses the correlation factor (1 year score 1, 2–3 years score 0, greater than or equal to 4 years score −1). *When there is no specific evidence on time to reproduction, assume herbaceous, fast-growing species reproduce in 1 year or less; herbaceous, slow- or medium-growing species in 2–3 years; woody, fast-growing species in 2–3 years; and woody, slow- or medium-growing in four or more years. Do not make these assumptions, though, for vines or for species that reproduce vegetatively.*	This is the time from germination to production of viable seed, or the time taken for a vegetatively reproduced plant to duplicate itself. The shorter the time span, the more weedy a plant is likely to be. The score for this trait uses the correlation factor (1 year score 1, 2–3 years score 0, greater than or equal to 4 years score −1). *Answer 'don't know' where there is no evidence regarding generative time.*
7.05 Propagules water dispersed	*Evidence that the propagule is carried by and survives in water, or is buoyant, is sufficient for a 'yes' response to this question. Assume 'no' where there is no evidence of buoyancy or water dispersal.*	*For a 'yes' response to this question, use only direct evidence that the propagule is carried by and survives in water long enough to be dispersed. Answer 'no' where there is evidence in the literature that the species is not water dispersed. Where there is no evidence regarding water dispersal, answer 'don't know'.*
7.06 Propagules bird dispersed	Any propagule that may be transported and/or consumed by birds. An example is small red berries with indigestible seeds. *Evidence of bird dispersal is sufficient for a 'yes' response; evidence of post-dispersal viability is not required. Where there is no information on dispersal, assume 'yes' for reasonably sized fleshy fruits. Assume 'no' where there is evidence of wind dispersal or external dispersal. Also assume 'no' for ferns, grasses, and sedges, even if direct evidence is lacking.*	Any propagule that may be transported and/or consumed by birds, and will grow after defecation. An example is small red berries with indigestible seeds. *Evidence of post-dispersal viability is required for a 'yes' response. Assume 'no' for fern species even if direct evidence is lacking. For other species, do not infer based on fruit morphology; use only direct evidence for positive or negative answers.*
7.08 Propagules dispersed by other animals (internally)	The propagules are eaten and dispersed by animals. *Evidence of animal dispersal is sufficient for a 'yes' response; evidence of post-dispersal viability is not required. Where there is no information on dispersal, assume 'yes' for all fleshy fruits. Assume 'no' where there is evidence of wind dispersal or external dispersal. Also assume 'no' for ferns, grasses, and sedges, even if direct evidence is lacking.*	The propagules are eaten by animals, dispersed, and will grow after defecation. *Evidence of post-dispersal viability is required for a 'yes' response. Assume 'no' for fern species even if direct evidence is lacking. For other species, do not infer based on fruit morphology; use only direct evidence for positive or negative answers.*
8.01 Prolific seed production	The level of seed production must be met under natural conditions and applies only to viable seed. For grasses and annual species, this rate should be (>5,000–10,000 $m^{-2}y^{-1}$), for woody species a rate of (>500 $m^{-2}y^{-1}$) would be considered high. Specific data on this attribute may be unavailable; however, an estimate can be made from the seed/plant and the average size of the plant. *Accept quantitative evidence or qualitative descriptions indicating high or low seed production for 'yes' or 'no' answers. Assume 'yes' for fern species.*	The level of seed production must be met under natural conditions and applies only to viable seed. For grasses and annual species, this rate should be (>5,000–10,000 $m^{-2}y^{-1}$), for woody species a rate of (>500 $m^{-2}y^{-1}$) would be considered high. Specific data on this attribute may be unavailable; however, an estimate can be made from the seed/plant and the average size of the plant. *Accept only quantitative evidence for a 'yes' answer unless the species is a fern, for which 'yes' should be assumed. Accept quantitative data or qualitative descriptions of low seed production for a 'no' answer.*

Table 3. Comparison of the percentage of species in each *a priori* category (major invader, minor invader, non-invader) falling into each WRA outcome (accept, evaluate, reject), using two approaches to data required for eight WRA questions (see Table 2), in the Florida test of the WRA (Gordon *et al.* 2008b). Results are shown after the secondary screen (Daehler *et al.* 2004) has been applied.

		Some assumptions allowed			Explicit evidence required		
		A priori category			*A priori* category		
		Major invader	Minor invader	Non-invader	Major invader	Minor invader	Non-invader
WRA Outcome	Accept	2%	36%	73%	2%	27%	71%
	Evaluate further	6%	6%	19%	6%	8%	21%
	Reject	92%	58%	8%	92%	65%	8%
Mean WRA scores		14.9	7.3	0.6	15.4	8.4	1.5

greatest difference can be seen in the minor invader category, with more species rejected and fewer accepted when explicit evidence is required for answers.

This comparison demonstrates that WRA outcomes are generally robust to differing data requirements for answering the WRA questions, particularly when the secondary screen is applied. Furthermore, accuracy has been fairly consistent among the various tests of the WRA system (Gordon *et al.* 2008a), despite likely differences in interpretation of questions. Since differing interpretations can apply to more questions than the eight compared here, and since different screens of the same species can result in different scores and outcomes (Barney and DiTomaso 2008), efforts to increase consistency in addressing the WRA questions (e.g. Gordon *et al.* 2010) should increase ease of both application and interpretation.

Different approaches to score weighting

Two slightly different approaches to the weighting of questions 3.01–3.05 by the climate questions (2.01 and 2.02) were introduced in Australia (Pheloung *et al.* 1999, see Gordon *et al.* 2010). One approach rounds the scores to integer values, while the other assigns some points in 0.5 increments. The systems give the same scores to questions 3.01–3.05, however, if the default climate matching answers are used. We found no consistent differences in scores between the Florida test, which used the system with only integers, and tests that used the alternative version (Gordon *et al.* 2008a). However, those testing or applying the WRA should be aware of the differences between the two approaches, as the integer system can result in slightly higher WRA scores.

Evidence used for climate matching questions

Biosecurity Australia's instructions for answering the WRA questions (http://www.daffa.gov.au/ba/reviews/weeds/system) suggest using a climate matching computer analysis to answer the climate matching questions (2.01 and 2.02). Where this analysis is not possible, users are directed to default to the highest value of 2 for each question. Alternatively, some tests of the WRA have used a qualitative climate match to answer these questions (e.g. Daehler *et al.* 2004). Applications that use the default climate matching answers will potentially have higher scores, since answers to these questions weight the scores for 'yes' answers to questions 3.01–3.05. Although no test's scores were consistently higher or lower than any other test's scores (Gordon *et al.* 2008a), the use of different methods for answering the climate matching questions is another source of variation in testing and implementation of the WRA.

Reporting of WRA results

General differences in reporting

In addition to differences in application of the WRA, the way in which results are reported has also varied among the tests. Often only partial results are reported, leaving the breakdown between all outcomes in each *a priori* category unclear. For example, if the focus of a study is on false positives and negatives, only rejected non-invaders and accepted invaders might be reported. The breakdown between accepted non-invaders and those needing further evaluation, and between rejected invaders and those needing further evaluation, is then unclear. Furthermore, percentages are sometimes reported without the numbers from which they were derived. Since the ability to compare the accuracy of the WRA system across studies is crucial for making policy arguments regarding its implementation, results should be reported as completely and clearly as possible (Gordon *et al.* 2008a).

Number of questions answered

The WRA system is designed so that not all 49 questions need to be completed to generate an outcome for a species put through the screen. A minimum of ten questions is required, with specified distribution through three categories of questions (http://www.daffa.gov.au/ba/reviews/weeds/system). The number of questions answered per species can vary among applications of the WRA depending on the evidence required to answer questions. For example, if a lack of evidence is often taken as a 'no' answer, more questions will be answered than if stronger evidence is required, even with the same data gathered. The Hawaii test (Daehler and Carino 2000) reports a range of 15–45 questions answered per species (mean not given), while Hawaii and other Pacific Islands (Daehler *et al.* 2004) gives a range of 31–44 with a mean of 37. The Czech Republic test (Křivánek and Pyšek 2006) reports a mean number of questions answered of 37 (range not given), and the Florida test (Gordon *et al.* 2008b) had a range of 25–44 questions with a mean of 35. Despite likely differences in evidence required to answer questions, the number of questions answered appears not to have differed greatly among tests reporting these numbers. This result supports the conclusion that the WRA is robust to inconsistencies in application.

Rarely answered questions

Some WRA questions are answered more frequently than others. Questions such as whether or not a species is a grass (5.02) or is aquatic (5.01) will always be answered as long as basic information on taxonomy, growth form, and habit is available. Other questions, such as whether the species has natural enemies present in the region of interest (8.05), will rarely be answered because the data required to answer the question are not generally available (Gordon *et al.* 2010). If the same questions are consistently left unanswered, perhaps they could be eliminated from the WRA system, thus reducing research effort.

In the Florida test of the WRA system, 9 of the 49 WRA questions were answered for ≤30% of the species tested (Table 4). To examine whether these nine questions were integral to the overall scores and outcomes of the test species, we compared the scores and outcomes of our main dataset to those when these nine questions were eliminated. We were still able to answer the minimum number of questions for all 158 species. Without the rarely answered questions, 86 scores were different than with the full set of questions (16 increased and 70 decreased). Prior to application of the secondary screen, the outcomes for six species were different than with the full set of questions; only four outcomes were different after application of the screen. Of these four different outcomes, one improved in accuracy (a non-invader changed from 'reject' to 'accept'), two

decreased in accuracy (a non-invader changed from 'accept' to 'evaluate', and an invader changed from 'reject' to 'accept'), and one change was indeterminate (a minor invader changed from 'reject' to 'evaluate', the accuracy of which depends on whether minor invaders are expected to be accepted or rejected (Gordon *et al.* 2008a)). Thus, eliminating these nine questions did not have a major impact on the outcomes for our 158 test species. Two of these nine questions (1.02 and 1.03) are meant to be answered only when the answer to question 1.01 (Is the species highly domesticated?) is 'yes', so it is not unexpected that they are rarely answered. Other tests did not report on the frequency with which questions were answered, so the generality of these results is unknown. While we would not recommend exclusion of these questions without further investigation, tracking this type of information would provide data that might eventually allow some simplification of the WRA system.

Table 4. Questions answered for ≤30% of the species in the Florida test of the Australian WRA (Gordon *et al.* 2008b).

WRA question	% of species with answer
1.02 Has the species become naturalized where grown?[A]	6
1.03 Does the species have weedy races?[B]	1
2.03 Broad climate suitability (environmental versatility).	27
4.04 Unpalatable to grazing animals.	17
6.01 Evidence of substantial reproductive failure in native habitat.	1
6.03 Hybridizes naturally.	17
7.01 Propagules likely to be dispersed unintentionally.	20
8.04 Tolerates or benefits from mutilation, cultivation, or fire.	30
8.05 Effective natural enemies present in *Florida*.[C]	1

[A] Answered only when the answer to questions 1.01 (Is the species highly domesticated?) is 'yes'.
[B] Answered only when the answer to questions 1.01 (Is the species highly domesticated?) is 'yes', and when the taxon is a sub-species, cultivar, or registered variety.
[C] The italicized portion of this question is modified to address the area of interest (Gordon *et al.* 2010).

Relationship between WRA score and number of questions answered

Ideally, the WRA score for a species should be unrelated to the number of questions answered. However, the Florida test found a positive correlation between these factors ($r^2 = 0.36$, $P < 0.001$), supporting a generally weak correlation found in other tests (Daehler and Carino 2000, Daehler *et al.* 2004, Nishida *et al.* 2009). This relationship likely results because more information is available for invasive species, and more information typically leads to higher scores, for two reasons. First, positive answers usually raise the score, while negative answers often do not change the score (Daehler *et al.* 2004). Second, even where negative answers decrease the score, negative evidence is often not reported, resulting in more positive than negative answers. This explanation is supported by the finding by Kato *et al.* (2006) of a positive correlation between WRA score and number of questions answered only for rejected species. Similarly, the correlation between score and number of questions answered in the Florida data was found to be driven by the invasive species (Figure 1). When we separated the data by *a priori* category, the correlation was highly significant for major invaders ($r^2 = 0.43$, $P < 0.001$), moderately significant for minor invaders ($r^2 = 0.08$, $P = 0.03$), and not significant for non-invaders ($r^2 = 0.00$, $P = 0.37$).

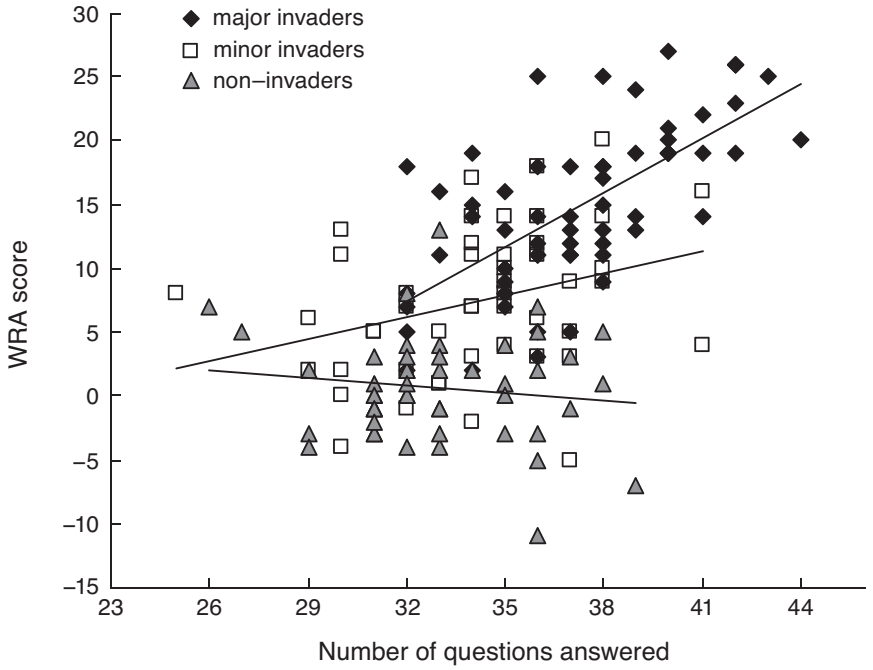

Figure 1. The relationship between WRA score and the number of questions answered for non-invaders (n = 48, $r^2 = 0.00$, $P = 0.37$), minor invaders (n = 48, $r^2 = 0.08$, $P = 0.03$) and major invaders (n = 62, $r^2 = 0.43$, $P < 0.001$) in the Florida test of the WRA (Gordon *et al.* 2008b).

Conclusions

Our analyses of the Florida data presented here and our comparison of WRA tests across geographies (Gordon *et al.* 2008a) demonstrate that the WRA is a robust screening tool, despite the differences in how the system has been applied. However, greater consistency in implementation and reporting of the WRA will strengthen comparisons of tests of the WRA in different geographies, as well as comparisons of the WRA with new methodologies. Further, the WRA system is likely to be adopted for regulatory use in more locations if it has been demonstrated that it can be applied consistently and with little ambiguity. As a result, we suggest that application of this tool be conducted as uniformly and transparently as possible and that researchers and implementers clearly report their methods and results.

Acknowledgments

We are grateful for funding for portions of this project from the Florida Department of Environmental Protection – Bureau of Invasive Plant Management, USDA APHIS – Plant Protection and Quarantine, and the Florida Department of Agriculture and Consumer Services – Division

of Plant Industry. The Florida Chapter of The Nature Conservancy and the Florida Agricultural Experiment Station also supported this work.

References

Barney, J.N. and DiTomaso, J.M. (2008). Nonnative species and bioenergy: are we cultivating the next invader? *BioScience* 58, 64-70.

Caley, P., Groves, R.H. and Barker, R. (2008). Estimating the invasion success of introduced plants. *Diversity and Distributions* 14, 196-203.

Daehler, C.C. (1998). The taxonomic distribution of invasive angiosperm plants: ecological insights and comparison to agricultural weeds. *Biological Invasions* 84, 167-80.

Daehler, C.C. and Carino, D.A. (2000). Predicting invasive plants: prospects for a general screening system based on current regional models. *Biological Invasions* 2, 93-102.

Daehler, C.C., Denslow, J.S., Ansari, S. and Kuo, H. (2004). A risk-assessment system for screening out invasive pest plants from Hawaii and other Pacific islands. *Conservation Biology* 18, 360-8.

Gordon, D.R., Onderdonk, D.A., Fox, A.M. and Stocker, R.K. (2008a). Consistent accuracy of the Australian weed risk assessment across varied geographies. *Diversity and Distributions* 14, 234-42.

Gordon, D.R., Onderdonk, D.A., Fox, A.M., Stocker, R.K. and Gantz, C. (2008b). Predicting invasive plants in Florida using the Australian weed risk assessment. *Invasive Plant Science and Management* 1, 178-95.

Gordon, D.R., Mitterdorfer, B., Pheloung, P.C., Ansari, S., Buddenhagen, C., Chimera, C., Daehler, C.C., Dawson, W., Denslow, J.S., LaRosa, A., Nishida, T., Onderdonk, D.A., Panetta, F.D., Pyšek, P., Randall, R.P., Richardson, D.M., Tshidada, N.J., Virtue, J.G. and Williams, P.A. (2010). Guidance for addressing the Australian Weed Risk Assessment questions. *Plant Protection Quarterly* 25, 56-74.

Groves, R.H., Hosking, J.R., Batianoff, G.N., Cooke, D.A., Cowie, I.D., Johnson, R.W., Keighery, G.J., Lepschi, B.J., Mitchell, A.A., Moerkerk, M., Randall, R.P., Rozefelds, A.C., Walsh, N.G. and Waterhouse, B.M. (2003). 'Weed categories for natural and agricultural ecosystem management' (Bureau of Rural Sciences, Canberra. http://live.greeningaustralia.org.au/nativevegetation/pages/pdf/Authors%20G/5_Groves_et_al.pdf. Accessed 7 July 2008).

Holm, L.G., Pancho, J.V., Herberger, J.P. and Plucknett, D.L. (1979). 'A geographical atlas of world weeds'. (John Wiley and Sons, New York).

Kato, H., Hata, K., Yamamoto, H. and Yoshioka, T. (2006). Effectiveness of the weed risk assessment system for the Bonin Islands. *In* 'Assessment and control of biological invasion risk', eds F. Koike, M.N. Clout, M. Kawamichi, M. De Poorter and K. Iwatsuki, p. 65. (Shoukadoh Book Sellers, Kyoto, Japan and IUCN, Gland, Switzerland).

Kitajima, K., Fox, A.M., Sato, T. and Nagamatsu, D. (2006). Cultivar selection prior to introduction may increase invasiveness: evidence from *Ardisia crenata*. *Biological Invasions* 8, 1471-82.

Křivánek, M. and Pyšek, P. (2006). Predicting invasions by woody species in a temperate zone: a test of three risk assessment schemes in the Czech Republic (Central Europe). *Diversity and Distributions* 12, 319-27.

Mack, R.N. (2005). Predicting the identity of plant invaders: future contributions from horticulture. *HortScience* 40, 1168-74.

Nishida, T., Yamashita, N., Asai, M., Kurokawa, S., Enomoto, T., Pheloung, P.C. and Groves, R.H. (2009). Developing a pre-entry weed risk assessment system for use in Japan. *Biological Invasions* 11, 1319-33.

Pheloung, P.C., Williams, P.A. and Halloy, S.R. (1999). A weed risk assessment model for use as a biosecurity tool evaluating plant introductions. *Journal of Environmental Management* 57, 239-51.

Smith, C.S., Lonsdale, W.M. and Fortune, J. (1999). When to ignore advice: invasion predictions and decision theory. *Biological Invasions* 1, 89-96.

Williamson, M. and Fitter, A. (1996). The varying success of invaders. *Ecology* 77, 1661-5.