Testing the effectiveness of the Forest Integrity Assessment: a field-based tool for estimating the condition of tropical forest

Andrew J. Suggitt [1,2,*], Kok Loong Yeong [3,4], Anders Lindhe[5], Agnes Agama[4], Keith C. Hamer[6], Glen Reynolds[4], Jane K. Hill[2] & Jennifer M. Lucey[7,*].

* To whom correspondence should be addressed. Emails: jennifer.lucey@zoo.ox.ac.uk (J.M.L.), andrew.suggitt@northumbria.ac.uk (A.J.S.).

1 Department of Geography and Environmental Sciences, Northumbria University, Newcastle-upon-Tyne, NE1 8ST, UK.

2 Leverhulme Centre for Anthropocene Biodiversity, Department of Biology, University of York, Wentworth Way, York, YO10 5DD, UK.

3 Leverhulme Centre for Climate Change Mitigation (LC3M), University of Sheffield, Department of Animal and Plant Sciences, Western Bank, Sheffield, S10 2TN, UK.

4 South East Asia Rainforest Research Partnership, Danum Valley Field Centre, 91112, Lahad Datu, Sabah, Malaysia.

5 High Conservation Value Resource Network, Frewin Court, Oxford, OX1 3HZ, UK.

6 School of Biology, University of Leeds, Leeds, LS2 9JT, UK.

7 Department of Zoology, University of Oxford, Mansfield Road, Oxford, OX1 3SZ, UK.

Running title: Testing a rapid field-based forest condition assessment tool.

Article type: Research article.

Keywords: Rapid assessment, tropical ecology, human-modified, forest set-aside, High Carbon Stock, ecological integrity, forest quality.

ABSTRACT

1. Global targets to halt biodiversity losses and mitigate climate change will require protecting rainforest beyond current protected area networks, necessitating responsible forest stewardship from a diverse range of companies, communities and private individuals. Robust assessments of forest condition are critical for successful forest management, but many existing techniques are highly technical, time-consuming, expensive, or require specialist knowledge.
2. To make assessment of tropical forests accessible to a wide range of actors, many of whom may be limited by resources or expertise, the High Conservation Value Resource Network (HCVRN), with the SE Asia Rainforest Research Partnership (SEARRP), developed a South East Asian version of the Forest Integrity Assessment (FIA) tool as a rapid (< 1 hour) method of assessing forest condition in the field, where non-experts respond to 50 questions about characteristics of the local environment while walking a site transect. Here, we examined the effectiveness of this survey tool by conducting ~ 1,000 assessments of forest condition at 16 tropical rainforest sites with varying levels of disturbance in Sabah, Malaysian Borneo.
3. We found good agreement (R-squared range: 0.50 – 0.78) between FIA survey scores and independent measures of forest condition, including biodiversity, vegetation structure, aboveground carbon, and other key metrics of ecosystem function, indicating that the tool performed well. Although there was variation among assessor responses when surveying the same forest sites, assessors were consistent in their ranking of those sites, and prior forest knowledge had a minimal effect on the FIA scores. Revisions or further training for questions where assessors disagree, for example on the presence of fauna at a site, could improve consistency.
4. We conclude that the FIA survey tool is a robust method of assessing forest condition, providing a rapid and accessible means of forest conservation assessment. The FIA tool could be incorporated into management practices in a wide range of forest conservation schemes, from sustainability standards, to community forestry and restoration initiatives. The tool will enable more organisations and individuals to understand the conservation value of the forests they manage, and to identify areas for targeted improvements.

1 INTRODUCTION

Globally, forests are at increasing risk of degradation or conversion to agriculture as the need for food and other resources continues to rise (Hosonuma et al., 2012). Yet, these complex ecosystems support high levels of biodiversity, harbour rare, threatened and endangered species, store and sequester large amounts of carbon, regulate local and global climate systems, and maintain soil, hydrological and other ecosystem services (Watson et al., 2018). Provision of ecosystem services is at its highest where forests are in the best condition, by which we mean that they more closely resemble intact, or primary, habitat (e g. in terms of carbon: Wang et al., 2001; biodiversity: Tawatao et al., 2014; water quality: Luke et al., 2017). Continued assessment and monitoring of forests, coupled with the ongoing maintenance and enhancement of forests to improve their condition, are therefore essential if global targets on biodiversity and climate mitigation are to be met.

To curb continued deforestation and degradation, conservation initiatives incentivise companies, communities and individuals to manage forest areas outside of protected areas. Industry-based certification standards, such as the Roundtable on Sustainable Palm Oil (RSPO) and the Forest Stewardship Council (FSC), require companies to set aside and manage natural forests within their management units. The High Conservation Value (HCV) approach has been widely adopted by certification standards such as the RSPO, and by corporations aiming to conserve biodiversity and meet avoided deforestation commitments (www.hcvnetwork.org), while additional areas are now being set aside by the palm oil, pulp and paper, and cocoa sectors, under the High Carbon Stock (HCS) Approach (www.highcarbonstock.org). Community forest stewardship, for example via REDD+ (Reducing Emissions from Deforestation and Forest Degradation) or ecotourism-based forest restoration schemes, is also becoming a frequent means of achieving forest conservation while also benefiting local livelihoods (Holck et al., 2008; Kunjuraman & Aziz, 2019). The widespread adoption of these schemes creates new opportunities for increased forest protection and improved management, but constraints may exist with respect to the ongoing stewardship of these vital ecosystems – which, in some cases, falls to institutions or individuals that have limited capacity in forest and conservation management.

Existing forest monitoring techniques often require a high level of technical knowledge, and can be both time-consuming and expensive (Gibbs et al., 2007). Over recent years interest has grown in remote sensing (e.g. by using satellites or drones) as a means of delivering forest monitoring (Finer et al., 2018). Whilst these techniques provide information about forest structure and biomass over large areas, they also require technical expertise, and are thus often inaccessible to smaller operations or local communities. Data derived from remotely-sensed imagery can also mask important sub-canopy aspects of conservation value and disturbance (such as hunting or the presence/absence of endangered species, Green et al., 2019). Existing community-based forest monitoring approaches focus largely on counting and measuring trees. While relatively simple, these approaches require a degree of expert knowledge, are time-consuming, and conservation managers can find it challenging to translate outputs into information of practical relevance (Holck et al., 2008). Other tools such as SMART (https://smartconservationtools.org/), focus on monitoring of threats to wildlife. SMART relies on patrols, and has been shown to be effective in enhancing protection of endangered species, but is contingent on continuous on-the-ground patrolling, has far less emphasis on wider ecosystem quality, and is only an option for well-funded and staffed conservation programmes (Hoette et al., 2016; Critchlow et al., 2017). Many of these techniques focus on individual aspects of the forest ecosystem, such as forest structure or specific species, and – as a consequence - encourage forest managers to maintain a narrow focus, rather than a broader view of the whole ecosystem. Ecological integrity assessment methods recognise the need to understand the multiple interacting characteristics that contribute to ecosystem functioning and thus the provisioning of key ecosystem services (Tierney et al., 2009; Wurtzebach & Schultz, 2016). However, these processes can be complex to assess and monitor, thus a more simple, low-cost and rapid forest assessment technique is needed to assess forest condition.

The Forest Integrity Assessment (FIA) tool assesses multiple facets of forest condition, while also addressing issues of time, resource and the need for technical expertise. The tool was developed by the HCVRN, in partnership with SEARRP in the Malaysian context, (https://hcvnetwork.org/library/forest-integrity-assessment-tool/) as a rapid (< 1 hour to complete) means of conducting broad assessments of forest condition via a cheap and efficient approach that does not require expert knowledge or extensive resources. Until now, however, the robustness of scores generated by the tool has not been tested. To enable such a test, we first conducted a large-scale field trial of the survey tool, completing 967 assessor surveys across 16 sites in the dipterocarp rainforests of Sabah, Malaysian Borneo. We use this trial data to test three key performance aspects of the tool relevant to its wider application: 1) how well the scores derived from the FIA survey tool correlate with independent metrics of forest condition, 2) if the characteristics or prior knowledge of assessors affect their scoring, and 3) if the survey is efficient as it can be, i.e. do all questions within the survey tool discriminate between sites in good or poor condition effectively. We conclude by discussing where and how the FIA tool might be deployed as a rapid and low-cost means of assessing forest condition.

2 MATERIALS AND METHODS

2.1 The survey tool

The FIA survey tool has been adapted for different forest types globally (Lindhe & Drakenberg, 2019) and we tested the version developed for lowland dipterocarp rainforest (Lindhe et al., 2015). Until recent clearances, Lowland dipterocarp rainforest was the dominant forest type across South East Asia, and it represents the modal pre-clearance forest type in areas managed by RSPO member oil palm companies. Dipterocarp forest has also been surveyed by a substantial number of independent research projects, and we utilise the resulting datasets in this study (Table 1, see also Methods 2.3 below).

In the FIA questionnaire (hereafter referred to as ‘the survey’, Supplementary Info 1), assessors respond to a series of 50 ‘yes’ or ‘no’ questions regarding the site they are surveying, and the final score is calculated by finding the total number of ‘yes’ responses. The survey targets a range of seven criteria known to be associated with forest condition, namely: landscape, topography, water, trees, flora, fauna and disturbance. Features that are especially important for indicating forest condition, such as the occurrence of large trees, are allocated more questions to provide greater weighting for these important aspects, and all questions are designed to be insensitive to the time of day or year (i.e. wet/dry season, Lindhe & Drackenberg, 2019).

2.2 Collecting survey test data

A total of 62 people was recruited to test the survey. Among this group we recruited a range of levels of expertise in forestry and conservation to test whether survey results were affected by prior knowledge, experience, education or other characteristics. We recruited 14 oil palm plantation staff from estates owned by Wilmar International (RSPO member since 2005), a total of 19 first year undergraduate conservation biology students from Universiti Malaysia Sabah, and 17 field research assistants and 12 graduates and post-graduates affiliated with SEARRP. Information on age, gender, nationality, educational background, profession and prior knowledge or experience of tropical forests was obtained. Assessors could choose to remain anonymous, leave the survey process at any time, or refrain from answering any of the questions if they wished, in line with the ethical approval granted for the study.

Assessors conducted surveys at 16 forest sites in Sabah, Malaysian Borneo (Supplementary Table 1, Supplementary Fig. 1), varying in size and disturbance level from continuous tracts of primary forest in Danum Valley Conservation Area to small degraded and previously logged forest fragments within oil palm plantations (fragment size ranged from 12ha to 120ha). These sites were chosen because: a) they represented a gradient of forest degradation on similar soil types, topography and elevation typical of the region, and b) they fell within the spatial coverage of datasets quantifying forest integrity (see below), to enable comparison with independent measures of forest condition at survey sites. Wherever possible we used the same transect or trail as that used by the original study. On the day of the survey, each assessor was given a set of guidelines to study for 30 minutes (Supplementary Info 2). One of the authors (K.L.Y.) accompanied assessors to the starting point of the survey, and during those 30 minutes was available to answer any questions and provide clarifications. The guidelines and the survey form were available in English and Bahasa Malaysia languages. Once the assessment began, the assessors were asked not to discuss answers with one another or seek any further clarifications. The survey was conducted along a pre-designated 500m transect at each site. Assessors spent a total of one hour walking each transect and considering their answers.

2.3 Test #1: Comparison of survey test scores to independent forest condition metrics

To ascertain the ability of the survey to correctly identify sites of good or poor forest condition, we compared scores from the full survey test dataset with data from independent published studies that were conducted at, or whose spatial coverage overlapped with, our 16 test sites. We used a total of 967 assessor surveys for this test- note that not every assessor visited every site, but nearly all did. The validation data were derived from independent studies, and encompassed a variety of aspects of forest condition or conservation value (Table 1, Supplementary Fig. 2), namely: the species richness of dipterocarp trees (Yeong et al., 2016a) and of ants (Tawatao et al., 2014), aboveground carbon stocks (Asner et al., 2018), and vegetation structure complexity, decomposition rate, and various aspects of ecosystem function in dipterocarp trees (Yeong et al. 2016a, 2016b, *in review*). In almost all cases these independent data were collected on the same forest trails, plots or coordinates as our assessments. The exception was the generation of aboveground carbon estimates (which were provided as a 30m x 30m resolution raster), where the mean value across a circular buffer of diameter 1.5km was extracted using the coordinates for each site. We used these datasets to calculate R-squared statistics between these components of forest condition and the survey scores generated by our assessors (Table 1).

2.4 Test #2: Examining assessor effects on survey scoring

In order to establish if the FIA tool could be rolled out more widely, irrespective of the background and other characteristics of the assessors involved, we needed to understand how survey scoring may have been affected by prior knowledge and expertise. We also needed to assess the extent to which individual assessors scored consistently across sites. We therefore used the test data to examine the effect of these factors on survey scoring using a Generalised Linear Mixed Modelling (GLMM) approach. For these analyses, and to link scoring of particular individuals across sites, we constructed GLMMs using test data from a subset of 34 assessors who: 1) chose not to remain anonymous, and 2) had identified themselves across at least 5 survey sites (n= 34 assessors, n = 493 total assessor surveys). Explanatory variables in models included the highest educational qualification obtained (four categories- 1) ‘primary school’, 2) ‘secondary school’, 3) ‘pre-university’, Diploma or Malaysian Matriculation Programme, and 4) ‘degree’- university undergraduate degree); prior forest knowledge (two self-assessed categories: prior knowledge or no prior knowledge); age (three categories: aged <31, 31-50, or 50<); gender (M/F). Candidate models were constructed using all possible combinations of these variables, with individual assessor identity (assessor 1, assessor 2, etc) specified as a random effect (intercept) in all models. GLMMs were constructed using the ‘lme4’ package in R (Bates et al., 2015) , model selection was performed using AIC (Akaike Information Criterion, Burnham & Anderson, 2002), and the associated marginal and conditional R-squared values were estimated using the Delta method (Bartoń 2020).

2.5 Test #3: Determining survey effectiveness and identifying improvements

We tested two key performance aspects of the survey, namely: 1) survey effectiveness, i.e. how effective the survey questions were at discriminating between sites of better or worse condition (hereafter ‘discriminative ability'), and 2) response consistency, i.e. how consistent different assessors were when surveying the same sites (hereafter 'agreement rate').

To estimate the discriminative ability of questions, we first calculated the mean answer for each of the 50 questions at each site ($0 \leq x \leq 1$), across all assessors. We then calculated the standard deviation in these means across all 16 sites to obtain a discriminative ability score for each question (hereafter referred to as the 'discriminative ability' measure). Using the mean answer for each question at each site ensured that variability associated with the assessors' answers at each site did not contribute towards our measure of the question's ability to discriminate between sites.

To estimate an agreement rate score that accounts for scorers agreeing simply by chance, we calculated Fleiss's kappa score for each survey question using the 'irr' package in R (Gamer et al., 2019). Fleiss's kappa is a dimensionless score from zero to one, where one indicates a high level of agreement, and zero a low level of agreement.

Both these measures were derived from a subset of the full data, wherein 30 responses- the maximum number of responses available for all 16 sites- were randomly chosen and analysed for each of the 16 sites. This was to ensure that each site was represented by the same number of survey responses, i.e. a balanced design, with each site contributing equally to the overall scoring for each aspect of survey performance (30 responses per site x 16 sites = 420 total assessor surveys).

Finally, in the interest of optimising the survey, we tested the effect of removing survey questions where agreement and/or discriminative ability was low, using the balanced subset of 420 assessor surveys. Questions were removed in groups (see results for details of these), and final scores were recalculated, before up-weighting them to the equivalent of a score out of 50, to enable comparisons with scores from the full survey. R-squared statistics were also recalculated to test for any change in agreement with the independent forest condition metrics (see Methods 2.3).

3 RESULTS

3.2 Test #1: Comparison of survey test scores to independent forest condition metrics

Overall, survey scores ranged from 10 to 47 out of 50, indicating that our sample of sites spanned a wide range of forest condition. The survey scores generated by our assessors agreed with the majority indicators of forest condition that we extracted from the independent studies specified in the methods (Fig. 1, Table 1). Foremost among these was a strong positive association with vegetation structure ($R^2 = 0.75$ vs. forest condition score, Yeong et al., 2016a, Fig. 1c). Scores were also positively associated with aboveground carbon stocks ($R^2 = 0.69$ vs. carbon in Mg C per ha, Asner et al., 2018, Fig.1d), although here there were signs that the association could be non-linear. R-squared statistics also showed good agreement between FIA scores and important aspects of ecosystem functioning in forests (decomposition $R^2= 0.74$, Fig. 1e; dipterocarp seedling prevalence $R^2=0.78$, Fig. 1f; Table 1). There was reasonable agreement between scores and the alpha diversity of the forests they were collected in ($R^2 = 0.50$ vs. dipterocarps, 0.54 vs. ants, Fig.1a,b), considering the high variability of measured diversity at these sites (Tawatao et al., 2014, Yeong et al., 2016a). Finally, some aspects of functioning did not seem to be associated with survey score (dipterocarp fruiting, Fig. 1g, dipterocarp seedling survival, Fig. 1h). Overall these results offer good support for the use of the FIA tool as an estimate of the relative condition of forest sites.

3.2 Test #2: Examining assessor effects on survey scoring

We found limited evidence for assessor effects on survey scoring. A GLMM containing the prior knowledge variable (alone) was the 'best' performing model in the set (Burnham & Anderson, 2002), but five other models with different formulations achieved an AIC score within two points of this model, thus also achieving 'substantial' support (Supplementary Table 2). Importantly, the marginal R-squared statistics for these six models (as for all models in the set) were low- max 0.06, minimum 0.03- indicating that assessor characteristics accounted for no more than 6% of the variation in scoring (Supplementary Table 2). The conditional R-squared statistic for the participant identity random effect was 0.18, indicating that 18% of the variation in scoring was due to participant identity. It was therefore likely that individual-level variability between assessors had a stronger influence on scoring than the other characteristics (education, prior knowledge, age or gender).

3.3 Test #3: Determining survey effectiveness and identifying improvements

Discriminative ability of questions varied widely (Fig. 2). There was a large amount of variation in answers to questions on landscape, topography or trees, whereas answers to disturbance questions were more consistent. One question had zero spread of answers: Q4 asking if the fragment was 1ha or above. This question therefore had zero discriminative power across the range of sites we visited. The average Fleiss's kappa score across questions was 0.25, indicating fair agreement between scorers across sites (Fig. 2). The level of agreement for two questions was worse than that expected by chance alone; these were questions relating to the presence of waterfalls (Q14) and off-trail visibility (Q46). Overall, questions with a higher level of agreement tended to also be better at discriminating across sites (Fig. 2).

Removing questions tended to have a negligible effect on the relative ranking of sites based on the average survey score (Fig. 3a), and also had little effect on the value of R-squared statistics calculated against independent metrics of forest condition (Fig. 3b). Removing every question appearing in the lower left-hand panel of Fig. 2 (labelled 'C'), corresponding to all survey questions with below average discriminative ability and below average agreement rate (n = 23 questions), tended to raise scores for sites in good condition, and in two instances this would have changed the ordering of site rankings (by ±1 rank, Fig. 3a). Removing Fig. 2 panel 'C' questions also mildly worsened R-squared scores, by a mean of –0.008 (Fig. 3b). Using only those questions with above average discriminative ability and agreement rate (n = 20 questions, upper right hand panel labelled 'B' in Fig. 2) had a similar but slightly stronger effect on site rankings (Fig. 3a), but again reduced the value of r-squared statistics vs. the forest condition metrics (by a mean of –0.026, relative to the full survey). Removing the 'worst' ten questions (as per assessor agreement) had a negligible effect on site rankings (Fig. 3a) and on R-squared scores (Fig. 3b, mean change: +0.001).

4 DISCUSSION

4.1 Agreement with independent forest condition metrics

The FIA tool was effective in measuring forest condition across our criteria of vegetation structure complexity, biodiversity and ecosystem functioning (Fig. 1). Strong agreement with vegetation structure and aboveground carbon estimates might have been expected, given the emphasis on structure-related questions in the survey, such as the size and number of trees (n = 13 structure questions of 50 in total). This nevertheless demonstrated that assessors were able to accurately identify critical aspects of forest structure that reflect the forest condition and conservation value without spending large amounts of time taking detailed measurements of tree size and identity. The strength of these associations also indicated the potential applicability of this tool to REDD+ schemes and the HCS Approach, which both use carbon and vegetation structure as proxies for forest 'value' (Brofeldt et al., 2014; Rosoman et al., 2017).

Vegetation structure and carbon stocks have been shown to correlate closely with biodiversity in tropical forest habitats (Lindenmayer et al., 2000; Gao et al., 2014; Magnago et al., 2015; Deere et al., 2018), and this was the case among our test sites (Tawatao et al., 2014). For the two biodiversity datasets available at our study sites (dipterocarp tree diversity and ant diversity), the survey responses were reasonably well correlated with these datasets, particularly given that the FIA assessors were not required to count or identify species. We would expect that the survey would perform similarly against many other groups of species that are forest dependent, and hence vary in occurrence in relation to forest condition. We were not able to test the FIA tool against datasets of vertebrate biodiversity, which is often of concern to conservation initiatives, but sections of the survey that ask questions about evidence of human disturbance including hunting, as well as signs or sightings of mammals, provide potentially important insights into threat levels to vertebrates (Green et al., 2019; Brodie et al., 2015).

The functioning of forest ecosystems is critical to their longer term ability to support biodiversity and ecosystem services (Tierney et al., 2009; Wurtzebach & Schultz, 2016). Although assessors were not asked to identify specific forest functions, our survey scores showed good agreement with aspects of ecosystem functioning, including regeneration and decomposition rates, indicating that the characteristics covered in the survey are highly relevant to ecological processes as well as structure and diversity.

It should be noted that the characteristics of a forest that make it of conservation value differ depending on the particular conservation goal, and are ultimately a value judgement. Our aim was to measure forest condition against the assumption that an intact forest ecosystem will provide the widest range of important services. We believe the tool goes some way to both address the complexity of the forest ecosystem and reduce the impact of value judgments that any individual metric might place on what constitutes 'good condition' by including a range of important elements such as vegetation structure, fauna and indicators of human disturbance. Therefore, although a site cannot achieve a perfect score if any of these elements are reduced, a site can still score well based on a range of different criteria. Questions such as those relating to saplings and fruits also indicate potential for the site to recover. It will be useful for managers to scrutinise the elements of the survey that contribute to the score, rather than simply use the total score, in order to understand the site condition in relation to specific conservation goals and to develop effective management. Developing supporting guidance on interpreting scores for subsequent conservation management would therefore be beneficial.

4.2 Usability and consistency

Overall, assessors were consistent at ranking sites by their relative condition. Accuracy and usability are key criteria for the success of the tool in improving the uptake of effective forest monitoring among a wider range of forest stewards with varied backgrounds and levels of expertise. Scoring by individuals had strong internal consistency across sites, and there was no evidence that prior knowledge or experience of forest ecology influenced the ability of assessors to discriminate between sites of different condition. Community forest monitoring for REDD+, which uses more involved vegetation measurement protocols, was found to be similarly reliable when undertaken by non-experts (Holck et al., 2008). The FIA tool enables the inclusion of other forest properties beyond vegetation structure, with a quicker and simpler approach, and without requiring additional expertise or training.

It should however be noted that the use of a subset of assessors who had chosen not to remain anonymous could have introduced some level of bias into the derived scores. For example, those who were comfortable in identifying themselves may have been more confident in their knowledge of forests, and those less confident may have opted to remain anonymous. It may also be the case that the assessors that opted to self-identify were less likely to deviate from what they perceived to be the consensus for the site, i.e. to provide some sort of 'right' answer ('impression management', Drescher et al., 2013). It is difficult to speculate on what the net effect of these biases might be, but we would however highlight that our tests on the full dataset (anonymous or otherwise) showed that assessor characteristics did not affect scoring in any substantial way.

The assessors in our study had the opportunity to ask for clarifications from a scientist before starting the survey. Questions were fully understood, on the whole, and clarifications were requested for unfamiliar technical words like 'ravine', 'ephemeral', or 'cicada chimney'. This could have some effect on surveys conducted without an expert present, however, we would expect that future revisions to the FIA, including simpler wording, further written explanation, or provision of photos or diagrams would largely solve this issue.

Although assessors were consistent in their ranking of sites, our results also show that there is substantial variability in the absolute scores recorded by each assessor at each site, and thus scores for intact forest sometimes overlapped scores for sites likely to have been in poorer condition. Therefore, individual scores cannot currently be used to ascertain a specific level of forest condition. The FIA tool can, however, be used to rank sites, to understand which areas may be of higher or lower conservation value, and to identify suitable targets for restoration activities. It may also be possible to use the FIA to monitor the condition of a site over time to understand if degradation or recovery is taking place, although more research is needed to test the sensitivity of the tool to changes over time, and to determine the period of time over which changes may become detectable. To make these sorts of ranks and comparisons, it is important to either use the same assessor- as indicated by our results, which show strong internal consistency in survey responses- or perhaps to generate an average of scores across a number of assessors for each site or time period. It should be noted that we did not test the consistency of individual assessors in scoring the same site, therefore the ability of the tool to accurately detect changes over time is unclear.

Reducing variability in responses between assessors is important if we wish to compare the site scores recorded by different individuals, and to increase our confidence in assigning scores to levels of forest condition (e.g. that a score of > 35 indicates primary forest). To understand whether there were aspects of the survey that could be improved to reduce this variability, we investigated the discriminative ability and level of agreement of the 50 survey questions. For questions where there was a low level of agreement, additional explanations, graphics or training could reduce variability without compromising the range of aspects of forest condition considered by the survey. Pre-survey calibration between assessors could also be employed to improve agreement further. Limited discriminative ability may have been caused by characteristics that are either too common at all sites, or too rare to be detected even at the most intact forest sites. One option might be to remove these questions altogether, but this would have negligible effect on overall efficiency (Fig. 3), or on the overall time spent surveying a site. Furthermore, although some questions may not have been important across the range of sites we tested, they could become important at other sites- particularly those sites of very good or very poor condition.

4.3 Applications

The FIA tool is able to rank the condition of forest areas based on our key criteria. For this reason, the tool is competitive alongside alternative methods such as tree enumeration, remote sensing, or wildlife monitoring, because it is able to capture vital information from across all these aspects in a fraction of the cost or time, and without the need for technical expertise. These attributes would enable projects to rapidly assess forests on the ground in a way that allows for engagement and participation by the wider community. Our results also indicate a continuing need for field surveys, even with full access to remote sensing data, because a number of aspects of forest condition for which measurement via remote sensing is either difficult or impossible- such as signs of leeches (Q41), epiphytes (Q33), and sub-canopy tree metrics (Q15-25)- scored highly for discriminative ability at our sites (Fig. 2).

The FIA tool may have applications for any project which requires information about forest condition. These could include eco-tourism restoration initiatives or monitoring of conservation set-asides, such as those in RSPO certified oil palm plantations. Some conservation initiatives require specific information on particular aspects of forest condition, such as aboveground carbon stocks (e.g.in carbon accounting schemes) or the abundance of particular focal species of conservation concern (e.g. orang-utan conservation programmes). The FIA tool is no substitute for the detailed and focused measurements required for these sorts of projects.

It may, however, complement or supplement such measurements, enabling field staff and local communities to cheaply, efficiently, and systematically capture information about the wider condition of the forest which could be pertinent to the focus of the project. For example, it may provide information about vegetation degradation that may affect habitat for focal species, or identify human disturbance that could impact on carbon stocks in the future. We argue that regularly monitoring the full range of aspects of forest condition, through the use of the FIA, could therefore contribute towards conservation goals, as well having wider non-target benefits.

5 CONCLUSIONS

1. The FIA tool is effective in ranking sites in terms of condition, but variation among assessors means that it is important that the same individual is used to conduct comparison of sites over space or time. Alternatively, taking a mean score across multiple assessments of the same site is likely to improve the robustness of condition estimates.
2. More information and training would enhance the accuracy of the survey. Some common sources of inaccuracy could be mitigated by the provision of photos and other visual aids to help understand survey questions. However, to maximise uptake it is important to balance the need for improved accuracy with the need for the survey to remain quick, cost-effective and accessible for non-experts.
3. The tool was shown to be effective in discriminating among forests of varying condition, but we did not test whether it is sufficiently sensitive to detect changes over time, or how repeatable scores are by individuals for the same site, which are important factors for monitoring purposes. More testing is needed to understand whether it can be used to monitor restoration projects, for example, and if so, the requisite frequency and intensity of surveying that would be required.
4. While its simplicity may not provide the detail needed for focussed conservation projects, the FIA tool provides a robust and systematic means of monitoring forest set-asides, providing rapid monitoring data that are accessible to a wide range of potential users.

Authors' contributions

J.M.L., G.R. & A.L. conceived research, J.M.L, A.J.S., J.K.H, K.H. & A.A. designed research;, A.J.S., K.L.Y. & J.M.L. performed research; A.J.S., K.L.Y. & J.M.L. analysed data; J.M.L. & A.J.S. led the writing of the paper, and all authors provided comments.

Acknowledgements

Thanks to Wilmar International, Yayasan Sabah, Sabah Forestry Department, the Danum Valley Management Committee, Sabah Biodiversity Council and Borneo Rainforest Lodge for site access and work permissions. Thanks also to SEARRP and Wilmar International for logistical support, and to the survey volunteers. All statistical analyses were conducted on the R statistical platform (version 3.4.4), and we thank the authors of this software and associated base packages for making this platform available and free to use.

Data availability statement

All supporting data will be archived via Dryad (http://datadryad.org/) and a DOI will be supplied upon publication.

Competing interests

The authors have no competing interests.

ORCID

Andrew J. Suggitt: 0000-0001-7697-7633.

Kok Loong Yeong: 0000-0001-8193-2130.

Keith C. Hamer: 0000-0002-2158-2420.

Jane K. Hill: 0000-0003-1871-7715.

Jennifer M. Lucey: 0000-0001-5224-091X.

References

Asner, G. P., Brodrick, P. G., Philipson, C., Vaughn, N. R., Martin, R. E., Knapp, D. E., ... Coombes, D. A. (2018). Mapped aboveground carbon stocks to advance forest conservation and recovery in Malaysian Borneo. *Biological Conservation, 217,* 289-310. https://doi.org/10.1016/j.biocon.2017.10.020

Bartoń, K. (2020). MuMIn: Multi-Model Inference. R package version 1.43.17. https://CRAN.R-project.org/package=MuMIn

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

Brodie, J. F., Giordano, A. J., Zipkin, E. F., Bernard, H., Mohd-Azlan, J., & Ambu, L. (2015). Correlation and persistence of hunting and logging impacts on tropical rainforest mammals. *Conservation Biology, 29,* 110-121. https://doi.org/10.1111/cobi.12389

Brofeldt, S., Theilade, I., Burgess, N., Danielsen, F., Poulsen, M., Adrian, T., ... Kurniawan, Y. (2014). Community monitoring of carbon stocks for REDD+: does accuracy and cost change over time? *Forests, 5,* 1834-1854. https://doi.org/10.3390/f5081834

Burnham, K. P., & Anderson, D. R. (2002). Model selection and multimodel inference: a practical information-theoretic approach. New York: Springer-Verlag.

Critchlow, R., Plumptre, A. J., Alidria, B., Nsubuga, M., Driciru, M., Rwetsiba, A., ... Beale, C. M. (2017). Improving law-enforcement effectiveness and efficiency in protected areas using ranger-collected monitoring data. *Conservation Letters, 10,* 572-580. https://doi.org/10.1111/conl.12288

Deere, N. J., Guillera-Arroita, G., Baking, E. L., Bernard, H., Pfeifer, M., Reynolds, G., ... Struebig, M. J. (2018). High Carbon Stock forests provide co-benefits for tropical biodiversity. *Journal of applied ecology, 55,* 997-1008. https://doi.org/10.1111/1365-2664.13023

Drescher, M., Perera, A. H., Johnson, C. J., Buse, L. J., Drew, C. A., Burgman, M. A. (2013). Toward rigorous use of expert knowledge in ecological research. *Ecosphere, 4,* 1-26. https://doi.org/10.1890/ES12-00415.1

Finer, M., Novoa, S., Weisse, M. J., Petersen, R., Mascaro, J., Souto, T., ... Martinez, R. G. (2018). Combating deforestation: From satellite to intervention. *Science, 360,* 1303-1305. https://doi.org/10.1126/science.aat1203

Gamer, M., Lemon, J., & Fellows, I. (2019). irr: Various Coefficients of Interrater Reliability and Agreement. R package version 0.84.1. https://CRAN.R-project.org/package=irr

Gao, T., Hedblom, M., Emilsson, T., & Nielsen, A. B. (2014). The role of forest stand structure as biodiversity indicator. *Forest Ecology and Management, 330,* 82-93. https://doi.org/10.1016/j.foreco.2014.07.007

Gibbs, H. K., Brown, S., Niles, J. O., & Foley, J.A. (2007). Monitoring and estimating tropical forest carbon stocks: making REDD a reality. *Environmental Research Letters, 2,* 045023. https://doi.org/10.1088/1748-9326/2/4/045023

Green, E. J., McRae, L., Freeman, R., Harfoot, M. B., Hill, S. L., Baldwin-Cantello, W., & Simonson, W. D. (2019). Below the canopy: global trends in forest vertebrate populations and their drivers. *PeerJ Preprints, 7,* e27882v1. https://doi.org/10.1098/rspb.2020.0533

Hoette, M. H., Kolodin, I. A., Bereznuk, S. L., Slaght, J. C., Kerley, L. L., Soutyrina, S. V., ... Miquelle, D. G. (2016). Indicators of success for smart law enforcement in protected areas: A case study for Russian Amur tiger (*Panthera tigris altaica*) reserves. *Integrative Zoology, 11,* 2-15. https://doi.org/10.1111/1749-4877.12168

Holck, M. H. (2008). Participatory forest monitoring: an assessment of the accuracy of simple cost–effective methods. *Biodiversity and Conservation, 17,* 2023-2036. https://doi.org/10.1007/s10531-007-9273-4

Hosonuma, N., Herold, M., De Sy, V., De Fries, R. S., Brockhaus, M., Verchot, L., ... Romijn, E. (2012). An assessment of deforestation and forest degradation drivers in developing countries. *Environmental Research Letters, 7,* 044009. https://doi.org/10.1088/1748-9326/7/4/044009

Kunjuraman, V., & Aziz, R. C. (2019). Community-Based Ecotourism (CBE) development in Lower Kinabatangan of Sabah, Malaysian Borneo. *Revista Turismo: estudos e práticas, 8,* 13-43.

Lindenmayer, D. B., Margules, C. R., & Botkin, D. B. (2000). Indicators of biodiversity for ecologically sustainable forest management. *Conservation Biology, 14,* 941-950. https:// doi.org/10.1046/j.1523-1739.2000.98533.x

Lindhe, A., & Drakenberg, B. (2019). *Forest integrity assessment: A simple and user-friendly tool for assessing and monitoring biodiversity conditions in forests and forest remnants*. Oxford: HCV Resource Network.

Lindhe, D., Payne, J., Reynolds, G., Drakenberg, B., Lucey, J. M., & Yeong, K. L. (2015). *Forest Integrity assessment template for Lowland Dipterocarp Rainforest*. Oxford: HCV Resource Network.

Luke, S. H., Barclay, H., Bidin, K., Chey, V. K., Ewers, R. M., Foster, W. A., ... Walsh, R. P. (2017). The effects of catchment and riparian forest quality on stream environmental conditions across a tropical rainforest and oil palm landscape in Malaysian Borneo. *Ecohydrology, 10,* e1827. https://doi.org/10.1002/eco.1827

Magnago, L. F. S., Magrach, A., Laurance, W. F., Martins, S. V., Meira-Neto, J. A. A., Simonelli, M., & Edwards, D. P. (2015). Would protecting tropical forest fragments provide carbon and biodiversity cobenefits under REDD+?. *Global Change Biology, 21,* 3455-3468. https://doi.org/10.1111/gcb.12937

Rosoman, G., Sheun, S. S., Opal, C., Anderson, P., & Trapshah, R., editors. (2017). *The HCS Approach Toolkit*. Singapore: HCS Approach Steering Group.

Tawato, N., Lucey, J. M., Senior, M., Benedick, S., Khen, C. V., Hill, J. K., & Hamer, K. C. (2014). Biodiversity of leaf-litter ants in fragmented tropical rainforests of Borneo: the value of publically and privately managed forest fragments. *Biodiversity Conservation, 23,* 3113-3126. https:// doi.org/10.1007/s10531-014-0768-5

Tierney, G. L., Faber-Langendoen, D., Mitchell, B. R., Shriver, W. G., & Gibbs, J. P. (2009). Monitoring and evaluating the ecological integrity of forest ecosystems. *Frontiers in Ecology and the Environment, 7,* 308-316. https://doi.org/10.1890/070176

Wang, X., Feng, Z., & Ouyang, Z. (2001). The impact of human disturbance on vegetative carbon storage in forest ecosystems in China. *Forest Ecology and Management, 148,* 117-123. https://doi.org/10.1016/S0378-1127(00)00482-5

Watson, J. E., Evans, T., Venter, O., Williams, B., Tulloch, A., Stewart, C., ... McAlpine, C. (2018). The exceptional value of intact forest ecosystems. *Nature Ecology & Evolution, 2,* 599. https://doi.org/10.1038/s41559-018-0490-x

Wurtzebach, Z., & Schultz, C. (2016). Measuring ecological integrity: history, practical applications, and research opportunities. *BioScience, 66,* 446-457. https://doi.org/10.1093/biosci/biw037

Yeong, K. L., Lucey, J. M., Jelling, A., Benedick, S., Bernadus, M., Ola, B., ... Hill, J. K. Habitat fragmentation disrupts rainforest regeneration. *In review*.

Yeong, K. L., Reynolds, G., & Hill, J. K. (2016a). Leaf litter decomposition rates in degraded and fragmented tropical rain forests of Borneo. *Biotropica, 48,* 443-452. https://doi.org/10.1111/btp.12319

Yeong, K. L., Reynolds, G., & Hill, J. K. (2016b). Enrichment planting to improve habitat quality and conservation value of tropical rainforest fragments. *Biodiversity Conservation, 25,* 957-973. https://doi.org/10.1007/s10531-016-1100-3

Supporting information

Ethical approval was granted by the University of York Ethics Committee (reference number: 20160404). Permissions to enter the field sites were granted by Wilmar International Ltd for privately owned sites, by Sabah Forestry Department for publically managed Virgin Jungle Reserves and Malua forest reserve, and Danum Valley Management Committee for the Danum Valley Conservation Area.

Additional supporting information may be found online in the Supporting Information section.

Funding information

A.J.S. was supported by the Natural Environment Research Council (NERC, Grant number NE/M013030/1). J.M.L. was supported by a NERC KE fellowship (Grant number NE/M006840/1-2). A.A., G.R., J.M.L., K.C.H., J.K.H. & Y.K.L. were supported by the Socially and Environmentally Sustainable Palm Oil

Research (SEnSOR) programme, which receives funding from the Roundtable on Sustainable Palm Oil (RSPO) and is facilitated by the South East Asia Rainforest Research Partnership (SEARRP).

Table 1 Indicators of forest condition, their description and reference in the ecological literature, and the value of the R-squared statistic when compared to the Forest Integrity Assessment (FIA) scores of our assessors.

Indicator of forest condition	Description	Reference(s)	Display Item	R-squared (R^2) vs. FIA scores
Alpha diversity (species richness)	Field sampling of dipterocarps (> 30cm DBH) and ants (quadrats).	Dipterocarps- Yeong *et al.* 2016a	Fig. 1a	0.50
		Ants- Tawatao *et al.* 2014	Fig. 1b	0.54
Vegetation structure	Vegetation quality score (dimensionless).	Yeong *et al.* 2016a	Fig. 1c	0.75
Aboveground Carbon	Aboveground carbon (units Mg C per Ha) mapped by combining airborne Light Detection and Ranging (LiDAR) with satellite imaging and other geospatial data.	Asner *et al.* 2018	Fig. 1d	0.69
Decomposition rate	Percentage loss of litter over 18 months.	Yeong *et al.* 2016a	Fig. 1e	0.74
Dipterocarp seedling prevalence	Presence in 1 hour searches, expressed as presence/absence per station (max 4, search radius 30m)	Yeong *et al. In review*	Fig. 1f	0.78
Dipterocarp fruit prevalence			Fig. 1g	0.26
Dipterocarp survival	Planted seedling survival (% over 18 months).	Yeong *et al.* 2016b	Fig. 1h	0.03.

Figure 1 Agreement between Forest Integrity Assessment (FIA) score (x-axes throughout) and independent metrics of forest condition (y-axes) sampled across a range of patch sizes in Sabah, Malaysian Borneo. The species richness values represent raw (non-bootstrapped) counts of: a) dipterocarp species (> 30cm DBH, Yeong et al. 2016a), and b) ant species (quadrats, Tawato et al. 2014). c) vegetation structural complexity or 'forest quality' score (dimensionless, Yeong et al. 2016a), d) aboveground carbon (Mg C per ha, Asner et al. 2018), e) Litter decomposition (% leaf litter mass lost over 120 days, Yeong et al. 2016a), f) dipterocarp seedling prevalence (n seedlings, max 4, Yeong et al. *In review*), g) fruit prevalence (n fruits, max 4, Yeong et al. *In review*), h) dipterocarp seedling survival (%, Yeong et al. 2016b). Error bars represent ± 95% confidence intervals. Where they appear, the dark green circles represent surveys conducted in continuous forest, while the lighter green circles represent surveys conducted in forest fragments, where symbol size is in proportion to the cube root of their area.

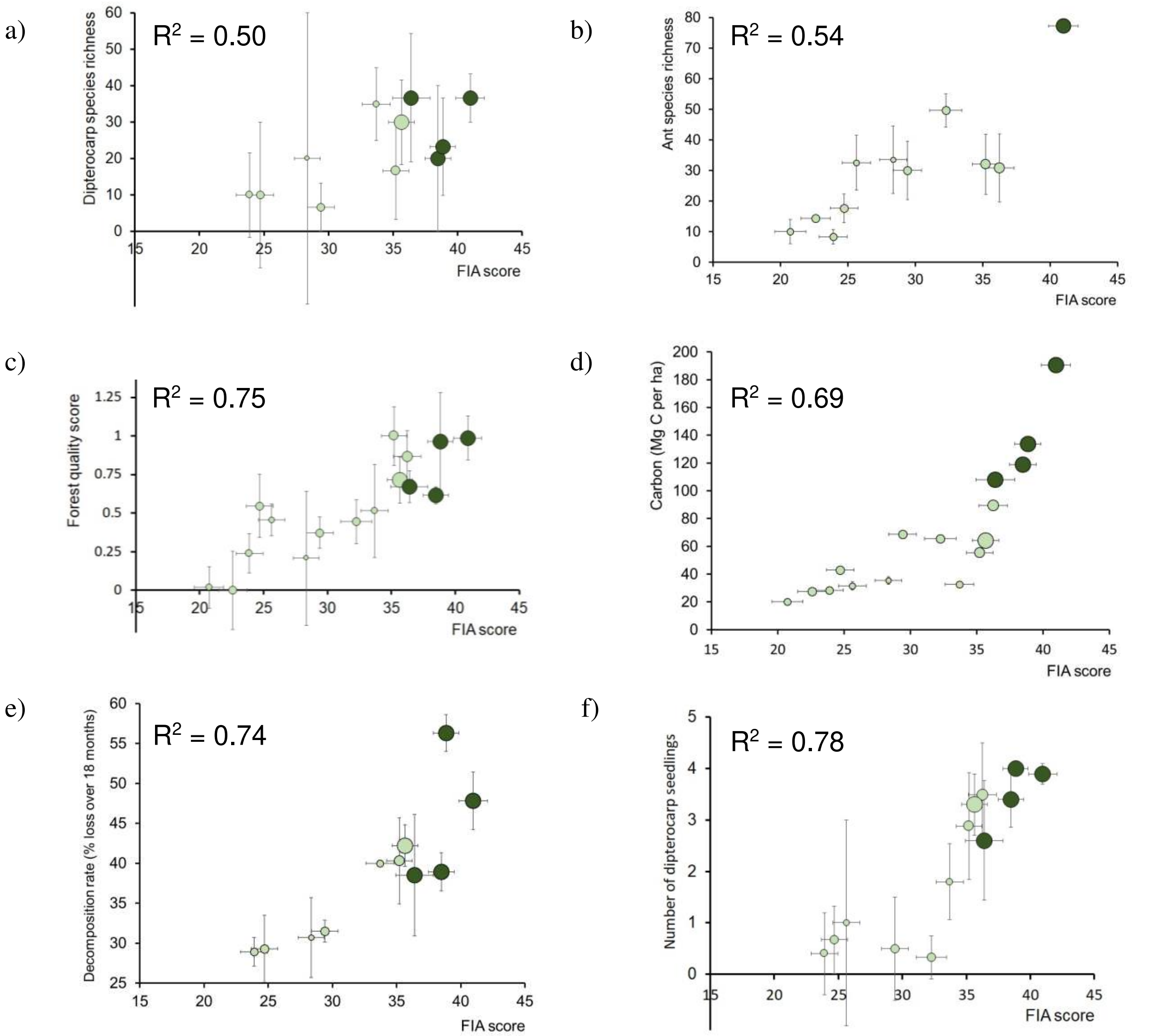
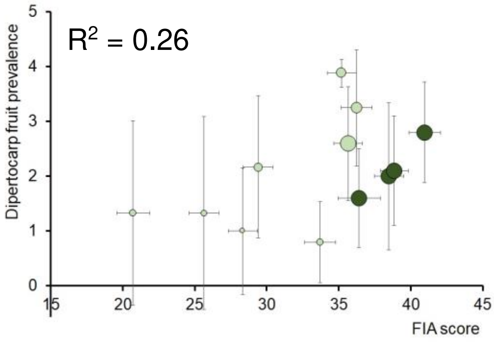


Figure 1 (continued)

g)



h)

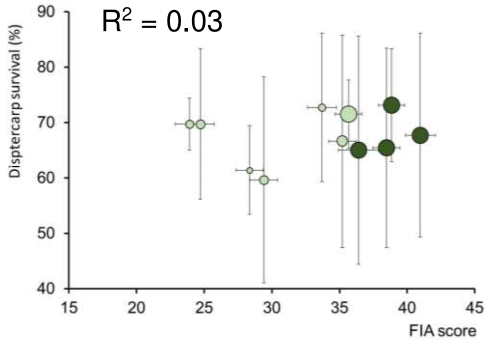


Figure 2 Performance of the Forest Integrity Assessment (FIA) questionnaire. Each question (n = 50 questions) is represented by a smaller circle, and scored based on its discriminative ability (y-axis) and the rate at which recorders agreed on the answer at each site, Fleiss's Kappa score (x-axis). Grand means for each criterion (n = 7 criteria) appear as larger circles, while the global mean score for the whole questionnaire appears as a black diamond (see key). Panels (A to D, illustrated by dotted lines) demark question scores relative to the global mean.

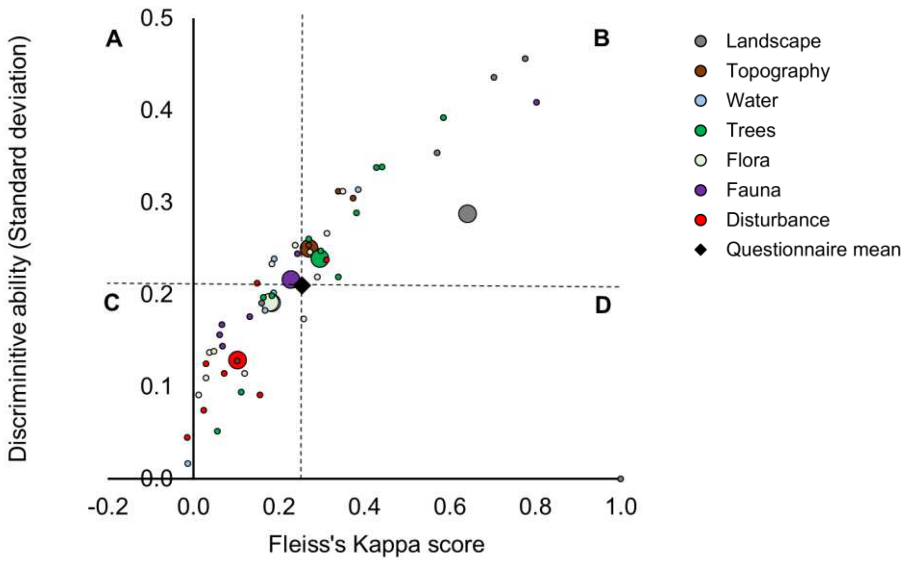


Figure 3 Effect of survey question removal on a) Forest Integrity Assessment (FIA) score and b) R-squared scores vs. independent metrics of forest condition. Neither the removal of the worst 10 questions for agreement, or the removal of all 'Panel C' questions (i.e. questions with low agreement and low discriminative ability, n = 23 questions), had a substantial effect on the ranking of sites by FIA score (a) or the estimates for R-squared, calculated using independent metrics of forest condition (b). Using only Panel B questions (those with high agreement rate, high discriminative ability, n = 20 questions) to calculate FIA scores had a mild effect on the order of site rankings, but an overall adverse effect on R-squared estimates.

a)

b)