England's statutory biodiversity metric offers lessons in ensuring biodiversity metrics for nature markets measure true change

Authors

Ivonne Salamanca Leon*[1], Tyler Hallman[1], Julia Baker[2], Ben Benatt[2], Eleanor Warren-Thomas[1], Julia P G Jones[1, 3]

School of Environmental and Natural Sciences, Bangor University, LL57 2UW, UK.

[2] Mott MacDonald, 10 Fleet Place, London EC4M 7RB, UK

[3]Department of Biology, Utrecht University, 3584 CS Utrecht, Netherlands.

Corresponding author: vns22sbv@bangor.ac.uk

Abstract

The need for standardized metrics for measuring losses and gains in biodiversity has resulted in many countries, and private sector initiatives, looking to adapt the England's Statutory Biodiversity Metric for mandatory Biodiversity Net Gain (BNG). BNG requires a minimum 10% uplift in biodiversity units out of infrastructure development, and these number of biodiversity units depends in part on field-based habitat condition assessments. We carried out simulations to explore the influence of uncertainty in condition assessments and found that there is inherent variability that introduce bias, even when a habitat's true condition remains unchanged. To explore likely real-world variability, we used an online test targeting ecologists (n=155) involved in BNG assessments and follow-up interviews with a sub-set of respondents (n=21). In our on-line test, BNG ecologists were more likely to correctly assess habitats in poor condition than good condition. Depending on the scenario and habitat, assessment variability could be significant enough to lead to a reported 10% uplift in biodiversity units even if there is no uplift. We acknowledge that our online test might not capture all true variability, but our findings underpin the focus on training and support for ecologists to reduce variability, and more detailed industry guidance to ensure that BNG legalisation in England delivers for nature. It is key to recognise the role of habitat condition assessment, tor the metric to measure true change and when adapted elsewhere.

Key words

BNG, biodiversity credits, observer bias, variability, biodiversity offsetting, biodiversity metrics

1. INTRODUCTION

There is increasing recognition that all sectors need to contribute to nature conservation and restoration if society is to successfully halt and reverse biodiversity loss by 2030 (Convention on Biological Diversity, 2022). The need to measure loss and gains in biodiversity to support national policies, or private sector initiatives, has resulted in a proliferation of biodiversity metrics (Hawkins et al., 2024; Marshall et al., 2020; Wauchope et al., 2024). However, research on how these are operationalized and how consistently they are applied is lacking (Simpson et al., 2022; Zu Ermgassen et al., 2022b).

Infrastructure development has an important role to play in boosting economic development, reducing poverty, and advancing the green transition (Zu Ermgassen et al., 2022a), but reconciling infrastructure development with nature restoration and enhancement remains a significant challenge. The mitigation hierarchy requires that negative impacts from development are avoided as much as possible, mitigated, restored as far as possible and that unavoidable impacts are compensated for through biodiversity offsetting to deliver at minimum a no net loss for biodiversity (Bull et al., 2016). Incorporation of the mitigation hierarchy has long been seen as best practice in environmental policy, originally with a focus on delivering no net loss of biodiversity, but there is growing interest around the world in policies requiring development to leave nature in a measurably better state than before (Jones et al., 2022; Marshall et al., 2023).

One example is England's Biodiversity Net Gain legalisation. Introduced as part of the UK Government's Environment Act of 2021, most new developments in England are required to deliver a 10% uplift in biodiversity units for at least 30 years (see for details DEFRA, 2024a; Rampling et al., 2024). BNG came into force for most developments requiring planning permission (with some exceptions) in February 2024. A similar requirement is expected for Nationally Significant Infrastructure Projects. The underpinning legislation and associated government guidance includes that developers follow the 'Biodiversity Gain Hierarchy' to avoid and reduce impacts on habitats (both direct and indirect), before delivering net gain through enhancing or creating habitats on-site. If on-site measures are not sufficient (or developers identify other options that are better for biodiversity), they can purchase biodiversity units from land managers/owners with legal

agreements in place for registered Biodiversity Gain Sites. Only as a last resort, developers can apply to purchase statutory biodiversity credits from the government.

One challenge for policies such as England's mandatory BNG is that they require a metric to consistently measure biodiversity (Marshall et al., 2024). Biodiversity, however, is a complex concept and cannot be captured with a single number (Marshall et al., 2020; Purvis & Hector, 2000; Wauchope et al., 2024). Condition-area metrics, which combine the area and the condition of a habitat, are commonly used as they require relatively limited data (Duffus et al., 2024). England's BNG legalisation requires that biodiversity unit change from baseline to post development is measured using the Statutory Biodiversity Metric (hereafter 'the metric') which is a condition-area metric (DEFRA, 2024a).

England's statutory biodiversity metric is attracting substantial international interest from jurisdictions including Saudi Arabia, Netherlands, and India which are in the process of introducing similar policies to Biodiversity Net Gain (White and Panks, 2024). Private sector initiatives are also developing metrics, drawing on England's statutory biodiversity metric, to support companies to assess and manage their impact on nature. For example, the Global Biodiversity Metric was launched at COP16 in Colombia following development of an Americas Biodiversity Metric (Ramboll, 2024). A growing number of operators are also developing metrics to support tradable biodiversity credits (Wauchope et al., 2024).

England's statutory metric was developed over nearly a decade from earlier iterations (Stuart et al., 2024). The metric measures 'biodiversity units' from individual parcels of habitats (e.g. grasslands, woodlands), linear habitats (e.g. hedgerows) and watercourses. For area and linear habitats, the metric has four components: area, distinctiveness, strategic significance, and condition. Habitat types in the UK's habitat classification system, known as UKHab, have each been preassigned distinctiveness scores. To incentivise enhancement or creation of habitats that support local conservation efforts, or Local Nature Recovery Strategies, all baseline habitats are assigned a low strategic significance score. Habitat condition is assessed on site by professional ecologists who score a habitat parcel against a set of criteria which are specific for each group of habitat types (SI 1). Criteria for condition assessment are based on earlier methodologies used in the UK for nature conservation purposes, such as the Common Standards Monitoring guidance (CSM) for grasslands, heathlands and wetlands (JNCC, 2004) and the England Woodland Biodiversity Group Woodland Condition Survey Method for woodlands (Woodland Wildlife Toolkit, 2020).

Recent work has revealed that, while the metric is correlated with most plant biodiversity variables (Marshall et al., 2024), it is not tightly correlated with common measures of biodiversity (e.g., species richness, individual abundance) for invertebrates, birds, or butterflies (Duffus et al., 2024; Marshall et al., 2024). However, it is important to note that the metric is only one of several statutory instruments and government guidance underpinning mandatory BNG (DEFRA, 2024a).

The habitat condition assessment influences the number of biodiversity units calculated, and therefore has implications for the number of biodiversity units a developer must deliver or purchase from BNG providers to achieve the 10% uplift in biodiversity units. As with all ecological surveys, there is likely to be variability among ecologists undertaking habitat condition assessments (Kelly et al., 2011). While training and experience can greatly reduce observer variability, even among experienced observers individual variability can impact results (Morrison, 2016).

We use simulations to explore how variability in habitat condition assessments affects measured uplifts in biodiversity units under four scenarios. We then estimate how much variability is likely to occur in the real-world using an online test, where ecologists were asked to apply the habitat condition criteria to videos and pictures of a random combination of four habitats in poor or good condition. We then carried semi-structured interviews with a sub-set of respondents to better understand the perceived challenges faced by those involved in habitat condition assessment in the field. Finally, we combined our simulations with our data on the level of variability in habitat condition assessment (estimated from the online test) to explore the likely impact on estimated biodiversity uplift.

2. METHODS

2.1. Simulations to explore how variability in habitat condition assessment influences measurement of biodiversity unit uplift

We simulated what would happen if a habitat parcel was surveyed twice by the same surveyor, under scenarios of real change in habitat condition, incorporating the probability that the surveyor sometimes incorrectly assessed condition. We simulated this for 10,000 habitat parcels. The Statutory Biodiversity Metric (eq. 1) calculates the baseline number of biodiversity units (BU) for a habitat parcel (DEFRA, 2024b) as follows:

$$BU = A \times D \times C \times S \quad (1)$$

Area (*A*) is the number of hectares or km of the parcel. Distinctiveness (*D*) is pre-assigned based on habitat type as very low, low, medium, high, and very high (with corresponding multipliers of 1, 2, 4, 6 and 8). Condition (*C*) is assessed on a parcel-by-parcel basis and is scored as poor, moderate and good (with corresponding multipliers of 1, 2, and 3). Strategic significance (*S*) is scored low for all baseline habitats where local nature recovery strategies are published (with multiplier of 1) then high if, post-development, BNG habitat creation or enhancement supports these strategies (or low if not).

Uplift within a single parcel is calculated as the change in biodiversity units from the baseline state before works (BU_1) to post development (BU_2):

$$Uplift = \frac{BU_2 - BU_1}{BU_1} \times 100 \quad (2)$$

We examined four scenarios in our simulations of a baseline and subsequent survey across 10000 parcels: (a) poor to poor condition, (b) poor to good condition, (c) good to good condition, and (d) good to poor condition. In each scenario, we simulated the initial baseline (*t = 1*) and subsequent (*t = 2*) condition assessments and calculated biodiversity units (1) and percent uplift (2) for all 10000 parcels. As we are interested in the effects of variability in condition assessments, we held the values of all variables except condition constant (area = 1 hectare, medium distinctiveness and low strategic significance). As the statutory biodiversity metric is a simple product of all components, the effects of condition assessment on percent uplift would be the same for any area, distinctiveness, or strategic significance selected as long as they are consistent between surveys. As with area, the additional multipliers present in the post-development assessment (time to target condition and difficulty risk) were held constant.

We examined the effects of variability in condition assessment by incorporating two probabilities: 1) the probability that an ecologist incorrectly assesses the condition (P_w), and 2) given that an ecologist incorrectly assesses the condition, the probability that that ecologist incorrectly assesses the condition by two levels e.g. poor to good rather than poor to moderate (P_v).

$$C_{ti} = \begin{cases} T_{ti}, & \text{with probability} \quad (1 - P_w) \\ T_{ti} \pm 1, & \text{with probability} \quad P_w(1 - P_v) \\ T_{ti} \pm 2, & \text{with probability} \quad P_w P_v \end{cases} \quad (3)$$

Where C_{ti} is the assessed condition and T_{ti} is the true condition, both at time t for parcel *i*. We ran simulations for values of P_w ranging from 0-1 (e.g., 0-100% probability of an incorrect assessment).

For ease of interpretation, we only examined P_v values of 0, 0.1, and 0.5 (i.e., 0, 10, and 50% probability of the assessor being wrong by two levels if they are wrong). As condition is bounded by 1 and 3, errors in condition assessment in true poor condition can only result in higher assessed condition and errors in true good condition can only result in lower assessed condition.

2.2. Estimating real-world variability in habitat condition assessment by BNG ecologists

To estimate the likely degree of variability in habitat condition assessments, we developed an online test (n=168). While an online test is inevitably different from conducting a habitat condition assessment in the field, doing this online allowed us to reach a sample of BNG ecologists working across England. The test was deployed as part of a survey targeted at professional ecologists who are or could be undertaking habitat condition assessments for mandatory BNG in England (hereafter referred to as BNG ecologists) (n=155). For details of the survey, see Table SI 1. We recruited respondents through posts on our own social media profiles and through those of the Association of Local Government Ecologists (ALGE), the Chartered Institute of Ecology and Environmental Management (CIEEM), the Botanical Society of Britain and Ireland (BSBI), the Field Studies Council, British BNG ecologists Facebook groups, and by emailing professional contacts and asking them to share the survey with their network. The survey was open from August 2024 to April 2025. One hundred and fifty-five of those who completed the survey met our criteria of being currently involved or likely to be involved in BNG assessments and were included in our analysis. Respondents varied in their professional backgrounds, level of experience, training levels and education (Figure SI 1). Seventy one percent worked for ecological or environmental consultancy firms or were independent consultants, 15% worked for statutory bodies and local governments, 9% for charities and 5% for developers (Figure SI 1d). There was a positive relationship between respondents' confidence in consistently applying the criteria when undertaking habitat condition assessments and their confidence in others applying them consistently (Figure SI 1c).

Respondents were asked to assess the condition of four area-based habitat types which are commonly impacted by infrastructure development in England (Table SI 2): Mixed Scrub, Other Neutral Grassland (Neutral Grassland), Modified Grassland and Lowland Mixed Deciduous Woodland (Deciduous Woodland). Each person was presented with videos and photographs of each habitat and were randomly assigned to be shown habitats in poor or good condition (FIGURE 1). Videos and photographs of each habitat were provided by an expert ecologist with 18 years' experience who is very familiar with the statutory metric and associated guidance. They were specifically selected as representative examples of good and poor condition. Assessments were checked by additional

experts in our network. In subsequent analysis we treat this assessment as "truth". There was debate among experts as to the true condition represented by the pictures we had selected for Mixed Scrub in good condition and so we considered both good and moderate condition to be the true condition in this case.

For each habitat, respondents were asked questions that reflected the criteria from the Technical Annex 1: Condition Assessment Sheets and Methodology, with minor modifications where needed to make the questions work in an online survey (Table SI 3). As in the metric, the responses to the criteria are added together and, depending on thresholds, the habitat is assigned a condition score of either poor, moderate or good (see Table SI 2). This means that an ecologist might get one criterion incorrect, but this does not affect the overall condition score. Respondents received feedback on how their responses compared to our assessment of habitat condition. This gamification made the survey engaging and more likely to be shared by BNG ecologists in their networks. For each question, respondents were given the option to select "*Don't know (not enough experience)*" or "*Don't know (pictures aren't clear enough or I need to visit the site)*." Respondents were invited to provide contact information for follow-up semi-structured interviews.

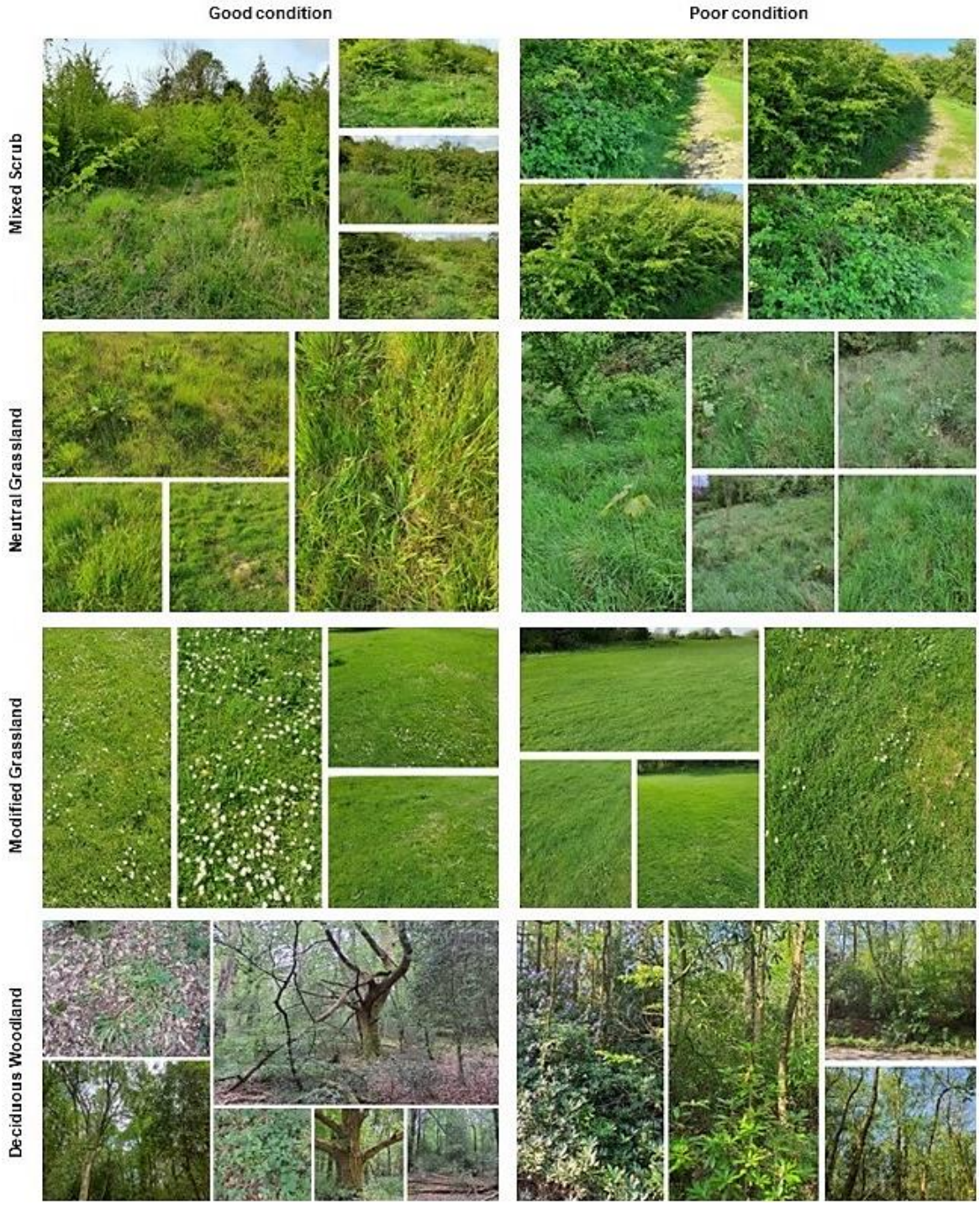

FIGURE 1 Composite images used for the online test to assess habitat condition of the four habitats. Images on the left are the "Good condition" set, while images on the right are the "Poor condition" set. Neutral Grassland refers to "Other Neutral Grassland" and Deciduous Woodland refers to Lowland Mixed Deciduous Woodland. Pictures by (we did not include to avoid breaching anonymity). The photos and videos of each habitat can be accessed here.

We used a logistic generalized linear mixed-effects model (GLMM) to explore whether BNG ecologists correctly matched habitat condition with our expert's condition assessment, noting the

limitations of an online survey. We removed data where respondents answered '*don't know*' to at least one criterion, created a new binary response variable classifying responses as correct or incorrect, and modelled the effects of predictors on the probability of correct habitat assessment (P_c). Predictor variables included habitat, true condition, years of experience, training, graduate education and self-confidence. An interaction effect between habitat and true condition was included to allow for habitat- and condition-specific differences. Participant ID was included as a random effect as each participant assessed four habitats. Education was converted to a binary variable of graduate and non-graduate education and types of training were treated as separate binary variables.

2.3. Semi-structured interviews to explore experiences with habitat condition assessment in the field

Those who indicated their willingness to participate were approached for an interview (see SI 6) to explore their experience in applying habitat condition assessment criteria in the field. Interviews were conducted with 21 people from various professional backgrounds (Table SI 4). Most were consultants (15), but we also interviewed local planning authority ecologists (4), those that worked for a developer (3), ecologists working for charities (4), and a student training in this area (Table S5, the latter not included in Table S5). Semi-structured interviews were transcribed using the Microsoft Teams AI transcription tool, corrected manually and edited for clarity and anonymisation. We carried out a thematic analysis to search for important emergent themes (Braun and Clarke, 2022). The codes were grouped into seven distinct themes with sub-themes as appropriate.

2.4. Incorporating our estimates of real-world variability in condition assessments into simulations to explore how variability affects biodiversity uplift measured

We used additional simulations to examine the effects of our estimated variability in condition assessment (from the online test) on uplift. As with the previous simulations, we included parcels that were surveyed at two time-steps, for example from baseline (pre-development) to post-development during the minimum 30-year BNG maintenance period. True habitat condition included the same four scenarios: (a) poor to poor condition, (b) poor to good condition, (c) good to good condition, and (d) good to poor condition, and the outcome of each assessment was again dependent upon the probability of incorrectly assessing condition (P_w), and the probability of subsequently incorrectly assessing the condition by two levels (P_v). As the online test results showed

that P_c was highly dependent upon habitat and true condition, we used habitat and condition specific P_c to inform P_w in our simulations (see SI 10 for parameters extracted from our GLMM analysis). For each habitat and true condition from our online test, we calculated P_w as $1\text{-}P_c$ as predicted by the GLMM described above. P_w in these simulations is therefore the actual estimated P_w for each habitat and condition combination from participant responses to our online test and represents real-world variability in condition assessments. As our GLMM was binomial, we did not have a similar estimate of P_v. We therefore used a rough estimate of P_v by calculating the percentage of incorrect assessments in each habitat and condition combination that were incorrect by two levels (i.e., of those that were incorrect, the percentage that were incorrect by two). We used the habitat distinctiveness defined in the metric for Mixed Scrub, Neutral Grassland, Modified Grassland, and Deciduous Woodland (4, 4, 2, and 6, respectively). We ran simulations for 10000 parcels, keeping area and strategic significance constant at 1 hectare and low significance. All analysis was conducted in R V4.4.1 (R Core Team, 2024).

Ethics approval statement: This research involved human participants in an online survey and follow-up interviews with ecologists and received full ethical approval from University Research Ethics (approval number CE202401A). All participants provided informed consent prior to their involvement, and their anonymity and confidentiality were strictly maintained throughout the study, in accordance with UKs General Data Protection Regulation (UK GDPR).

3. RESULTS

3.1. Variability in habitat condition assessment results in bias in measured biodiversity unit uplift

In each scenario, the magnitude of the bias depended on the probability of being wrong (P_w) and the probability of being wrong by two condition levels given the assessment is wrong (P_y). In some cases, where there is no change in condition (e.g. scenarios good-good and poor-poor), this bias could be sufficient for a 10% uplift in biodiversity units to be recorded (FIGURE 2a-c). While it is counterintuitive that the scenario good-good can be recorded as an increase in biodiversity units, this is because the variability in habitat condition assessment means the habitat may have been recorded as moderate or poor condition at baseline. The bias is more pronounced when assessors are wrong by two conditions 10% or 50% of the time, compared to when they are only ever wrong by one condition.

There was also positive bias in the good to poor scenario (FIGURE 2d), where true change is a decrease in biodiversity units of 67%. In this scenario a very high probability of being wrong (>55% or 90% chance of being wrong if wrong by two conditions 10% or 50% of the time respectively), would be required to incorrectly measure a 10% uplift. In the poor to good scenario, there is a negative bias in uplift as variability in condition assessments can only cause a reduction in biodiversity units.

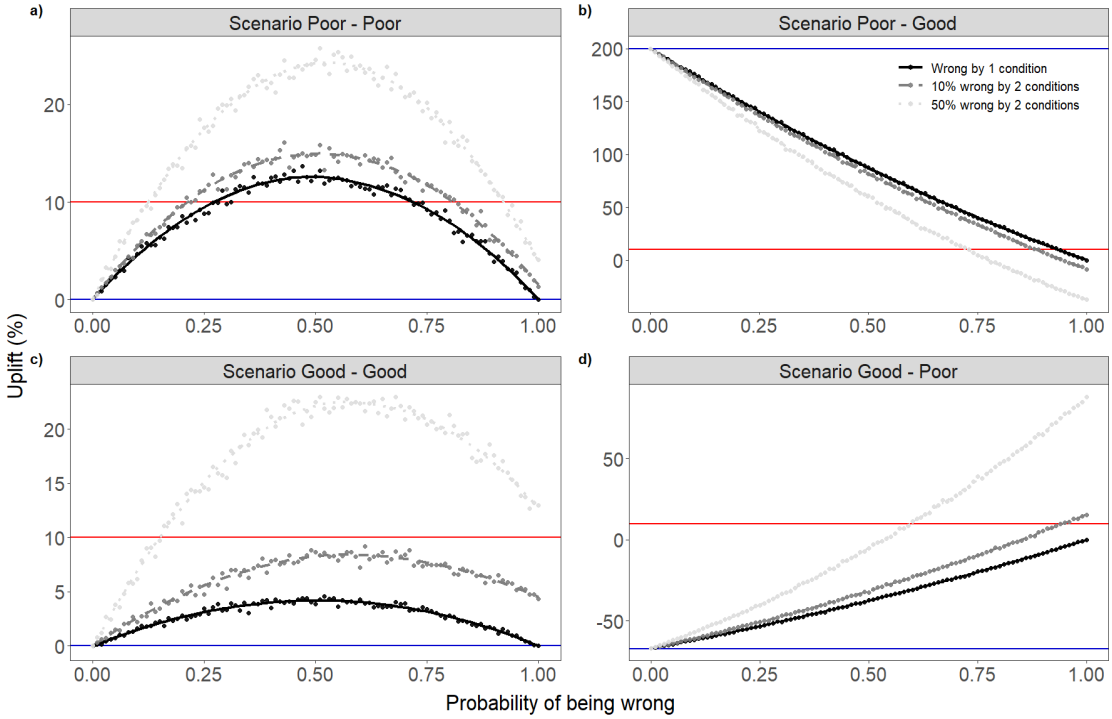

FIGURE 2 Interaction between the probability of being wrong (P_w) when assessing habitat condition and the estimated percent uplift for four scenarios: a) poor to poor condition, b) poor to good condition c) good to good condition and d) good to poor condition. Each point represents the mean result across 10,000 simulations for a habitat parcel assessed twice. The red solid lines indicate the 10% uplift required under mandatory Biodiversity Net Gain, and the blue solid lines indicate the true uplift per scenario. Black lines give the results assuming that when an assessor is wrong, they are only wrong by 1 condition (e.g. poor to moderate), grey dashed lines show results assuming that when assessors are wrong then for 10% or 50% of the time, they are wrong by two conditions (e.g. poor to good) i.e. P_v = 0, 0.1, 0.5.

3.2. Potential real-world variability in habitat condition assessment by BNG ecologists

In our online test, respondents correctly assessed poor condition habitats more frequently than good condition habitats (FIGURE 3). There were a substantial number of don't know responses for at least one criterion. The most common reason to answer '*Don't know*' was because the respondent felt the images were not sufficient and they needed to go to the field (89% of all don't knows given across all respondents and habitats, see Figure SI 2).

277

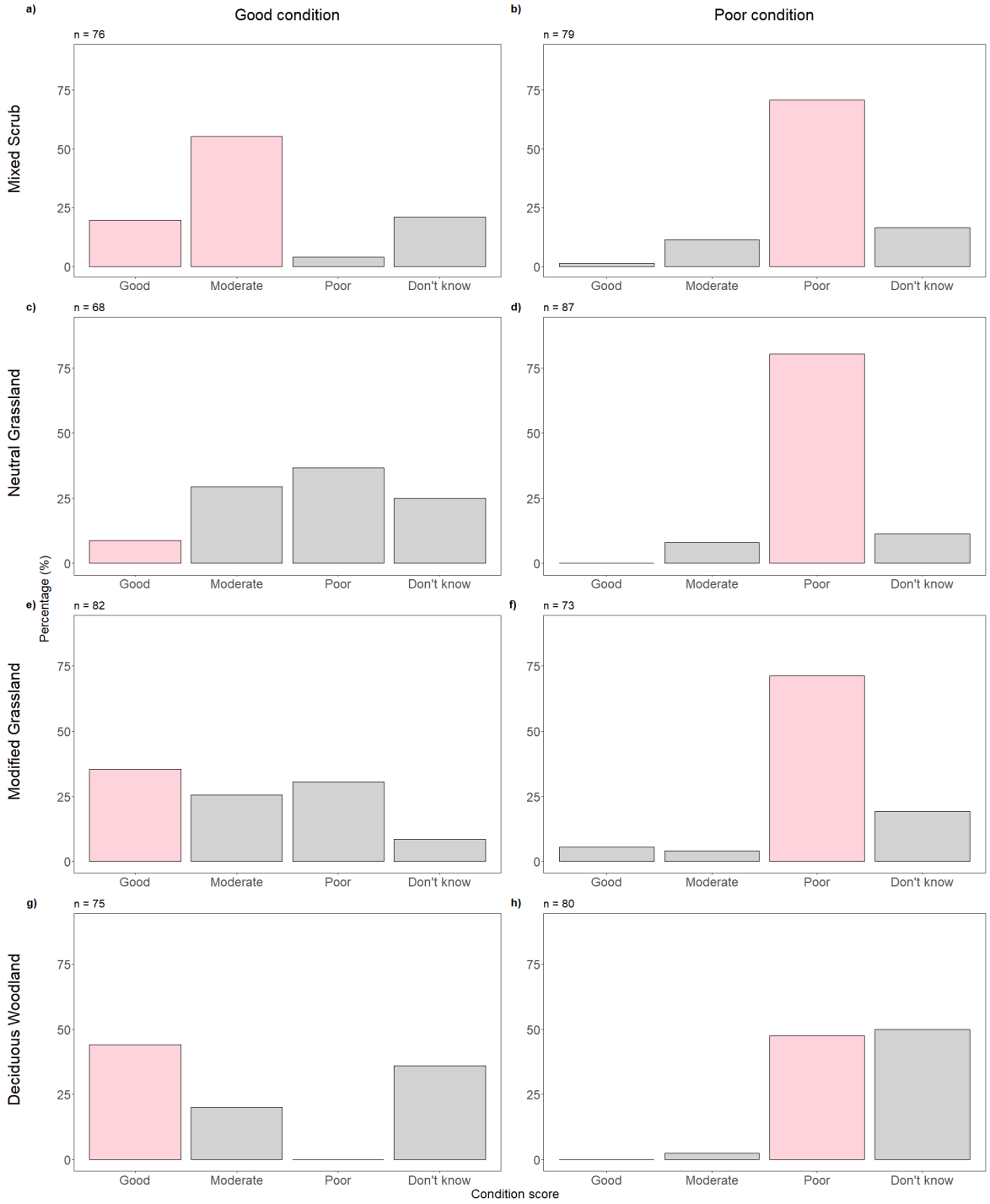

FIGURE 3 Percentage of condition scores given to each habitat by online test respondents shown images of Mixed Scrub (a and b), Neutral Grassland (Other Neutral Grassland; c and d), Modified Grassland (e and f), and Deciduous Woodland (Lowland Mixed Deciduous Woodland; g and h). Panels on the left show results from respondents shown images from parcels in "Good condition" (a, c, e, and g) and on the right from parcels in "Poor condition" (b, d, f, and h). Bars in pink show what we consider as truth based on expert assessment. Note that for Mixed Scrub the pictures were ambiguous, so we accept either good or moderate as 'correct'. Sample sizes are shown above each panel.

Some criteria were less consistently answered than others (see Figure SI 3). In good condition Neutral Grassland, for example, respondents should have answered yes to both criteria A and F (see Table SI 2), but only 45% and 8% of respondents answered yes to each of these criteria respectively. This contributed to the low percentage of respondents who correctly assessed good condition Neutral Grassland.

We found a strong interaction between habitat and true condition (FIGURE 4b). Respondents were much more likely to correctly assess Neutral Grassland, Modified Grassland, and Deciduous Woodland in poor condition than in good condition (this did not apply to Mixed Scrub). Even when using a conservative approach that accepts both good and moderate assessed conditions as correct for habitats in good condition, the probability of correctly assessing good condition Neutral Grassland and Modified grassland was only 53% and 74%, respectively (see Figure SI 4). Training, graduate education and years of experience, as captured by our online survey, were not significant predictors of correctly assessing habitat condition (FIGURE 4a) however these aspects were highlighted by respondents during interviews. Our estimates of the probability of being wrong (P_w) and the probability of being wrong by two conditions when wrong (P_y) estimated from the online test for each habitat are given in Table SI 7.

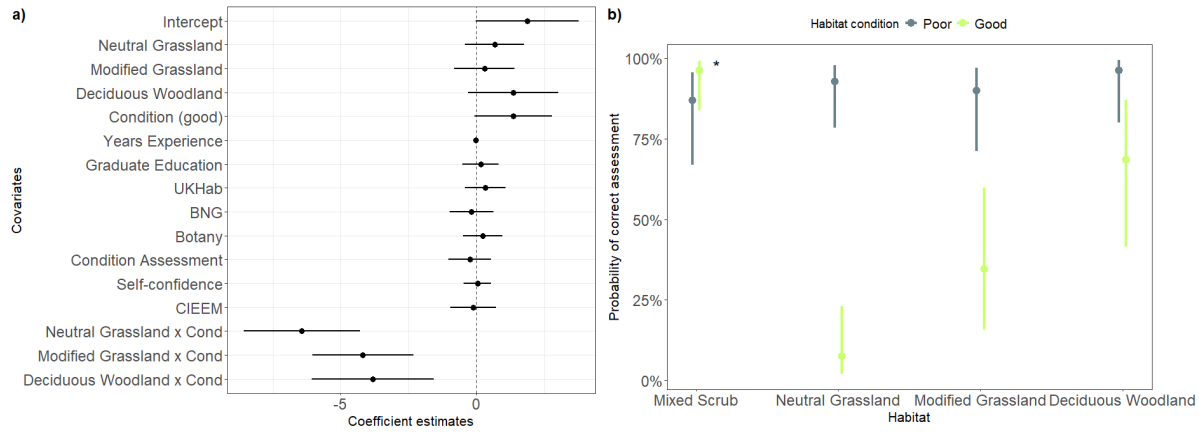

FIGURE 4 Coefficients a) and effects plots b) for the interaction between habitat and condition from a GLMM examining the effects of habitat, habitat condition, assessor education, assessor training, and assessor self-confidence on the correct assessment of habitat condition by respondents in the online test. The reference levels for this model (intercept) are Mixed Scrub in good condition, no graduate education, no training, and no years of experience (see Table SI 6).

3.3. Experiences of habitat condition assessment in the field

Interviewed participants (n=21) highlighted a range of challenges they reported facing when conducting habitat condition assessments (see Table SI 8). Challenges with accessing appropriate training, the importance of field experience, issues around funding to spend sufficient time in the field undertaking assessments, and time pressures meaning assessments sometimes were undertaken outside of the optimal survey season were all highlighted. Participants also highlighted challenges they reported facing with specific criteria (especially in grassland and woodland assessments), and how further guidance could make assessments more standardized. Some also explained how they, or their firm, have developed methods to increase consistency when undertaking habitat condition assessments.

3.4. Incorporating real-world variability into simulations to explore the influence on measured biodiversity uplift

Parameterizing our simulations using data from the online test reveals that, for some habitats and in some scenarios, a 10% uplift in biodiversity units may be recorded, on average across multiple parcels, even if there is no change in habitat condition (FIGURE 5). For parcels in poor condition which remain in poor condition or those in good condition which remain in good condition (where true uplift is zero), the variability introduced from condition assessments means that uplift is likely to be overestimated. In the poor-poor scenario this would not exceed 10% uplift (FIGURE 5a), however in the good-good scenario for two habitats (Neutral Grassland and Modified Grassland) the mean uplift across multiple parcels could exceed 10% despite no uplift occurring (FIGURE 5c). While this feels unintuitive, it is because variability can occur between the baseline assessment (before development) and then subsequent assessments during the minimum 30-year BNG maintenance period.

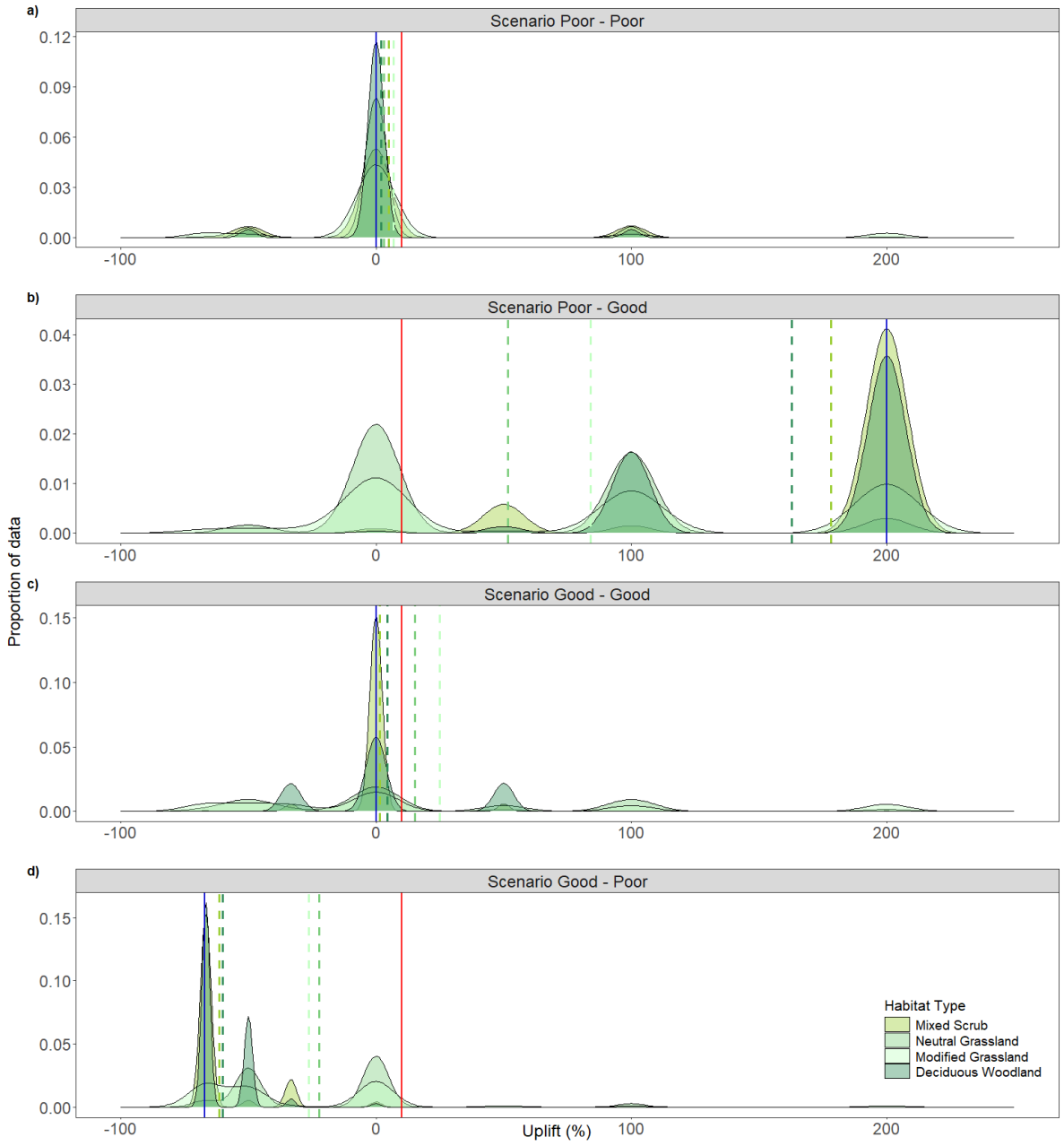

FIGURE 5 Simulated percentage uplift in biodiversity units across four habitat types. Panels show four scenarios; poor-poor, poor-good, good-good, and good-poor using values of probability of being wrong (P_w) and the probability of being wrong by two conditions where the surveyor is wrong (P_v) estimated from the online test for each habitat (Table SI 6). Red solid lines indicate 10% uplift, blue solid lines indicate the true uplift per scenario, and the green dashed lines indicate the mean uplift per habitat.

4. DISCUSSION

Our simulations exploring the influence of variability in habitat condition assessment on the biodiversity unit uplift calculated by England's statutory biodiversity metric reveals a directional bias. This bias suggests that, over multiple assessments, for example over the minimum 30 years of

mandatory BNG habitat management and monitoring, there will be a tendency for more uplift to be detected than has occurred. To understand how important this bias is likely to be in the real-world, we wanted to estimate how much variability there is likely to be in habitat condition assessments conducted by BNG ecologists. Using an online test, we found inter-observer variability in habitat condition assessments. For some habitats and in some scenarios, this would be sufficient to record a 10% uplift in biodiversity units even if no uplift occurred.

Observer variability is inevitable in ecological surveys, whether carried out by researchers, conservation practitioners or in the context of infrastructure development. Even where objectively correct answers exist, a broad range of factors from observer experience to the local weather conditions, can influence the observation process and add variability and bias to ecological metrics (Loosen et al., 2022; Sunde and Jessen, 2013). Such bias is pervasive across taxonomic groups and habitats, prompting the development of increasingly sophisticated statistical models (Kéry et al., 2024). Habitat condition assessment whether for nature conservation purposes or mandatory BNG, requires professional judgement, which inevitably includes a degree of subjectivity which in turn increases variability (Harwood et al., 2016). Our research highlights how important this variability is in the application of metrics for measuring loss and gain in biodiversity.

There are important caveats to our research. First, an assessment from pictures and videos in an online test is not the same as a field visit. While the photos and videos were carefully taken by an experienced ecologist to reflect the specific features that the criteria require, it is certainly possible that inter-observer variability may be lower in real field conditions. Second, to avoid the test becoming too long, we included four habitats commonly impacted by development in England, variability may be higher or lower in other habitats not considered here. However, we believe that the test, whilst imperfect, reveals that variability in habitat condition assessments is likely and may be sufficient to influence the results where the metric is used to measure change in biodiversity.

The interviews provide valuable insights into experiences of a few individuals, who suggested how variability in condition assessments could be reduced.

4.1. High-quality training, and mentoring will likely help to reduce variability in habitat condition assessment

The Statutory Biodiversity Metric User Guide is a statutory instrument. It makes clear that the metric calculation and habitat condition assessments should be undertaken by competent people (DEFRA, 2024b), i.e. professionals with the experience, skills and knowledge needed to carry out these tasks

reliably. The wider literature on observer variability in ecological surveys highlights the importance of training for reducing variability (Morrison, 2021, 2016). Although we did not detect an association between BNG training and the chance of making a correct assessment, this might have been because it was not detectable using the crude measures in our statistical analysis (if respondents had taken, or not taken, specific training courses). The importance of high-quality training and on-going mentoring to improving the quality and consistency of habitat condition assessment came across clearly in our interviews (see Table SI 8 theme a).

4.2. Good practice guidance could further help reduce variability

The diligence of BNG ecologists came across during our interviews and respondents described how further guidance could help BNG ecologists to more consistently assess criteria when conducting habitat condition assessment (see Table SI 8 theme c). Some firms and individuals have developed their own good practice guidance to reduce variability (see Table SI 8 theme c). Even simple improvements to the layout of forms could potentially reduce variability (see Table SI 8 themes c and f). In addition, there could be a role for industry professional bodies in developing and making available guidance and standardization to reduce variability.

4.3. Refining the wording of condition criteria could reduce variability

The Statutory Biodiversity Metric is being reviewed every three to five years in order to make recommendations of possible improvements. Our research suggests that this review should include the wording of habitat condition criteria, with the aim for refinements to be clearer and reduce the subjectivity causing variability. For example, respondents to our interviews suggested that the criterion "*The parcel represents a good example of its habitat type*" which appears in Mixed Scrub and Neutral Grassland, is challenging to apply consistently. Respondents also highlighted the challenge of distinguishing between other neutral grassland in poor condition and modified grasslands in good condition, as an example. This review might not only reduce variability for England's mandatory BNG under the Town and Country Planning Act 1990, but also for the possible forthcoming mandatory BNG requirement for Nationally Significant Infrastructure Projects.

4.5. Lessons for those adapting England's statutory biodiversity metric

There are valuable lessons from this work for those adapting England's statutory biodiversity metric to support policies similar to Biodiversity Net Gain, or private sector initiatives to measure and address biodiversity impacts. Reducing variability in habitat condition assessments is key to ensuring that the metric can measure real change in biodiversity. The way the multipliers in the metric work mean that variability can result in bias (measuring an increase in biodiversity when none has occurred) rather than simply introducing noise. Variability in habitat condition assessment carried out by ecologists in the field can be reduced through training, high-quality guidance targeted to specific habitats, and ensuring the wording of condition criteria are clear. There is also a potential role for the responsible use of Artificial Intelligence (AI) to increase consistency (Reynolds et al., 2025; Tuia et al., 2022). There has long been interest in exploring the potential for assessing habitat condition from remote sensed data (Nagendra et al., 2013; Wallace et al., 2006), and advances in AI make this more practical (Harwood et al., 2016). The use of remote sensed data (from drones or satellites) combined with AI could, if properly trained against high-quality field data, help reduce variability in condition assessments and could allow ecologists to focus efforts on habitats of higher ecological value. However, while these approaches have potential, they should not prematurely replace field surveys, and the role of professional ecologists to check AI-based classifications will not be easily replaced.

England's mandatory BNG is impressive and ambitious legalisation resulting in the involvement of ecologists earlier in the planning process than before, and likely to result in better outcomes for biodiversity than previous legalisation (Stuart et al., 2024). Government agencies, academia, NGOs, and industry need to work together so it can deliver on its potential to contribute to nature's recovery. The results are relevant to efforts around the world to develop standardized metrics to measure losses and gains in biodiversity: while condition-area metrics such as England's statutory metric are a pragmatic and usable tool, their ability to measure change is vulnerable to how repeatable the condition assessments are.

5. REFERENCES

Braun, V., Clarke, V., 2022. Conceptual and design thinking for thematic analysis. Qualitative Psychology 9, 3–26. https://doi.org/10.1037/qup0000196

Bull, J.W., Gordon, A., Watson, J.E.M., Maron, M., 2016. Seeking convergence on the key concepts in 'no net loss' policy. Journal of Applied Ecology 53, 1686–1693. https://doi.org/10.1111/1365-2664.12726

Convention on Biological Diversity, 2022. Kunming-Montreal Global Biodiversity Framework. Presented at the CONFERENCE OF THE PARTIES TO THE CONVENTION ON BIOLOGICAL DIVERSITY Fifteenth meeting – Part II, Convention on Biological Diversity, Montreal, Canada.

DEFRA, 2024a. Biodiversity net gain [WWW Document]. GOV.UK. URL https://www.gov.uk/government/collections/biodiversity-net-gain (accessed 6.19.24).

DEFRA, 2024b. The Statutory Biodiversity Metric-User Guide.pdf.

Duffus, N.E., Atkins, T.B., Zu Ermgassen, S.O.S.E., Grenyer, Richard., Bull, J.W., Castell, D.A., Stone, Ben., Tooher, Niamh., Milner-Gulland, E.J., Lewis, O.T., 2024. Metrics based on habitat area and condition are poor proxies for invertebrate biodiversity. https://doi.org/10.1101/2024.10.02.616290

Ecogain, 2024. CLIMB. Changing land use impact on biodiversity. URL https://climb.ecogain.se/ (accessed 6.16.25).

Harwood, T.D., Donohue, R.J., Williams, K.J., Ferrier, S., McVicar, T.R., Newell, G., White, M., 2016. Habitat Condition Assessment System: a new way to assess the condition of natural habitats for terrestrial biodiversity across whole regions using remote sensing data. Methods Ecol Evol 7, 1050–1059. https://doi.org/10.1111/2041-210X.12579

Hawkins, F., Beatty, C.R., Brooks, T.M., Church, R., Elliott, W., Kiss, E., Macfarlane, N.B.W., Pugliesi, J., Schipper, A.M., Walsh, M., 2024. Bottom-up global biodiversity metrics needed for businesses to assess and manage their impact. Conservation Biology 38, e14183. https://doi.org/10.1111/cobi.14183

JNCC, 2004. Common Standards Monitoring guidance | JNCC - Adviser to Government on Nature Conservation [WWW Document]. Joint Nature Conservation Committee. URL https://jncc.gov.uk/our-work/common-standards-monitoring-guidance/#guidance-documents (accessed 5.1.25).

Jones, K.R., von Hase, A., Costa, H.M., Rainey, H., Sidat, N., Jobson, B., White, T.B., Grantham, H.S., 2022. Spatial analysis to inform the mitigation hierarchy. Conservat Sci and Prac 4. https://doi.org/10.1111/csp2.12686

448 Kelly, A.L., Franks, A.J., Eyre, T.J., 2011. Assessing the assessors: Quantifying observer variation in
449 vegetation and habitat assessment. Ecological Management & Restoration 12, 144–148.
450 https://doi.org/10.1111/j.1442-8903.2011.00597.x
451 Kéry, M., Royle, J.A., Hallman, T., Robinson, W.D., Strebel, N., Kellner, K.F., 2024. Integrated distance
452 sampling models for simple point counts. Ecology 105, e4292.
453 https://doi.org/10.1002/ecy.4292
454 Loosen, A., Devineau, O., Zimmermann, B., Marie Mathisen, K., 2022. The importance of evaluating
455 standard monitoring methods: Observer bias and detection probabilities for moose pellet
456 group surveys. PLoS ONE 17, e0268710. https://doi.org/10.1371/journal.pone.0268710
457 Marshall, C.A.M., Wade, K., Kendall, I.S., Porcher, H., Poffley, J., Bladon, A.J., Dicks, L.V., Treweek, J.,
458 2024. England's statutory biodiversity metric enhances plant, but not bird nor butterfly,
459 biodiversity. Journal of Applied Ecology 61, 1918–1931. https://doi.org/10.1111/1365-
460 2664.14697
461 Marshall, E., Southwell, D., Wintle, B.A., Kujala, H., 2023. A global analysis reveals a collective gap in
462 the transparency of offset policies and how biodiversity is measured. Conservation Letters
463 17, e12987. https://doi.org/10.1111/conl.12987
464 Marshall, E., Wintle, B.A., Southwell, D., Kujala, H., 2020. What are we measuring? A review of
465 metrics used to describe biodiversity in offsets exchanges. Biological Conservation 241,
466 108250. https://doi.org/10.1016/j.biocon.2019.108250
467 Morrison, L.W., 2021. Nonsampling error in vegetation surveys: understanding error types and
468 recommendations for reducing their occurrence. Plant Ecol 222, 577–586.
469 https://doi.org/10.1007/s11258-021-01125-5
470 Morrison, L.W., 2016. Observer error in vegetation surveys: a review. JPECOL 9, 367–379.
471 https://doi.org/10.1093/jpe/rtv077
472 Nagendra, H., Lucas, R., Honrado, J.P., Jongman, R.H.G., Tarantino, C., Adamo, M., Mairota, P., 2013.
473 Remote sensing for conservation monitoring: Assessing protected areas, habitat extent,
474 habitat condition, species diversity, and threats. Ecological Indicators 33, 45–59.
475 https://doi.org/10.1016/j.ecolind.2012.09.014
476 Ohlson, J., Johansson, A., Rydin, L., Bökmark, S., Hägglund, T., Olsson, S.B., Berg, A., Jonasson, C.,
477 Härdmark, E., 2023. PROJECT MANAGEMENT, WAYS OF WORKING WITH CLIMB [WWW
478 Document]. CLIMB. Changing land use impact on biodiversity.
479 Purvis, A., Hector, A., 2000. Getting the measure of biodiversity. Nature 405, 212–219.
480 https://doi.org/10.1038/35012221

R Core Team, 2024. _R: A Language and Environment for Statistical Computing_. R Foundation for Statistical Computing.

Ramboll, 2024. Ramboll's biodiversity metrics [WWW Document]. Measuring Biodiversity. URL https://www.ramboll.com/measuring-biodiversity (accessed 6.12.25).

Rampling, E.E., Zu Ermgassen, S.O.S.E., Hawkins, I., Bull, J.W., 2024. Achieving biodiversity net gain by addressing governance gaps underpinning ecological compensation policies. Conservation Biology 38, e14198. https://doi.org/10.1111/cobi.14198

Reynolds, S.A., Beery, S., Burgess, N., Burgman, M., Butchart, S.H.M., Cooke, S.J., Coomes, D., Danielsen, F., Di Minin, E., Durán, A.P., Gassert, F., Hinsley, A., Jaffer, S., Jones, J.P.G., Li, B.V., Mac Aodha, O., Madhavapeddy, A., O'Donnell, S.A.L., Oxbury, W.M., Peck, L., Pettorelli, N., Rodríguez, J.P., Shuckburgh, E., Strassburg, B., Yamashita, H., Miao, Z., Sutherland, W.J., 2025. The potential for AI to revolutionize conservation: a horizon scan. Trends in Ecology & Evolution 40, 191–207. https://doi.org/10.1016/j.tree.2024.11.013

Simpson, K.H., de Vries, F.P., Dallimer, M., Armsworth, P.R., Hanley, N., 2022. Ecological and economic implications of alternative metrics in biodiversity offset markets. Conservation Biology 36, e13906. https://doi.org/10.1111/cobi.13906

Stuart, A., Bond, A., Franco, A., Gerrard, C., Baker, J., Ten Kate, K., Butterworth, T., Bull, J., Treweek, J., 2024. How England got to Mandatory Biodiversity Net Gain: A Timeline. https://doi.org/10.2139/ssrn.4883170

Sunde, P., Jessen, L., 2013. It counts who counts: an experimental evaluation of the importance of observer effects on spotlight count estimates. Eur J Wildl Res 59, 645–653. https://doi.org/10.1007/s10344-013-0717-8

Tuia, D., Kellenberger, B., Beery, S., Costelloe, B.R., Zuffi, S., Risse, B., Mathis, A., Mathis, M.W., Van Langevelde, F., Burghardt, T., Kays, R., Klinck, H., Wikelski, M., Couzin, I.D., Van Horn, G., Crofoot, M.C., Stewart, C.V., Berger-Wolf, T., 2022. Perspectives in machine learning for wildlife conservation. Nat Commun 13, 792. https://doi.org/10.1038/s41467-022-27980-y

Wallace, J., Behn, G., Furby, S., 2006. Vegetation condition assessment and monitoring from sequences of satellite imagery. Eco Management Restoration 7. https://doi.org/10.1111/j.1442-8903.2006.00289.x

Wauchope, H.S., Zu Ermgassen, S.O.S.E., Jones, J.P.G., Carter, H., Schulte To Bühne, H., Milner-Gulland, E.J., 2024. What is a unit of nature? Measurement challenges in the emerging biodiversity credit market. Proc. R. Soc. B. 291, 20242353. https://doi.org/10.1098/rspb.2024.2353

White, N., Panks, S., 2024. International use of England's biodiversity metric having global impact – Natural England. Natural England. URL https://naturalengland.blog.gov.uk/2024/11/08/international-use-of-englands-biodiversity-metric-having-global-impact/

Woodland Wildlife Toolkit, 2020. Woodland Condition Assessment [WWW Document]. URL https://woodlandwildlifetoolkit.sylva.org.uk/ (accessed 5.1.25).

Zu Ermgassen, S.O.S.E., Drewniok, M.P., Bull, J.W., Corlet Walker, C.M., Mancini, M., Ryan-Collins, J., Cabrera Serrenho, A., 2022a. A home for all within planetary boundaries: Pathways for meeting England's housing needs without transgressing national climate and biodiversity goals. Ecological Economics 201, 107562. https://doi.org/10.1016/j.ecolecon.2022.107562

Zu Ermgassen, S.O.S.E., Howard, M., Bennun, L., Addison, P.F.E., Bull, J.W., Loveridge, R., Pollard, E., Starkey, M., 2022b. Are corporate biodiversity commitments consistent with delivering 'nature-positive' outcomes? A review of 'nature-positive' definitions, company progress and challenges. Journal of Cleaner Production 379, 134798. https://doi.org/10.1016/j.jclepro.2022.134798