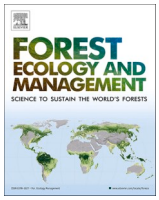
Forest Ecology and Management 595 (2025) 123006

Contents lists available at ScienceDirect

Forest Ecology and Management

journal homepage: www.elsevier.com/locate/foreco



Inter-observer reliability in forest conservation value assessments

Mari Jönsson [a,b,*], Anne-Maarit Hekkala [c], Karina Clemmensen [d], Julia Kyaschenko [b], Simon Kärvemo [b], Louis Mielke [d], Jörgen Sjögren [c], Joachim Strengbom [b]

[a] Swedish University of Agricultural Sciences, Swedish Species Information Centre, Box 7007, Uppsala 75007, Sweden
[b] Swedish University of Agricultural Sciences, Department of Ecology, Box 7044, Uppsala 75007, Sweden
[c] Swedish University of Agricultural Sciences, Department of Wildlife, Fish and Environmental Studies, Skogsmarksgränd, Umeå 90183, Sweden
[d] Swedish University of Agricultural Sciences, Department of Forest Mycology and Plant Pathology, Box 7026, Uppsala 75007, Sweden

ARTICLE INFO

Keywords:
Boreal forest
Forest and site indicators
Observer bias
Qualitative method
Sampling errors
Semi-quantitative method

ABSTRACT

Identifying and safeguarding forests of high conservation value is central to sustainable forest management. Qualitative and semi-quantitative surveys of forest conservation indicators often form the evidence base for management decisions. However, it remains unclear how consistently different surveyors assess such indicators using these methods. In this study, we evaluated inter-observer reliability (IOR) among triplets of professional biologists conducting independent surveys in 14 boreal, conifer-dominated forest stands in south-central Sweden. Surveyors recorded 50 qualitative indicators (presence-absence) and 20 semi-quantitative indices (counts and ordinal scores). We hypothesized that semi-quantitative assessments would yield higher IOR, as they are based on structured counts and ordinal scales applied within defined plots, which may reduce subjectivity. Contrary to this expectation, several qualitative indicators — based on presence-absence observations at stand scale — showed equal or even higher IOR. For example, the overall IOR for the qualitative composite score was good (intra-class correlation coefficients; ICC = 0.84), while many semi-quantitative indicators reached only moderate levels (ICC = 0.50–0.70). Indicators related to downed deadwood exhibited moderate to substantial IOR across both methods, while indicators involving standing structures, such as high nature value (HNV) trees and tree microhabitats, showed lower IOR. Our findings highlight that indicator-specific characteristics (e.g., subjectivity, rarity), rather than assessment method alone, influence reliability. Excluding low-reliability structural (e.g., tree microhabitats) indicators from qualitative protocols slightly improved overall agreement. We recommend integrating IOR analyses to refine survey protocols, guide surveyor training, and improve consistency in forest conservation value assessments. Even small-scale IOR evaluations — such as those involving three independent surveyors — can yield valuable insights into observer bias within relatively homogeneous groups of professional surveyors. Future research should expand such analyses to a wider range of ecosystems, indicator types, and surveyor backgrounds to strengthen the robustness and credibility of qualitative and semi-quantitative forest conservation value assessments.

1. Introduction

Forests are complex socio-ecological systems, shaped by site conditions, internal dynamics, and natural or anthropogenic pressures (Berglund and Kuuluvainen, 2021; Dieler et al., 2017). Identifying and safeguarding forests of high conservation value is therefore central to sustainable forest management (Oettel and Lapin, 2021). Reliable assessment of conservation values is critical for prioritizing forest stands for management and protection, but this requires assessment tools that are transferable, measurable, cost-effective, and understandable to both practitioners and policy-makers (Bellamy et al., 2024; Lindenmayer and Likens, 2011). To meet these demands, conservation value assessments (CVAs) in forests often rely on multiple, and sometimes composite, qualitative and semi-quantitative indicators or proxies (features of conservation interest) and indices (metrics or scoring systems used to measure these features) (Bellamy et al., 2024; Drakenberg and Lindhe,

Abbreviations: CCC, concordance correlation coefficient; CVA, conservation value assessment; HNV, high nature value; IOR, inter-observer reliability; IOA, inter-observer agreement; ICC, intra-class correlation coefficient.

* Corresponding author at: Swedish University of Agricultural Sciences, Department of Ecology, Box 7044, Uppsala 75007, Sweden.
E-mail address: mari.jonsson@slu.se (M. Jönsson).

https://doi.org/10.1016/j.foreco.2025.123006
Received 15 April 2025; Received in revised form 8 July 2025; Accepted 14 July 2025
Available online 16 July 2025
0378-1127/© 2025 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

1999; Hekkala et al., 2023; Zeller et al., 2022). Unlike quantitative species inventories, these CVAs do not require specialist taxonomic expertise or equipment, and are typically based on presence-absence scoring, ordinal ratings, or counts of habitat features. This makes them efficient, affordable, and feasible tools for large-scale assessments in forest management and policy.

In Europe, habitat-based CVAs that assess habitat amount and structural variation as proxies for forest biodiversity are commonly used (e.g., Bellamy et al., 2024; Blicharska, 2005; Gao et al., 2015; Hekkala et al., 2023; Perhans et al., 2011; Zeller et al., 2022). For example, the widely used Index of Biodiversity Potential (IBP) combines key stand structures, composition, and habitat attributes known to support biodiversity (Zeller et al., 2022). Although IBP scores reflect biodiversity potential, complementary species surveys are often needed for more detailed information (Zeller et al., 2022), and observer expertise influences assessment precision (Gosselin and Larrieu, 2020). Other studies similarly show that habitat-based CVAs can predict species richness or abundance of some species groups (Hekkala et al., 2023) and the occurrence of red-listed fungi and bryophytes on deadwood (Larsson Ekström et al., 2025), while for others — such as soil fungi — they may fail to capture variation (Kyaschenko et al., 2025). Despite mixed evidence of their ecological validity and methodological framework (Greco et al., 2019), CVAs remain widely used and valuable tools for conservation planning, based on their cost-efficiency and knowledge that sites with higher habitat diversity values are generally more likely to support species of conservation interest (Hekkala et al., 2023; Larsson Ekström et al., 2025; Zeller et al., 2022).

Importantly, the accuracy and usefulness of CVAs also depend on their consistency among observers (Gosselin and Larrieu, 2020). Qualitative and semi-quantitative indicators involve subjective judgement, making them potentially sensitive to observer effects such as experience, knowledge, and perception (Kitahara et al., 2009; Milberg et al., 2008; Paillet et al., 2015). Low agreement among observers may indicate that an indicator is poorly defined, difficult to assess, or requires improved training or calibration (Cherrill, 2016; Gorrod and Keith, 2009; Kelly et al., 2011; Lindenmayer and Likens, 2011; Paillet et al., 2015). Inter-observer reliability, as well as other sources of bias or noise, are therefore important to understand and consider when applying CVAs in decision-making processes (Gorrod and Keith, 2009; Gosselin and Larrieu, 2020).

Inter-observer reliability (IOR) — also referred to as inter-observer agreement (IOA) — quantifies the degree to which multiple observers produce consistent assessments of the same forest indicators and indices across subjects (e.g., forest stands). IOR is a conceptual measure of consistency that can be quantified using various statistical metrics. These include intra-class correlation coefficients (ICC; Koo and Li, 2016) and concordance correlation coefficients (CCC; Lin, 1989), which evaluate consistency for continuous ratings, and Cohen's kappa, which examines agreement in categorical ratings (Cohen, 1968). These methods are widely applied in clinical research, animal behavior, and welfare studies (e.g., Kaufman and Rosenthal, 2009; Kottner et al., 2011) but are rarely applied in ecological field assessments. High IOR indicates that observers are interchangeable, with minimal observer bias, while low IOR suggests variability in interpretation or assessment of indicators and indices. In ecological studies, inter-observer variability has more commonly been evaluated using percentage agreement, coefficients of variation, or correlations between observers (e.g., Goodenough et al., 2020; Gorrod and Keith, 2009; Milberg et al., 2008; Morrison et al., 2016). However, unlike correlational studies, IOR methods provide a more robust evaluation of both the strength of association and the degree of absolute agreement between measurements in any assessment, scoring, or rating system (Kottner et al., 2011), making them particularly suitable for assessing the consistency of qualitative and semi-quantitative CVAs.

The aim of this study was to evaluate the IOR among professional biologists assessing forest conservation values, using both qualitative and semi-quantitative methods across a set of boreal forests. We conducted triple independent surveys of the same 14 forest stands, where each surveyor ('rater') assessed a broad set of conservation indicators and indices. We focused on two commonly applied CVAs in Sweden, both designed for use by forest managers in operational surveys and requiring no or limited taxonomic expertise. The qualitative assessment followed the standardized methodology developed by Drakenberg and Lindhe (1999) and evaluated by Hekkala et al. (2023) and Larsson Ekström et al. (2025), involving standardized presence-absence scoring (yes/no) of 50 conservation indicators adapted to the stand scale, forest type, and region. The semi-quantitative assessment followed the methodology applied across all state-owned forest land in Sweden, involving counts, frequency estimates, ordinal ratings (low to very high), and maximum age estimates of 20 structural and habitat-related indicators within sub-plots.

We hypothesized that IOR would be higher for semi-quantitative assessments than for qualitative assessments, based on the assumption that more structured and quantitative measures provide greater transparency and reduce observer bias compared to binary judgements such as presence-absence scoring (Lindenmayer and Likens, 2011). Specifically, we addressed the following research questions and expectations:

1. *What is the degree of IOR in qualitative and semi-quantitative forest CVAs?* We expected overall IOR to be moderate to high, with higher values for semi-quantitative indicators due to their more structured scoring and reduced subjectivity (Lindenmayer and Likens, 2011).
2. *Do certain groups of indicators and indices — such as summed downed deadwood vs standing structures — show consistently higher or lower IOR across the qualitative and semi-quantitative assessment methods?* We expected indicators related to downed deadwood to show higher IOR, because they are visually conspicuous and less open to interpretation (Gosselin and Larrieu, 2020; Kelly et al., 2011). In contrast, we expected standing structures such as HNV trees and tree microhabitats to show lower IOR due to their rarity and the greater subjectivity involved in identifying features like tree form, age, or small microhabitats (Harper et al., 2004; Kenning et al., 2005; Paillet et al., 2015).
3. *Can IOR analyses be used to refine forest CVA protocols, for example by flagging low-reliability indicators, and identifying needs for observer training and calibration?* We expected that small-n reliability studies could potentially detect observer bias and low-reliability indicators among professional biologists conducting forest CVAs (Kottner et al., 2011; Shoukri et al., 2004). Such analyses should therefore help pinpoint which indicators require clearer definitions, enhanced training, or potential exclusion to strengthen the reliability and interpretability of composite CVA scores (Greco et al., 2019; OECD/European Union/EC-JRC, 2008).

2. Material and methods

2.1. Study forests

We surveyed 14 forest stands located in south-central Sweden, covering a geographic range that corresponds approximately to 59.81°-60.37° N latitude and 13.50°- 16.07° E longitude (WGS84). The study area lies within a relatively homogeneous forest region dominated by managed coniferous stands, within the boreal vegetation zone (Ahti et al., 1968). The study region has a long history of industrial forestry (Angelstam, 1997), and old-growth forest remnants are scarce in the surrounding landscape (Kärvemo et al., 2021). Stands ranged in size from 2.1 to 26.7 ha (mean 8.9 ha). However, all surveys (Sections 2.3–2.4) were conducted within a standardized 2-ha study plot, systematically placed at the center of each stand. This ensured that structural measurements and species surveys were conducted within the same area across all sites and that stand size did not influence the results. Maximum tree age within each stand ranged from 106 to 294 years

(mean 165 years), based on tree-ring counts from increment cores taken at breast height (1.3 m) from four of the oldest-looking trees within each 2-ha plot. The selected stands represented a gradient from structurally simple production forests to woodland key habitats — defined in Sweden as forests of high conservation value, where red-listed species occur or are likely to occur (Nitare and Norén, 1992). Stands were randomly selected from forest owners' (Sveaskog) and Swedish Forestry Agency's databases to represent different conifer-dominated management classes, based on canopy composition: Norway spruce (*Picea abies* L. Karst.) dominated (>65 % spruce, n = 5), Scots pine (*Pinus sylvestris* L.) dominated (>65 % pine, n = 5), and mixed coniferous stands (>65 % combined spruce and pine, n = 4).

2.2. Surveyors

All assessments were conducted in 2018 by triplets of professional biologists within the 2-ha plots at each site. The surveyors were either employed as forest ecology consultants or worked as conservation experts within a forest company, with professional experience in field-based forest assessments. Prior to the surveys, all surveyors participated in preparatory training consisting of: (i) theoretical instruction covering both the qualitative CVA (see Appendix S1) and the semi-quantitative CVA (see Appendix S2), including survey design, indicator definitions, scoring protocols, and overall survey aims; and (ii) a supervised field training session introducing the practical application of both CVAs. No inter-calibration exercises or joint assessments among surveyors were conducted prior to the assessments, and all surveys were performed individually and separately in time, ensuring that the results reflected independent evaluations by each rater. In total, six surveyors conducted surveys (three women and three men). For each CVA method, three surveyors (n = 3) were randomly assigned to perform the assessments within each site, with both genders represented within each surveyor triplet. The use of three observers, rather than pairs, provides a more robust and informative evaluation of reliability, allowing assessment of agreement across multiple pairwise comparisons while capturing variability among raters (Koo and Li, 2016).

2.3. Qualitative conservation value assessments

The qualitative CVA used in this study followed the "Assessment of Forest Biodiversity Potential" method developed by AB Skogsbiologerna (Drakenberg and Lindhe, 1999), also referred to as the "Habitat Heterogeneity Score" (Hekkala et al., 2023). This method is widely applied in Swedish forest management and conservation planning and has also been adapted for use in other countries (Blicharska, 2005, Hekkala et al., 2023; Larsson, 2001; Perhans et al., 2011). The method is based on systematic walk-through surveys of forest stands, where surveyors record the presence (Yes = 1, No = 0) of 50 predefined qualitative conservation indicators. Indicators are grouped into six subsets reflecting different aspects of importance for forest biodiversity: (i) site characteristics, (ii) dynamics, (iii) habitats, (iv) trees, (v) structure, and (vi) deadwood (Table 1; Appendix S1). The set of 50 indicators is adapted to regional conditions and forest type, based on disturbance history and dominant tree species. In this study, we applied the protocol versions for two forest types in boreal regions: Fp = Frequent pine, pine -dominated stands with a history of frequent or intense fire disturbance and deciduous species in later successional stages, and S = Seldom, spruce-dominated stands with small-scale or infrequent disturbance (Appendix S1). Although none of the 14 study stands contained a mix of forest types within stands, the method recommends that, in such cases, only the most homogeneous part of the stand should be assessed (Drakenberg and Lindhe, 1999). The indicators cover a range of features including traces of natural disturbances (e.g., fire scars), habitat elements (e.g., large deadwood), tree characteristics (e.g., large and old veteran trees), and site attributes (e.g., rocky outcrops). Some indicators represent direct habitat structures for species of conservation concern,

Table 1
Overview of the qualitative (Q) conservation value indicators assessed for presence (yes/no) and summed within six conservation value categories, applied to both Frequent pine (Fp) and Seldom (S) forest types (with 50 central questions to each type) in boreal regions (Drakenberg and Lindhe, 1999). The table also shows the data distribution and the statistical method used to analyze inter-observer reliability (IOR) for each group of conservation indicators. Dbh = diameter at breast height (1.3 m). Diameter for downed logs it refers to the largest-end diameter.

Summed values assessed	Qualitative (Q) conservation value indicators	Data distribution /Statistics
Sum Q Range 0–50	All indicators below (50 for each forest type assessed, shown here as 69 for both forest types combined).	Normal/ICC
Sum Site Range 0–12	1. Conspicuously broken terrain/ varied topography 2. Vertical cliff/scree-slope > 10 m high 3. Forested gorge/ravine > 10 m deep 4. Site characterised by S-SW facing slope steeper than 15 % 5. Site characterised by N-NE facing slope steeper than 15 % 6. At least part of the site located above 450 m altitude/pre-alpine 7. Site surrounded by forest/terrain buffering local climate 8. Site characterised by normally wet/ very wet forest 9. Area > 0.1 ha of forested rocky outcrop/ground with very shallow soils 10. Lichens cover > 50 % of the ground 11. Site characterised by a conspicuous herb component/*Ribes*/ *Lonicera* 12. Lime-/hyperite-rich soils/ conspicuous amounts of orchids/ liverwort	Non-normal /CCC
Sum Dynamics Range 0–9	13. Signs of former-recent forest fire on stumps/ trees 14. Several living trees with fire-scars 15. Several living trees with scars from more than one fire 16. Recently burnt area > 0.1 ha with substantial amounts of living/dead trees 17. Spruce constitutes less than 10 % of the stand volume/basal area 18. Several canopy gaps less than 0.1 ha with natural regrowth of main species 19. Site characterized by a thick, continuous moss cover on rocks and boulders 20. Conspicuous signs of woodpecker activity on living trees/dead wood 21. Seasonally flooded area > 0.1 ha in forested surroundings	Non-normal /CCC
Sum Habitat Range 0–12	22. Boulder terrain > 0.1 ha/large boulders > 2 m high 23. A total of > 0.1 ha sandy, sun-exposed, sparsely vegetated ground 24. Shaded > 2 m high conspicuous vertical cliff with a mixed moss cover 25. Area > 0.1 ha of normally wet/ very wet forest 26. Area > 0.1 ha of wet very wet, conspicuously sloping forest 27. Area > 0.1 ha dominated by luxuriant herbs/tufted ferns 28. Forest in contact with open water/ wetland > 0.1 ha 29. Spring/spring brook in forested surroundings	Non-normal /CCC

(*continued on next page*)

Table 1 (*continued*)

Summed values assessed	Qualitative (Q) conservation value indicators	Data distribution /Statistics
	30. Non-seasonal brook/watercourse in forested surroundings 31. As above, and meandering in sand/silt 32. White-water/rapids/waterfall in forested surroundings 33. Conspicuous hollow tree/nest of coarse twigs/several nesting holes	
Sum Trees Range 0–10	34. Several > 2 m high hazels 35. Substantial amounts of > 2 m high junipers/shrubs 36. Occurrence of oak/lime/maple/ash 37. Substantial amounts of aspen/sallow/black alder > 10 cm dbh 38. Substantial amounts of broadleaf trees > 20 cm dbh 39. Several aspen/sallow/black alder > 40 cm dbh 40. Several broadleaf trees > 40 cm dbh 41. Several trees > 40 cm dbh 42. Substantial amounts of trees > 40 cm dbh 43. Several trees > 60 cm dbh	Non-normal /CCC
Sum Structure (e.g., tree microhabitats) Range 0–11	44. Trees characterized by a conspicuous girth/age variation of trees with > 10 cm dbh 45. Several trees stand out as conspicuous older/larger than the stand in general 46. Several trees with conspicuously thick branches and low/wide crowns 47. As above and in open, sun-exposed conditions 48. Substantial amounts of conspicuous retarded/stunted trees/biological old trees 49. Substantial amounts of formerly - recently snow-broken trees > 10 cm dbh 50. Substantial amounts of basally multi-stemmed trees/coppice > 10 cm dbh 51. Substantial amounts of trees on buttresses 52. Several stems with conspicuous occurrences of mixed mosses/lichens/*Lobaria* 53. Several trees with conspicuous occurrences of pendulous lichens 54. Substantial amounts of trees with conspicuous occurrences of pendulous lichens	Non-normal /CCC
Sum Deadwood Range 0–15	55. Conifers; several erect dying/dead trees/> 2 m high stumps > 20 cm dbh 56. As above and in sun-exposed conditions 57. Broadleaves; several erect dying/dead trees/> 2 m high stumps > 20 cm dbh 58. As above and in sun-exposed conditions 59. Substantial amounts of erect dying/dead trees/> 2 m high stumps > 20 cm dbh 60. Several windthrown trees with upturned roots 61. Several rot-broken trees 62. Several downed logs > 20 cm diameter 63. Several downed logs > 20 cm diameter in open sun-exposed conditions 64. Several downed logs > 20 cm diameter with a mixed, partly moss cover 65. Several downed logs > 20 cm diameter in various stages of decay 66. Substantial amounts of downed logs > 20 cm diameter 67. Several downed logs > 40 cm diameter 68. Several trees/stumps/logs with conspicuous occurrences of fungi 69. Subst. amounts of trees/stumps/logs with conspic. occurrences of fungi	Normal/ICC
Sum Q (*-dynamics*) Range 0–41	All 50 indicators minus nine dynamics indicators with low reliability among surveyors	Normal/ICC
Sum Q (*-site*) Range 0–38	All 50 indicators minus 12 site indicators with low reliability among surveyors	Normal/ICC
Sum Q (*-structure*) Range 0–39	All 50 indicators minus 11 structure indicators with low reliability among surveyors	Normal/ICC

while others reflect site heterogeneity or disturbance history (Drakenberg and Lindhe, 1999). Surveyors recorded the presence of conservation value indicators based on definitions provided in the original method, where terms such as "several" (on average more than two per hectare), "substantial amount of" (readily noticeable without active search), and "conspicuous" (clearly visible or eye-catching, characteristic of the stand) were used to guide assessments (Drakenberg and Lindhe, 1999). All 50 presence-absence indicators were recoded as Yes = 1 and No = 0 and then tallied to create the composite metric Sum Q (Table 1; possible range = 0–50), where higher scores indicate higher habitat richness and conservation value (Drakenberg and Lindhe, 1999; Hekkala et al., 2023). In practice, scores above 30 are rare and considered indicative of stands with particularly high conservation value. The same tallying procedure was applied to thematic subsets of indicators (site, dynamics, habitat, trees, structure, and deadwood), as well as to three modified totals that exclude dynamics, site, or structure indicators. This yielded subset scores such as Sum site and Sum structure (Table 1). Each subset score ranges from 0 to n for that group (e.g., 0–11 for structure indicators), with higher values indicating a greater richness of features within that theme. This qualitative assessment is designed to be simple, rapid, and applicable in operational forest management without the need for specialized taxonomic expertise. However, the final CVA score is a general proxy of biodiversity potential and should be interpreted with caution in management decisions, considering additional site-specific factors such as location, size, rarity, and conservation potential (Drakenberg and Lindhe, 1999).

2.4. Semi-quantitative conservation value assessments

The semi-quantitative CVA applied in this study was developed by several Swedish forest companies, including the state-owned Sveaskog, to support responsible forestry practices and fulfil Forest Stewardship Council (FSC) certification requirements. The method aims to assess whether a forest stand holds sufficiently high conservation value to warrant protection from harvesting (Sveaskog, unpublished protocol; Appendix S2). The assessment combines structural measurements in circular field plots with stand-scale estimates of habitat features. Within each 2-ha study plot, surveyors established one or more circular subplots with a radius of 25 m (≈0.2 ha) depending on forest heterogeneity, following guidelines to achieve a representative sample of the stand (Appendix S2 Section 1.3–1.4). In our study case, subplot placement was

determined independently by each surveyor without coordination, meaning the exact locations of the subplots varied between observers. This approach reflects operational field conditions for the 2-ha study forest conditions (Appendix S2 Sections 1.3–1.4) but may introduce spatial variability as a potential source of inter-observer differences. Within each subplot, surveyors counted the number of: (i) high nature value (HNV) trees — defined as trees clearly distinct in age, size, or growth form; (ii) high stumps and snags (>15 cm diameter at breast height); and (iii) downed logs (>15 cm diameter at breast height). Counts were scaled to per-hectare values by multiplying by five. Surveyors also visually estimated (iv) the maximum tree age of dominant species within the inventory plots. In addition to these plot-level counts, surveyors assessed 16 conservation value indicators across the inventory plots. These indicators included aspects such as tree continuity, deadwood continuity, stand age, topography, ground conditions, water environment habitats, forest dynamics, biotopes, and cultural or recreational values. Each indicator was scored on a five-point ordinal scale, where 1 indicated very high conservation value and 5 indicated low or no conservation value, based on detailed criteria provided in the assessment protocol (Appendix S2). Although this scale runs in descending order (1 = very high, 5 = low or no value), it was retained to remain consistent with the established assessment protocol.

Forest companies applying this method use the indices and structural measurements collectively to guide conservation value classification and management decisions. Stands with multiple high or very high indices, or exceptionally high values for individual indices (e.g., very high values for HNV trees or volumes of downed deadwood), are typically classified as set-aside for conservation. In this study, we summarized mean counts and frequencies of HNV trees and deadwood, maximum estimated tree ages, and the summed number of indicators assessed to be falling into different groupings of very high (1), high (2), or medium (3) conservation value, for IOR analyses (Table 2; Appendix S2 Sections 2.3–2.4). For 11 out of the 16 individual indicators, we analyzed the IOR in the assignment of ordinal values among surveyors (Table 2).

2.5. Statistical analyses of inter-observer reliability

Inter-observer reliability (IOR) was assessed using statistical methods appropriate for both continuous and ordinal-scale conservation value indices, following established recommendations for evaluating agreement among raters (Hallgren, 2012; Koo and Li, 2016; Lin, 1989). For conservation value indices with approximately normal distributions, we calculated intra-class correlation coefficients (ICCs) with 95 % confidence intervals to assess agreement among surveyors. Analyses were conducted in R (v4.2.0; R Development Core Team, 2022), using the package irr (Gamer et al., 2019). ICCs were calculated as single-rating, absolute agreement measures from a two-way random-effects model, treating forest stands as repeated units and surveyors as random factors (Hallgren, 2012; Gamer et al., 2019; Koo and Li, 2016). This approach was chosen because surveyors were randomly drawn from a larger pool of professional forest ecologists (i.e., generalizable to surveyors who possess the same competences), and all surveyors assessed the same 14 forest stands, themselves randomly selected from a larger population of conifer-dominated stands in the region. The absolute agreement model was applied since the aim was to assess similarity in scores, not just consistency in rank order (correlation). The ICCs reflect both the strength of agreement and the magnitude of disagreement, with larger differences among raters resulting in lower ICC values (Hallgren, 2012). Normality of indices was assessed visually and with Anderson-Darling tests. Where possible, variables were log-transformed (log(1 +x)) to improve normality (Table 2). For indices that remained non-normally distributed despite transformation, we applied the concordance correlation coefficient (CCC) to assess agreement between pairs of surveyors. The CCC measures how closely observations conform to the 45° line of perfect agreement, combining both precision (Pearson correlation) and

Table 2
Overview of the semi-quantitative (SQ) indices and indicators assessed within inventory plots of the forest stands. The upper part of the table summarizes the main conservation value indices, including a short description, data distribution, and the statistical method used to analyses inter-observer reliability (IOR). Sum SQ values represent the total number of individual indicators receiving scores in the specified range of ordinal values: 1 = very high, 2 = high, or 3 = medium conservation value. The lower part lists individual indicators, with corresponding descriptions, data distributions, and IOR analysis methods. Indicators analyzed individually are marked in bold. Dbh = diameter at breast height (1.3 m); HNV = High Nature Value. Detailed descriptions of all indices, indicators, and scoring criteria are provided in Appendix S2.

Conservation values assessed	Short description of indices	Data distribution /Statistics
Sum SQ medium-very high	Total number of individual indicators listed below scored with ordinal values 3, 2 *or* 1	Normal log (1 +data)/ICC
Sum SQ high-very high	Total number of individual indicators listed below scored with ordinal values 2 *or* 1	Normal log (1 +data)/ICC
Sum SQ very high	Total number of individual indicators listed below scored with ordinal value 1	Non-normal /CCC
Max age	Estimated average maximum age (years) from the oldest trees	Normal log (1 +data)/ICC
Number of HNV trees	Average number of trees per ha that are clearly deviant in age, diameter and growth form	Non-normal /CCC
No. standing deadwood	Average number of high stumps and snags > 15 cm dbh per ha	Normal/ICC
No. downed deadwood	Average number of logs > 15 cm dbh per ha	Normal/ICC
No. deadwood	Average total number of deadwood > 15 cm dbh per ha	Normal/ICC
Conservation value indicators assessed	**Short description of conservation indicator, scored with ordinal scale values from 1−5**	**Data distribution /Statistics**
HNV trees I	Trees that are clearly deviant in age, diameter and growth form. Examples of HNV trees are old trees, trees with nest holes, birds-of-prey nests, rotted trees, trees richly draped with pendulous lichens, and trees with clear fire scars	Weighted Fleiss' kappa
HNV trees II (very high values)	Trees that are exceptionally large and old (c. 40−60 cm dbh, c. 200−300 years old)	Weighted Fleiss' kappa
Tree species composition	Presence of rare tree species, high tree species richness, e.g. deciduous trees	Weighted Fleiss' kappa
Deadwood I	Logs > 15 cm dbh and of different tree species	Weighted Fleiss' kappa
Deadwood II (very high quality)	Very large (c. 30 cm dbh), decayed logs, and logs in many decay classes	Weighted Fleiss' kappa
Continuity	Continuity of the forest stand (for at least two tree generations), deadwood (logs for at least 50−100 years), and trees (e.g. old pine trees)	Weighted Fleiss' kappa
Forest stand age	Estimated age of the oldest tree layer in the forest stand	Weighted Fleiss' kappa
Topography	Calcareous rocks and ground, high humidity/special microclimate, rich sediments, boulders, ravines	Weighted Fleiss' kappa
Ground conditions	Calcareous, nutrient-rich, low shrubs, litter, or lichen ground conditions with/without red-listed species, and vegetation of nutrient-rich ground conditions	**Not analyzed**
Water environment/ aquatic conditions	Red-listed aquatic species, springs, wet forests, lake outlets, flooded old forest, pristine running water	**Not analyzed**
Dynamics, natural disturbances and processes	Fire/burned areas, flooded forest areas, storm canopy gaps, and natural succession in young/old forest	Weighted Fleiss' kappa

(*continued on next page*)

Table 2 (*continued*)

Conservation values assessed	Short description of indices	Data distribution /Statistics
Biotopes	Larger areas (>0.5–1 ha) of sensitive biotopes; flooded forests and forested rocky outcrops. Rare biotopes or landscape-level valuable forest type in young/old forest	**Not analyzed**
Cultural values	Meadows with/without species, cultural remains and values	Weighted Fleiss' kappa
Recreational values	Forests close (within 300 m) to schools and other public places, recreational (urban) forests, visited places, trails.	**Not analyzed**
Reindeer herding values	Particularly important or significant cultural remains or places of worship, values for moving, resting and food resources (lichen biomass) for reindeers	**Not analyzed**
Species	Sensitive interior old-growth forest indicator species (list Swedish Forestry Agency), light-demanding red-listed species, and red-listed species	Weighted Fleiss' kappa

accuracy (closeness to the identity line). It is particularly suitable for smaller and non-normally distributed data, where ICC assumptions may be violated. The CCC values were calculated using the epi.ccc() function in the epiR package (Stevenson et al., 2021) in R. We interpreted ICC and CCC estimates following Koo and Li (2016): < 0.50 = poor reliability, 0.50–0.75 = moderate reliability, 0.75–0.90 = good reliability, and > 0.90 = excellent reliability.

For ordinal-scale indices (e.g., conservation value ratings from low to very high for individual conservation indicators), we calculated weighted kappa coefficients (Kw; Cohen, 1968) with 95 % confidence intervals using the R package irrCAC (Gwet, 2022). Weighted Fleiss' kappa provides a measure of agreement (%) among multiple raters while penalizing disagreements according to their severity; taking the order of categories into account (Cohen, 1968). We applied a quadratic weighting scheme, which increases penalties for larger discrepancies in ratings (Brenner and Kliebsch, 1996). Kw values were interpreted following Landis and Koch (1977): < 0 = poor, 0.00–0.20 = slight, 0.21–0.40 = fair, 0.41–0.60 = moderate, 0.61–0.80 = substantial, and 0.81–1.00 = almost perfect agreement. All data visualizations were produced using the R package ggplot2 (Wickham, 2016).

3. Results

3.1. Inter-observer reliability in qualitative conservation value assessments

We assessed IOR for the qualitative conservation value assessments based on (i) the total summed score of all 50 indicators, and (ii) the summed scores for the six conservation value subsets: deadwood, dynamics, habitats, site, structure, and trees. Overall, agreement among surveyors for the composite sum of all 50 qualitative indicators was good (ICC = 0.84; Fig. 1A). The summed subset of deadwood indicators also showed moderate agreement (ICC = 0.72; Fig. 1A). Excluding the dynamics, site or structure indicator subsets, which showed the lowest reliability (Fig. 1B), resulted in a marginal decrease (minus dynamics or site) or increase (minus structures) in overall IOR for the composite score (ICCs = 0.81–0.85; Fig. 1A) compared to the full set of conservation indicators (ICC = 0.84). The pairwise agreement between surveyors for Sum dynamics was consistently poor across all surveyor pairs, with CCC values ranging from 0.07 to 0.51, indicating substantial variability in assessments (Fig. 1B). Sum site and Sum structures both showed poor to moderate agreement, with CCC values ranging from 0.34 to 0.65. In contrast, agreement was generally higher for Sum habitat and Sum trees, with CCC values between 0.52 and 0.81, indicating moderate to good agreement, depending on the pair. Visual inspection of indicator scores showed that discrepancies among surveyors varied across indicator

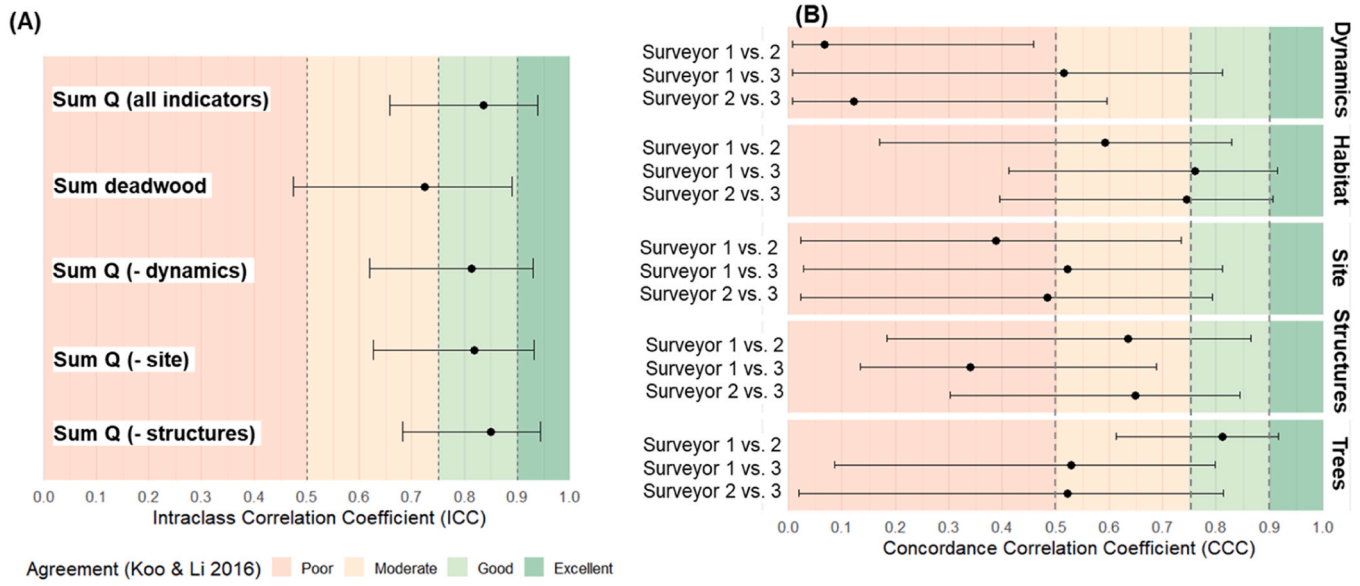


Fig. 1. Inter-observer reliability (IOR) among three surveyors assessing forest conservation values across 14 forest stands. (A) Intra-class correlation coefficients (ICC; points) with 95 % confidence intervals (whiskers) for normally distributed quantitative indicators: total sum of all 50 qualitative indicators (Sum Q), sum of deadwood indicators (Sum deadwood), and total Sum Q excluding dynamics indicators (Sum Q - dynamics), site indicators (Sum Q - site), respective structure indicators (Sum Q - structure). (B) Concordance correlation coefficients (CCC; points) with 95 % confidence intervals (whiskers) showing pairwise agreement between surveyors for five non-normally distributed indicator sums: dynamics, habitat, site, structures, and trees. Vertical dashed lines indicate threshold values for poor (<0.50), moderate (0.50–0.75), good (0.75–0.90), and excellent (>0.90) agreement (Koo and Li, 2016). Surveyor identifiers (Surveyor 1, 2, and 3) are anonymized and were assigned arbitrarily; they do not correspond to the same individuals across assessment methods (qualitative and semi-quantitative) as shown in Figs. 1 and 2.

subsets and were not consistently attributable to the same individual surveyor (Appendix S3; Fig. S1).

3.2. Inter-observer reliability in semi-quantitative assessment

The IOR for the semi-quantitative conservation value assessments was evaluated based on (i) eight semi-quantitative indices (e.g., counts or total number of conservation indicators), and (ii) eleven individual conservation indicators assessed on an ordinal five-step scale (Table 2). Agreement among surveyors for the total number of indicators rated as medium, high, or very high (Sum SQ medium-very high) was good (ICC = 0.80; Fig. 2A). When the analysis was restricted to the number of indicators rated as high or very high (Sum SQ high-very high), inter-observer reliability (IOR) remained moderate to good (ICC = 0.73). For the total number of indicators rated as very high (Sum SQ very high), agreement between Surveyor 1 and Surveyor 3 was the strongest (CCC = 0.85, 95 % CI: 0.62–0.95), which falls within the moderate to good agreement range (Fig. 1B). The agreement between Surveyor 1 and Surveyor 2 was moderate (CCC = 0.49, 95 % CI: 0.03–0.78), while the lowest agreement was found between Surveyor 2 and Surveyor 3 (CCC = 0.24, 95 % CI: 0.00–0.64), indicating poor agreement. Among the structural indices, the number of downed deadwood units per hectare showed moderate-to-good agreement (ICC = 0.74), while estimates of maximum tree age (ICC = 0.66) and the total number of deadwood units per hectare (ICC = 0.61) showed moderate agreement. Lower IOR was found for the number of standing deadwood objects (ICC = 0.50; Fig. 2A). For the HNV tree counts, the highest agreement was again found between Surveyor 1 and Surveyor 3 (CCC = 0.87, 95 % CI: 0.64–0.95), which represents good agreement. The agreement between Surveyor 1 and Surveyor 2 was poor to moderate (CCC = 0.46, 95 % CI: 0.11–0.71), and the agreement between Surveyor 2 and Surveyor 3 was similarly poor (CCC = 0.39, 95 % CI: 0.04–0.65). Visual inspection of the semi-quantitative data indicated that assessment discrepancies varied across indices and indicators, without a consistent pattern of deviation linked to any single surveyor (Appendix S3; Fig. S2).

3.3. Inter-observer reliability in classifying individual ordinal-scale indicators

Inter-observer reliability (IOR) for individual ordinal-scale indicators, each rated on a five-step conservation value scale (from low to very high), varied substantially across indicators (Fig. 3). Weighted Fleiss' kappa coefficients (Kw) ranged from 0.00 to 0.72, corresponding to IOR levels from slight to substantial agreement. The highest agreement among surveyors was observed for indicators related to deadwood, forest stand age, and continuity, all of which showed substantial IOR (Kw > 0.62). In contrast, conservation indicators related to species values, tree species composition, and very high-quality high nature value (HNV) trees II showed the lowest agreement, with slight IOR (Kw = 0.00–0.20). Inter-observer reliability was markedly lower for rarer, high-quality substrates (HNV trees II and deadwood II), falling up to two reliability threshold categories below that of more common counterparts (HNV trees I and deadwood I) (Fig. 3). Four indicators — ground conditions, water environments/aquatic habitats, recreational values, and biotopes — were not individually analyzed for IOR due to their consistently low conservation value scores across all sites. However, these indicators were assessed with near-complete agreement among all surveyors, suggesting minimal variation in classification.

4. Discussion

A major challenge for forest managers is to ensure consistency in conservation value assessments (CVAs) among surveyors, as poor agreement may lead to inappropriate management decisions or misclassification of valuable forest stands. This study provides a case study for utilizing triple surveys to assess inter-observer reliability (IOR) in CVAs, showing that even relatively simple qualitative presence-absence scorecard methods can achieve good reliability across professional surveyors. Contrary to our initial expectation, semi-quantitative assessments did not consistently yield higher IOR than qualitative assessments. Rather, the reliability of both methods was comparable — and in some cases slightly higher for qualitative indices — suggesting that IOR may depend more on the characteristics of specific indicators (e.g., downed vs standing substrate indicators) than on the overall assessment method used. This study illustrates that even limited inter-observer evaluations, such as our triplet surveyor approach, can serve as a useful tool for understanding consistency in CVAs and refining CVA protocols. Given that this was a smaller, exploratory case study focused on boreal conifer-dominated forest assessments and professional biologists, further research is needed to expand on these findings by identifying sources of IOR and evaluating strategies to enhance consistency in CVAs across different regions, forest types, seasons, and surveyor groups.

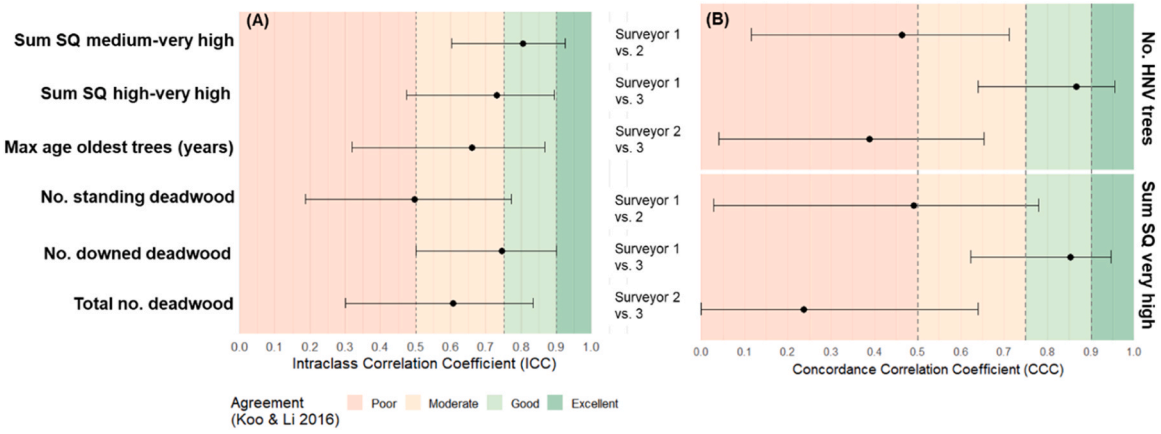


Fig. 2. Inter-observer reliability (IOR) among three surveyors assessing semi-quantitative conservation values across 14 forest stands. (A) Intra-class correlation coefficients (ICC; points) with 95 % confidence intervals (whiskers) for normally distributed indicators: number of indicators scored to have medium to very high values, number of indicators scored to have high to very high values, maximum estimated tree age, number of standing deadwood per hectare, number of downed deadwood per hectare, and total number of deadwood units per hectare. (B) Concordance correlation coefficients (CCC; points) and 95 % confidence intervals (whiskers) for pairwise agreement between surveyors on two non-normally distributed indices: number of High Nature Value (HNV) trees per hectare and number of structural indicators scored as having very high values. Vertical dashed lines indicate threshold values for poor (<0.50), moderate (0.50–0.75), good (0.75–0.90), and excellent (>0.90) agreement (Koo and Li, 2016). Surveyor identifiers (Surveyor 1, 2, and 3) are anonymized and were assigned arbitrarily; they do not correspond to the same individuals across assessment methods (qualitative and semi-quantitative) as shown in Figs. 1 and 2.

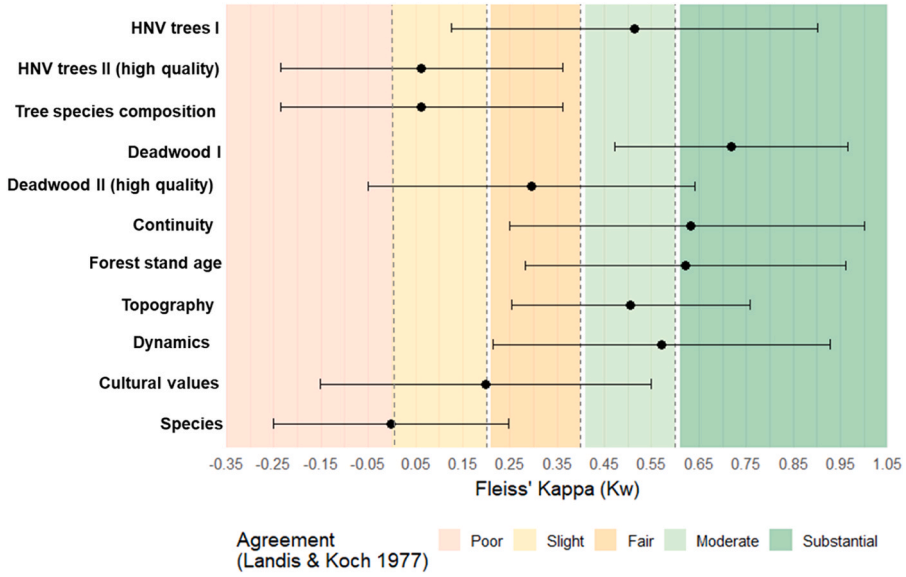


Fig. 3. Inter-observer reliability (IOR) among three surveyors classifying conservation values for eleven semi-quantitative ordinal-scale indicators across 14 forest stands. Weighted Fleiss' kappa coefficients (Kw; points) with 95 % confidence intervals (whiskers) are shown. Vertical dashed lines indicate threshold values for poor (Kw = <0.00), slight (0.00–0.20), fair (0.21–0.40), moderate (0.41–0.60), and substantial (>0.60) agreement following Landis and Koch (1977). The HNV trees I include structurally distinct trees, such as old trees and those with nest holes, raptor nests, fire scars, or draped in pendulous lichens. HNV trees II are exceptionally large and old (approx. 40–60 cm dbh, 200–300 years). Deadwood I refers to logs > 15 cm dbh from various species, while Deadwood II includes very large (~30 cm dbh), decayed logs representing multiple decay stages.

4.1. Qualitative vs semi-quantitative assessments - strengths and challenges

Our findings that summed scores of multiple qualitative or semi-quantitative indicators similarly achieved good IOR are encouraging, given the widespread use of both methods in forest management. This study provides novel evidence that well-structured qualitative CVAs — based on the presence-absence of multiple conservation indicators assessed across an entire forest stand — can achieve reliability levels comparable to semi-quantitative methods, which focus on smaller subplots and more measurable indicators. Our results align with previous research evaluating the French Index of Biodiversity Potential (IBP), which also found relatively low to moderate between-observer variation across multiple forest conservation indices (Gosselin and Larrieu, 2020). Taken together, these findings suggest that observer reliability in quantitative and semi-quantitative forest CVAs can be relatively high — provided that survey protocols are well-defined, and surveyors are well-trained.

Smaller differences in IOR between methods likely reflect different mechanisms for introducing observer error. Semi-quantitative approaches offer greater transparency through standardized counts or measurements within plots but may be sensitive to the placement of plots relative to forest heterogeneity, or to variability in how surveyors detect, identify, or count specific structures. In contrast, qualitative presence-absence surveys distribute observation effort across the whole stand, reducing the risk of missing uncommon features, but potentially increasing variation where indicator definitions are less clear, or when inclusion criteria are interpreted differently by surveyors. Similar patterns have been observed in vegetation surveys, where presence-absence recording yielded more consistent results among observers than visual cover estimates or point-frequency methods (Ringvall et al., 2005). In our study, the exact placement of subplots for semi-quantitative assessments was independently determined by each surveyor. This means that observed differences may partially reflect real spatial heterogeneity within stands, rather than solely differences in observer judgement. This design choice reflects common practice in forest inventories, but it likely contributed to lower IOR for patchily distributed indicators, and should be explicitly considered in future research. Overall, our findings highlight that both qualitative and semi-quantitative CVAs can provide a reliable basis for assessing forest conservation values — but that minimizing observer bias requires careful attention to protocol design, indicator clarity, and surveyor training, regardless of method used.

4.2. Indicator-specific reliability and observer bias

Across both assessment methods, deadwood indices consistently stood out for their relatively high inter-observer reliability (IOR), in line with previous studies highlighting deadwood as a relatively robust and easily distinguishable conservation feature (Gosselin and Larrieu, 2020; Kelly et al., 2011). However, this pattern was less consistent for deadwood type II — very large, decayed logs with high conservation value — which showed lower (fair) IOR, likely due to their rarity and greater variation in decay stage interpretation. Indicators involving rare standing structures, such as high nature value (HNV) trees in semi-quantitative assessments or the structure (tree microhabitats) subgroup in qualitative assessments, showed even lower agreement among observers. These features often require subjective judgments of tree quality, growth form, or rare characteristics, increasing the risk of inconsistent assessments (Gosselin and Larrieu, 2020; Kelly et al., 2011). Tree microhabitats associated with HNV trees — including hollows, woodpecker cavities, cracks, and bark characteristics — are particularly challenging to survey reliably, and are prone to underestimation unless standardized protocols, sufficient time, and calibration between surveyors are applied (Harper et al., 2004; Paillet et al., 2015). This is reflected in our results, where the reliability of ordinal scoring was consistently lower for indicators involving species values, tree species composition (rare tree species), or exceptionally old or large HNV trees II or deadwood II, often only reaching slight to fair agreement. In contrast, their more common counterparts - HNV trees I and deadwood I — showed moderate to substantial IOR. Similarly, within the qualitative CVA, surveyors showed moderate-to-good agreement when scoring habitat and tree indicators, whereas significant differences emerged in the scoring of the dynamics, site, and substrate indicators subgroup — which included several subjective features related to forest disturbance

histories, site/topographic conditions, tree quality and microhabitats. These patterns reinforce the need for clearer definitions, improved field protocols, and targeted training for indicators that are particularly subjective or rare, such as exceptionally large, old trees or highly decayed deadwood. Another consistent finding was that counts of standing deadwood (high stumps and snags) were more difficult to assess reliably than downed deadwood within the semi-quantitative CVA. This corroborates previous studies reporting that snag inventories are particularly prone to observer variability, especially during the leaf-on season when visibility is reduced (Gosselin and Larrieu, 2020; Kenning et al., 2005). While overall IOR were often very variable, several indicator groups exhibited reasonably high agreement only in the semi-quantitative assessment, such as those related to forest dynamics and site/topography — likely because these features are more readily understood and observed using this method. Conversely, the low IOR observed for the species indicator in the semi-quantitative assessment — which requires integrating information on species of conservation concern — was unsurprising, given the well-documented difficulties and high observer error in species detection and inventories (e.g., Archaux et al., 2006; von Hirschheydt et al., 2024; Lõhmus, 2009; Morrison, 2016).

4.3. Implications for improving CVA protocols

Composite CVA indices have been criticized for potentially being misleading or overly subjective — not only due to observational errors and data quality issues, but also because aggregating diverse indicators into grouped indices can obscure important underlying patterns or drivers (Greco et al., 2019; OECD/European Union/EC-JRC, 2008). The initial selection of indicators is itself a critical step, often shaped by expert judgement, with the potential for bias that can strongly influence composite outcomes. Therefore, understanding how uncertainty and observer variability in individual indicators propagate into composite CVA scores is essential for interpreting results and making informed management decisions. Our findings highlight the value of inter-observer reliability (IOR) analyses as a practical tool for refining CVA protocols. Even relatively simple steps — such as implementing triple-assessor surveys and testing for indicator-specific variation — can improve overall consistency and efficiency in field assessments. While the summed scores (e.g., Sum Q and Sum SQ medium-very high) showed relatively high inter-observer agreement, we acknowledge that these metrics reflect aggregated presence-absence or ordinal responses. Two surveyors may arrive at similar summed scores while identifying different sets of indicators, potentially masking discrepancies at the individual indicator level. This averaging effect may lead to an overestimation of agreement, particularly in qualitative assessments with many indicators. We also note, however, that the use of summed indices reflects standard practice in operational CVAs and forest monitoring protocols. Our approach is therefore aligned with how such assessments are typically applied in forest management settings, making it relevant for evaluating both methodological and applied implications. To address this limitation, we conducted additional analyses at finer resolution, including IOR calculations for indicator subsets (e.g., deadwood, structure) in the qualitative CVA, as well as ordinal scoring consistency across individual indicators in the semi-quantitative CVA. These analyses offer a more nuanced picture of where agreement breaks down and highlight the importance of combining aggregate and disaggregated approaches when evaluating the robustness of CVA protocols. In this study, excluding a subset of structural qualitative indicators with low IOR modestly improved overall agreement without compromising the assessment's holistic character. However, indicators that are difficult to assess — such as tree microhabitats or rare, very high-quality substrates — should not automatically be discarded. Instead, their inclusion requires appropriate calibration, clear protocols, and potentially more intensive training of surveyors to reduce subjectivity and error (Lõhmus et al., 2018; Morrison, 2016; Paillet et al., 2015). IOR analyses can thus play a key role in identifying where field manuals, protocols, or surveyor training needs to be refined — particularly for indicators prone to subjectivity, rare occurrence, or challenging detection. Where feasible, two-stage surveys or guided sampling approaches, including the use of remote sensing data, may further enhance the detection of rare features (Ringvall et al., 2007). Alternatively, when statistical models are used to analyze CVA data, explicit estimates of observer error could be incorporated (Paillet et al., 2015). Awareness of indicator-specific reliability is also crucial for interpreting CVA results, and for weighting evidence in management decisions. Indicators with consistently low IOR should be treated cautiously, and interpreted in the context of their uncertainty, or alternatively supported by complementary evidence from other sources.

Given the lack of IOR analyses in forest CVAs, and the exploratory nature of our study, we did not perform a formal sample size analysis. Nevertheless, we consider the dataset — comprising 42 values per composite metric (from 14 forest stands assessed by 3 observers, totaling over 2900 individual indicator or index observations) — to provide a robust basis for identifying patterns of observer bias. Still, our study is limited in scope, focusing on a small group of surveyors, conducting the surveys in a single region and two conifer-dominated forest types. Future research should extend this approach to a broader range of observers, forest types, and geographical settings, to explore how knowledge, experience, and habitat complexity influence observer reliability (e.g., Goodenough et al., 2020). Importantly, our study focused on trained professional biologists, who typically detect more features, measure more precisely, and show less variation in assessments than non-specialists (Butt et al., 2013; Gosselin and Larrieu, 2020; von Hirschheydt et al., 2024; Paillet et al., 2015). Whether similar levels of IOR would be obtained among less experienced or volunteer surveyors remains an open question. Although investigating the same forest stand by multiple surveyors is resource-demanding, smaller-scale assessments of IOR, like our triplet survey, can offer valuable insights into the degree and nature of observer bias. This approach is commonly applied in clinical, behavioral, and welfare research conducted by skilled professionals (e.g., Kaufman and Rosenthal, 2009; Kottner et al., 2011; Shoukri et al., 2004), where it has been shown that the precision of ICC estimates is optimized with only 2–3 observers per subject when true ICC is reasonably high (>0.60; Shoukri et al., 2004). In our study, multiple indices reached moderate-to-good agreement (ICC > 0.6), suggesting that triplet observer surveys can be a useful tool for testing CVA protocols, identifying patterns of observer bias, and informing protocol improvements — without requiring very large sample sizes. At the same time, the 95 % confidence intervals were often wide, sometimes spanning multiple categories of agreement. This imprecision reflects the limited number of triplets (n = 14) and indicates that our protocol yields informative but noisy estimates of congruence. Triplet surveys therefore remain a cost-effective screening tool for identifying clear cases of low or high observer agreement, but finer distinctions will require either more stands, additional observers, or repeated assessments to narrow the confidence intervals. Extending the design to more observers, forest types, and geographical settings would improve precision and help generalize these findings.

4.4. Conclusions and future research

This study demonstrates that IOR varies substantially across commonly used forest conservation indices, regardless of whether qualitative or semi-quantitative methods are employed. While many indicators and indices can be assessed with good consistency by professional biologists, others — particularly those involving rare structural features or species of high conservation value — showed low agreement and may require clearer definitions, more detailed protocols, or enhanced surveyor training. Given the relatively limited application of IOR analyses in ecological fieldwork, our findings highlight the value of such approaches for improving the robustness, transparency, and credibility of conservation value assessments. Even small-scale IOR

evaluations — such as the triplet survey design used in this study — can reveal important patterns of observer bias, guide improvements in indicator selection and protocol design, and ultimately support more evidence-based decision-making in sustainable forest management. Future research should extend this work across different forest types, geographical regions, and surveyor groups to better understand how factors, such as experience, training, survey context, and habitat complexity influence observer agreement. Additionally, further development of methods to quantify and incorporate observer error into composite index frameworks could enhance the interpretation and use of CVA results in conservation planning. Our analysis relied on standard simpler frequentist metrics of inter-observer reliability (ICC, CCC, and weighted kappa), which offer interpretable and widely used measures of agreement. However, we acknowledge that Bayesian hierarchical models are increasingly applied to CVA data to model observer effects and uncertainty more explicitly (e.g., Gosselin and Larrieu, 2020; Paillet et al., 2015). Given our limited sample size, such models were not feasible here, but future studies with larger datasets could benefit from Bayesian approaches that allow integration of observer-level covariates and generate full posterior distributions of agreement.

Beyond statistical evaluation, it is crucial to ask whether the observed variation has practical consequences for forest management decisions (Cherrill, 2016; Goodenough et al., 2020). In our dataset, differences in CVA scores among surveyors sometimes spanned multiple agreement categories, potentially crossing the operational threshold used in Swedish practice to flag "high conservation value." In such borderline cases, an IOR that appears moderate on paper, or a disagreement of one or more ordinal classes for semi-quantitative indicators, could translate into opposing management prescriptions — set-aside vs. retention forestry vs. harvest. These examples underscore that even small errors may have significant consequences when scores cluster near decision thresholds. We therefore recommend that future IOR studies complement reliability metrics with simple decision-sensitivity analyses — e.g., "Does observed disagreement change the stand's classification?" — to support forest managers to assess the real-world risk of misclassification and determine whether additional training, double-checking, or protocol refinements are needed.

CRediT authorship contribution statement

Anne-Maarit Hekkala: Writing – review & editing, Data curation, Conceptualization. **Mari Jönsson:** Writing – review & editing, Writing – original draft, Visualization, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation. **Louis Mielke:** Writing – review & editing. **Simon Kärvemo:** Writing – review & editing. **Julia Kyaschenko:** Writing – review & editing. **Karina Clemmensen:** Writing – review & editing. **Joachim Strengbom:** Writing – review & editing, Project administration, Investigation, Funding acquisition, Data curation, Conceptualization. **Jörgen Sjögren:** Writing – review & editing, Funding acquisition, Conceptualization.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the author(s) used ChatGPT (OpenAI) for language editing and readability improvements. After using this tool, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

Declaration of Competing Interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Joachim Strengbom reports financial support was provided by Swedish Research Council Formas.

Acknowledgement

We thank the forest company Sveaskog for providing access to their forests and forest-stand database and for conducting part of the surveys, Olle Kellner and Neil Cory from the Swedish Forest Agency for providing some of the structural data and access to woodland key habitat register, and biologists at Greensway AB for conducting part of the surveys. This work was supported by the Swedish Research Council for Environment, Agricultural Sciences and Spatial planning, Formas (2016–20029).

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.foreco.2025.123006.

Data availability

Data will be made available on request.

References

Ahti, T., Hämet-Ahti, L., Jalas, J., 1968. Vegetation zones and their sections in northwestern Europe. Ann. Bot. Fenn. 5, 169–211.
Angelstam, P., 1997. Landscape analysis as a tool for the scientific management of biodiversity. Ecol. Bull. 46, 140–170.
Archaux, F., Gosselin, F., Bergès, L., Chevalier, R., 2006. Effects of sampling time, species richness and observer on the exhaustiveness of plant censuses. J. Vege. Sci. 17, 299–306. https://doi.org/10.1111/j.1654-1103.2006.tb02449.x.
Bellamy, C., Rattey, A., Edwards, C., Kortland, K., Stringer, A., Tew, E., Bathgate, S., Kerecsenyi, N., Moseley, D., Watts, K., Broome, A., 2024. The forest biodiversity index (FOBI): monitoring forest biodiversity potential over space and time. Environ. Res. Ecol. 3, 035001. https://doi.org/10.1088/2752-664X/ad57cf.
Berglund, H., Kuuluvainen, T., 2021. Representative boreal forest habitats in northern Europe, and a revised model for ecosystem management and biodiversity conservation. Ambio 50, 1003–1017. https://doi.org/10.1007/s13280-020-01444-3.
Blicharska, M., 2005. Using a Swed. For. Biodivers. Assess. Pol. Cond. Inst. F. öR. skogens Prod. och Markn. Slu. Available from: ⟨https://stud.epsilon.slu.se/11611/1/blicharska_m_171002.pdf⟩.
Brenner, H., Kliebsch, U.L., 1996. Dependence of weighted kappa coefficients on the number of categories. Epidemiology 7, 199–202. ⟨https://www.jstor.org/stable/3703036⟩.
Butt, E., Slade, J., Thompson, Y., Malhi, T., Riutta, T., 2013. Quantifying the sampling error in tree census measurements by volunteers and its effect on carbon stock estimates. Ecol. Appl. 23, 936–943. https://doi.org/10.1890/11-2059.1.
Cherrill, A., 2016. Inter-observer variation in habitat survey data: investigating the consequences for professional practice. J. Environ. Plann. Manag. 59, 1813–1832. https://doi.org/10.1080/09640568.2015.1090961.
Cohen, J., 1968. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. Psychol. Bull. 70, 213–220. https://doi.org/10.1037/h0026256.
Dieler, J., Uhl, E., Biber, P., et al., 2017. Effect of forest stand management on species composition, structural diversity, and productivity in the temperate zone of Europe. Eur. J. For. Res. 136, 739–766. https://doi.org/10.1007/s10342-017-1056-1.
Drakenberg, B., Lindhe, A., 1999. Indirekt naturvärdesbedömning på beståndsnivå – en praktiskt tillämpbar metod. Skog Forsk. 2, 60–66.
Gamer, M., Lemon, J., Singh, I.F.P., 2019. irr: various coefficients of interrater reliability and agreement. R. Package Version 0. 84. 1. ⟨https://CRAN.R-project.org/package=irr⟩.
Gao, T., Nielsen, A.B., Hedblom, M., 2015. Reviewing the strength of evidence of biodiversity indicators for forest ecosystems in Europe. Ecol. Indic. 57, 420–434. https://doi.org/10.1016/j.ecolind.2015.05.028.
Goodenough, A.E., et al., 2020. The impact of inter-observer variability on the accuracy, precision and utility of a commonly-used grassland condition index. Ecol. Indic. 117, 106664. https://doi.org/10.1016/j.ecolind.2020.106664.
Gorrod, E.J., Keith, D.A., 2009. Observer variation in field assessments of vegetation condition: Implications for biodiversity conservation. Ecol. Manag. Restor. 10, 31–40. https://doi.org/10.1111/j.1442-8903.2009.00437.x.
Gosselin, F., Larrieu, L., 2020. Developing and using statistical tools to estimate observer effect for ordered class data: the case of the IBP (Index of Biodiversity Potential). Ecol. Indic. 110, 105884. https://doi.org/10.1016/j.ecolind.2019.105884.
Greco, S., Ishizaka, A., Tasiou, M., Torrisi, G., 2019. On the methodological framework of composite indices: a review of the issues of weighting, aggregation, and robustness. Soc. Indic. Res. 141, 61–94. https://doi.org/10.1007/s11205-017-1832-9.
Gwet, K.L., 2022. irrCAC Comput. ChanceCorrected Agreem. Coeff. (CAC). R. Package.
Hallgren, K.A., 2012. Computing inter-rater reliability for observational data: an overview and tutorial. Tutor. Quant. Methods Psychol. 8, 23–34.
Harper, M.J., McCarthy, M.A., van der Ree, R., Fox, J.C., 2004. Overcoming bias in ground-based surveys of hollow-bearing trees using double-sampling. For. Ecol. Manag. 190, 291–300. https://doi.org/10.1016/j.foreco.2003.10.022.

Hekkala, A.M., Jönsson, M., Kärvemo, S., Strengbom, S., Sjögren, J., 2023. Habitat heterogeneity is a good predictor of boreal forest biodiversity. Ecol. Indic. 148, 110069. https://doi.org/10.1016/j.ecolind.2023.110069.

Kärvemo, S., et al., 2021. Multi-taxon conservation in northern forest hot-spots: the role of forest characteristics and spatial scales. Landsc. Ecol. 36, 2703–2716. https://doi.org/10.1007/s10980-021-01205-x.

Kaufman, A.B., Rosenthal, R., 2009. Can you believe my eyes? The importance of interobserver reliability statistics in observations of animal behaviour. Anim. Behav. 78, 1487–1491. https://doi.org/10.1016/j.anbehav.2009.09.014.

Kelly, A., Franks, A., Eyre, T., 2011. Assessing the assessors: quantifying observer variation in vegetation and habitat assessment. Ecol. Manag. Restor. 12, 144–148. https://doi.org/10.1111/j.1442-8903.2011.00597.x.

Kenning, R.S., et al., 2005. Field efficiency and bias of snag inventory methods. Can. J. For. Res. 35, 2900–2910. https://doi.org/10.1139/x05-207.

Kitahara, F., Mizoue, N., Yoshida, S., 2009. Evaluation of data quality in Japanese National Forest Inventory. Environ. Monit. Assess. 159, 331–340. https://doi.org/10.1007/s10661-008-0632-8.

Koo, T.K., Li, M.Y., 2016. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J. Chiropr. Med. 15, 155–163. https://doi.org/10.1016/j.jcm.2016.02.012.

Kottner, J., et al., 2011. Guidelines for reporting reliability and agreement studies (GRRAS) were proposed. J. Clin. Epidemiol. 64, 96–106. https://doi.org/10.1016/j.jclinepi.2010.03.002.

Kyaschenko, J., et al., 2025. Complex relationship between soil fungi and conservation value assessments in boreal forests. Conserv. Biol., e70012 https://doi.org/10.1111/cobi.70012.

Landis, J.R., Koch, G.G., 1977. The measurement of observer agreement for categorical data. Biometrics 33, 159–174.

Larsson, J. 2001. Evaluation of the Forest Biodiversity Potential Method – An Indirect Biodiversity Assessment. Master thesis nr 70, Institutionen för naturvårdsbiologi, SLU, Uppsala.

Larsson Ekström, A., Jones, F.A.M., Hardenbol, Faith, Hekkala, A.A., Jönsson, A.M., Koivula, M., Strengbom, M., Sjögren, J., J, 2025. Habitat diversity as a taxon-dependent tool for predicting red-listed forest species. For. Ecol. Manag. 593, 122858. https://doi.org/10.1016/j.foreco.2025.122858.

Lin, L.I.-K., 1989. A concordance correlation coefficient to evaluate reproducibility. Biometrics 45, 255–268. https://doi.org/10.2307/2532051.

Lindenmayer, D.B., Likens, G.E., 2011. Direct measurement versus surrogate indicator species for evaluating environmental change and biodiversity loss. Ecosystems 14, 47–59. https://doi.org/10.1007/s10021-010-9394-6.

Lõhmus, A., 2009. Factors of species-specific detectability in conservation assessments of poorly studied taxa: the case of polypore fungi. Biol. Conserv 142 (11), 2792–2796. https://doi.org/10.1016/j.biocon.2009.05.022.

Lõhmus, A., Lõhmus, P., Runnel, K., 2018. A simple survey protocol for assessing terrestrial biodiversity in a broad range of ecosystems. PLoS ONE 13 (12), e0208535. https://doi.org/10.1371/journal.pone.0208535.

Milberg, P., et al., 2008. Observer bias and random variation in vegetation monitoring data. J. Veg. Sci. 19, 633–644. https://doi.org/10.3170/2008-8-18423.

Morrison, L.W., 2016. Observer error in vegetation surveys: a review. J. Plant Ecol. 9, 367–379. https://doi.org/10.1093/jpe/rtv077.

Nitare, J., Norén, M., 1992. Nyckelbiotoper kartläggs i nytt projekt vid Skogsstyrelsen. Sven. Bot. Tidskr. 86, 219–226.

OECD/European Union/EC-JRC. 2008. Handbook on Constructing Composite Indicators: Methodology and User Guide, OECD Publishing, Paris, https://doi.org/10.1787/9789264043466-en.

Oettel, J., Lapin, K., 2021. Linking forest management and biodiversity indicators to strengthen sustainable forest management in Europe. Ecol. Indic. 122, 107275. https://doi.org/10.1016/j.ecolind.2020.107275.

Paillet, Y., et al., 2015. Strong observer effect on tree microhabitats inventories: a case study in a French lowland forest. Ecol. Indic. 49, 14–23. https://doi.org/10.1016/j.ecolind.2014.08.023.

Perhans, K., et al., 2011. Fine-scale conservation planning outside of reserves: cost-effective selection of retention patches at final harvest. Ecol. Econ. 70, 771–777. https://doi.org/10.1016/j.ecolecon.2010.11.014.

R Development Core Team, 2022. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. ⟨https://www.R-project.org/⟩.

Ringvall, A., et al., 2005. Surveyor consistency in presence/absence sampling for monitoring vegetation in a boreal forest. For. Ecol. Manag. 212, 109–117. https://doi.org/10.1016/j.foreco.2005.03.002.

Ringvall, A., et al., 2007. Unrestricted guided transect sampling for surveying sparse species. Can. J. For. Res. 37, 2575–2586. https://doi.org/10.1139/X07-074.

Shoukri, M.M., et al., 2004. Sample size requirements for the design of reliability study: review and new results. Stat. Methods Med. Res. 13, 251–271. https://doi.org/10.1191/0962280204sm365ra.

Stevenson, M., Sergeant, E., Nunes, T., Heuer, C., Marshall, J., Sanchez, J., Thornton, R., Reichel, R., 2021. epiR: Tools for the Analysis of Epidemiological Data (& others). R. Package Version 2. 0. 50. ⟨https://CRAN.R-project.org/package=epiR⟩.

von Hirschheydt, G., Kéry, M., Ekman, S., Stofer, S., Dietrich, M., Keller, C., Scheidegger, C., 2024. Occupancy model reveals limited detectability of lichens in a standardised large-scale monitoring. J. Veg. Sci. 35, e13255. https://doi.org/10.1111/jvs.13255.

Wickham, H., 2016. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag, New York. ⟨https://ggplot2.tidyverse.org⟩. ISBN 978-3-319-24277-4.

Zeller, L., et al., 2022. Index of biodiversity potential (IBP) versus direct species monitoring in temperate forests. Ecol. Indic. 136, 108692. https://doi.org/10.1016/j.ecolind.2022.108692.