

Evaluating the usability and practicality of a proposed carbon footprint methodology and tool for National Olympic Committees

Yazeid Usama Aqqad  . Energy and Environment Institute. University of Hull. Hull, United Kingdom.

ABSTRACT

Robust carbon footprint methodologies and tools are increasingly required across the sport sector, yet National Olympic Committees lack a standardised, sector-specific approach. A bespoke carbon footprint methodology and accompanying calculation tool were developed for this purpose; however, technical rigour alone does not ensure that intended users can apply them. This study evaluates the usability and practicality of the proposed methodology and tool through direct engagement with end-users. Fourteen staff members from different National Olympic Committees applied the methodology and tool within their organisations over a four-to-six-week period, then completed the System Usability Scale and a semi-structured interview. The tool achieved a mean System Usability Scale score of 68.4, corresponding to a grade C and a marginal-high acceptability rating. Usability was rated positively, whereas learnability scored below the neutral midpoint, and scores correlated strongly with participants' years of experience ($r = .99$). Interviews indicated that effectiveness was the strongest dimension and learnability the weakest, with data collection identified as the main source of effort. Participants judged the methodology rigorous and credible but recommended improved onboarding, validation, and guidance to support less experienced users and wider adoption.

Keywords: sport management; carbon accounting; system usability scale; semi-structured interviews; sport sustainability; learnability; greenhouse gas reporting.

Cite this article as:

Aqqad, Y.U. (2026). Evaluating the usability and practicality of a proposed carbon footprint methodology and tool for National Olympic Committees. *Sustainability and Sports Science Journal*, 4(4), 289-338. <https://doi.org/10.55860/DNQB4576>

 **Corresponding author.** Energy and Environment Institute, University of Hull, Cottingham Road, Hull HU6 7RX, United Kingdom.

E-mail: Y.aqqad-2021@hull.ac.uk

Submitted for publication June 17, 2026.

Accepted for publication August 11, 2026.

Published October 01, 2026.

[Sustainability and Sports Science Journal](#). ISSN 2990-2975.

©Asociación Española de Análisis del Rendimiento Deportivo. Alicante. Spain.

Identifier: <https://doi.org/10.55860/DNQB4576>

INTRODUCTION

The development of robust Carbon Footprint (CF) methodologies and associated calculation tools has become an increasingly urgent priority across a wide range of sectors, including sport, as organisations face mounting pressure to measure, report, and reduce their greenhouse gas emissions (Adhikari et al., 2025; Onat et al., 2025). Within the sport sector specifically, reporting practices remain largely fragmented and voluntary, with National Olympic Committees (NOCs) lacking a standardised, sector-specific methodology capable of supporting consistent and credible CF assessments (Ciocoi-Pop et al., 2025; Marinelli et al., 2021). In response to this gap, a bespoke CF methodology and accompanying calculation tool have been developed for NOCs, grounded in established accounting frameworks and adapted to reflect their operational characteristics.

However, the development of a methodologically sound CF standard represents only one component of a successful implementation strategy. A methodology that cannot be applied consistently, correctly, and independently by its intended users risks undermining the very objectives it is designed to serve, irrespective of its technical rigour (Xia et al., 2025; Harangozo & Szigeti, 2017). The existing literature on carbon footprint methodologies and calculators has predominantly focused on methodological development, emissions estimation, and technical accuracy, with comparatively limited attention given to empirical usability testing involving independent end-users (Christy et al., 2023; Marlowe & Clarke, 2022; Bekaroo et al., 2019). Where evaluation has been undertaken, it has often relied on developer-led case studies or hypothetical applications that may not fully capture the operational challenges, learning requirements, and contextual constraints encountered by intended users in practice (Harangozo & Szigeti, 2017; Xia et al., 2025). Furthermore, many studies fail to distinguish between methodological robustness and tool usability, despite the fact that a technically sound methodology does not necessarily guarantee ease of use, accessibility, or practical implementation (Xia et al., 2025). As a result, there remains limited understanding of how CF methodologies and tools perform when applied by users with varying levels of expertise, organisational resources, and prior experience with carbon accounting.

This study addresses this gap by evaluating the usability and practicality of the proposed CF methodology and tool for NOCs through direct engagement with intended end-users. In doing so, the study moves beyond conceptual methodological development to examine how the proposed approach performs under real-world application conditions. The proposed CF methodology and tool are evaluated against the four core usability dimensions. These dimensions provide a structured framework for understanding not only whether participants were able to complete the carbon footprint assessment successfully, but also the extent to which the process was manageable, understandable, and appropriate for users with different levels of expertise. In addition, the study explores areas requiring refinement and improvement in order to strengthen the functionality, accessibility, and long-term adoptability of the proposed methodology and tool across NOCs.

To achieve this, participants from different NOCs were asked to apply the proposed CF methodology and tool within their organisations over a testing period before completing the SUS questionnaire and participating in semi-structured interviews. This dual-method design enables a more comprehensive and balanced evaluation than either approach could offer in isolation (Zapata et al., 2015) and is directly responsive to calls in the literature for greater end-user involvement in CF tool evaluation (Harangozo & Szigeti, 2017; Xia et al., 2025). The study therefore provides one of the few empirical evaluations within the carbon accounting literature that combines standardised usability assessment with qualitative inquiry involving independent end-users operating within authentic organisational environments.

The study begins by reviewing the importance of methodology and tool testing within carbon accounting research, followed by an overview of usability theory, usability assessment approaches, and the rationale for selecting the SUS and semi-structured interviews as the primary evaluation methods. The methodology adopted for participant recruitment, data collection, and data analysis is then presented. This is followed by the results and discussion sections, which examine the quantitative SUS findings alongside the qualitative interview themes relating to effectiveness, efficiency, satisfaction, learnability, and proposed refinements. Finally, the study concludes by reflecting on the implications of the findings for the further development, implementation, and potential standardisation of the proposed CF methodology and tool for NOCs.

Literature review

The importance of testing

The development of CF methodologies and associated calculation tools has expanded rapidly across sectors, driven by growing demands for emissions disclosure and accountability (Adhikari et al., 2025; Onat et al., 2025). However, the credibility and usefulness of such methodologies do not rest solely on their technical alignment with accounting standards or their theoretical completeness (Asam et al., 2025; Augoye et al., 2024). Increasingly, the effectiveness of CF approaches depends on whether they can be applied consistently and correctly by intended users within real organisational contexts (Bowen & Wittneben, 2011).

Methodology and tool testing play a critical role in identifying gaps between conceptual design and practical application. Without empirical testing, carbon accounting tools risk being overly complex, misaligned with user capabilities, or incompatible with existing organisational processes (Melville & Whisnant, 2014). This is particularly relevant for organisations with limited sustainability expertise or decentralised data structures, where methodological burden can directly undermine data quality, consistency, and long-term uptake (Okon & Marauche, 2025).

Existing carbon accounting literature has largely prioritised methodological development and emissions estimation accuracy, while paying comparatively little attention to systematic usability and applicability testing (Christy et al., 2023; Marlowe & Clarke, 2022; Bekaroo et al., 2019). Where evaluation does occur, it is often informal or retrospective, rather than treated as a distinct analytical exercise (Mulrow et al., 2018). As a result, limited insight is available regarding how users interact with CF tools, interpret methodological guidance, or navigate data requirements in practice (Mulrow et al., 2018).

Empirical testing of methodologies and tools provides an evidence base for refining methodological assumptions, simplifying calculation procedures, and improving guidance clarity (Breton & Aggeri, 2018; Mulrow et al., 2018). Usability testing, in particular, enables the identification of cognitive and operational barriers that may not be apparent during methodological design (Kong et al., 2021). Similarly, qualitative engagement with users allows organisational constraints such as data availability, internal coordination, and reporting responsibilities to be explicitly considered within methodological refinement processes (Ball et al., 2023; Chatty et al., 2021).

For standardisation efforts, methodology and tool testing is especially important. A proposed standard that is not empirically appraised risks inconsistent application across organisations and limiting its potential for sector-wide adoption (Strohmeier et al., 2025). Systematic testing therefore supports not only methodological robustness but also the legitimacy and scalability of carbon footprint standards (Marinelli et al., 2021), particularly in sectors such as sport where reporting practices remain fragmented and largely voluntary (Ciocoi-Pop et al., 2025).

Testing methods

CF methodologies and tools have been tested and evaluated through many different approaches. However, the majority of studies that have developed CF methodologies do not offer a supporting tool to apply the methodology, and in the rare cases where a tool is developed, the study does not differentiate between testing the methodology and assessing the usability of the tool itself (for example, see Mateo Pla et al., 2021; Loyarte-López et al., 2020). This distinction is critical, as methodological soundness does not guarantee practical usability, while tool usability does not guarantee methodological rigour (Xia et al., 2025).

Studies developing CF methodologies can be broadly grouped into two categories. The first comprises studies that propose a methodology conceptually, without subjecting it to empirical testing or practical application (see Abdul Aleem et al., 2021; Szatylowicz et al., 2021) or only offering detailed numerical examples based on hypothetical values (e.g., Gandola and Asdrubali, 2024). The second group includes studies that implement and test their proposed methodologies, often through pilot projects or case studies (e.g., Griffiths et al., 2024; Issa Zadeh et al., 2023). Testing methodologies through case studies offers a robust means of validating their practical applicability, technical feasibility, and calculation accuracy. However, such testing is often undertaken by the methodology developers themselves, which may introduce bias or fail to capture the operational challenges and contextual constraints that independent end-users encounter in practice (Harangozo & Sziget, 2017; Spandonidis, 2025). To mitigate this limitation, methodology evaluation should actively involve intended end-users through both qualitative and quantitative approaches, enabling systematic feedback on usability, clarity, and overall effectiveness.

The situation for CF tools is often the reverse, with developers placing greater emphasis on user interface and user experience than on the robustness and transparency of the underlying methodology. As a result, the current state of CF calculators is characterised by uncertain accountability and considerable methodological inconsistency (Harangozo et al., 2017; Malakahmad et al., 2015). Research shows that many calculators lack consistency and transparency (Birnik, 2013), leading to substantial variation in results even when identical inputs are used (Xia et al., 2025).

CF calculators are defined as software tools that employ quantitative models to estimate carbon emissions (Xia et al., 2025). These tools exhibit varying degrees of complexity and user interaction modalities, ranging from simple spreadsheet inputs to multi-step web-based interfaces (Barendregt et al., 2020). Studies have adopted a range of approaches to evaluate their functionality, including real-world case studies (e.g., Beery et al., 2011), hypothetical scenarios (e.g., Xia et al., 2025), usability and user experience testing (e.g., Bekaroo et al., 2019), expert validation (e.g., Spandonidis, 2025), and comparative analyses against established calculators (e.g., Tuomisto et al., 2015). In several studies, these testing approaches were complemented by additional qualitative methods, such as workshops, facility visits, and interviews to assess the CF calculators (e.g., Alromaizan et al., 2023; Azarkamand et al., 2021; Büchs et al., 2018). Employing multiple evaluation methods is recommended when assessing the usability of software tools, such as CF calculators, as the strengths of one approach can compensate for the limitations of another, providing a more comprehensive and balanced evaluation (Zapata et al., 2015).

Usability

In its simplest sense, usability refers to the capability of being used, or more plainly, the ease with which a system can be used (Bevan et al., 2015). However, there is no single universally accepted definition in the literature, as usability is a multifaceted concept encompassing numerous dimensions, making it difficult to capture comprehensively within one definition (Lewis, 2014). Usability is regarded as a general quality of appropriateness; however, it cannot be evaluated in isolation and must be assessed within a defined context

that specifies the intended users of the system, the tasks they will perform, and the characteristics of the environment in which it will be used (Brooke, 1996). The International Organization for Standardization (ISO) defines usability as *“the extent to which a system, product or service can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use”* (ISO, 2018).

Based on the ISO definition, usability is measured through three components. Effectiveness refers to the extent to which users can successfully complete tasks accurately, completely, and appropriately, reflecting the quality and correctness of the system's outputs (Bevan et al., 2015; Kortum & Peres, 2014). Efficiency relates to the number of resources required to complete a task, including time, human effort, costs, and materials (ISO, 2018). Together, effectiveness and efficiency assess the overall performance of the system (Bevan et al., 2015). Satisfaction concerns users' subjective experience, reflecting their comfort and acceptability when interacting with the system (Corsini et al., 2012). For a system to be considered highly usable, it must perform well across all three dimensions. For example, a user may successfully complete a task (effectiveness), but if doing so requires excessive time, cost, or effort, or results in dissatisfaction with the process, the system would still be regarded as having low usability, as it fails to meet the efficiency and satisfaction measures (Kortum & Peres, 2014).

Effectiveness, efficiency, and satisfaction are not the only measures used to evaluate usability, as additional criteria have also been applied in the literature (Goundar et al., 2024; Bevan et al., 2015). For example, ISO/IEC 25010 (2024) decomposes usability into several measurable attributes, including appropriateness, recognisability, learnability, operability, user error protection, user interface aesthetics, and accessibility. Similarly, Nielsen (1994) conceptualises usability through five key attributes: learnability, efficiency, memorability, error reduction, and satisfaction. However, this broader conceptualisation of usability is typically associated with the testing of software products and tools rather than systems as a whole (Sagar & Saha, 2017; Almazroi, 2021). Nevertheless, it is common practice to extend the core measures of effectiveness, efficiency, and satisfaction by including learnability, which refers to the time required or ease with which typical users can learn a system and achieve a certain level of competency, as a fourth usability dimension (Lee & Sah, 2020; Abran et al., 2003).

Efficiency and effectiveness are regarded as the objective components of usability and are typically measured through task completion times and task success rates, respectively (Kortum & Peres, 2014; Lewis, 2018). In contrast, satisfaction and learnability are considered subjective components, as they depend on users' perceptions and experiences and are therefore more complex to measure (Aldoihi, 2020; Golondrino et al., 2023). These dimensions are highly sensitive to the context of use, including frequency of use, user expertise, task type, and the operating environment (Lewis et al., 2015; Bevan et al., 2015).

Usability can be assessed through various methods, such as qualitative surveys, interviews, and observational techniques, with questionnaires being the most established and frequently used method (Blattgerste et al., 2022). Usability questionnaires are of three types: standardised usability questionnaires, heuristic evaluations, and custom questionnaires (Sagar & Saha, 2017). However, standardised questionnaires are more reliable and robust than the other two approaches (Lewis, 2014; Wang et al., 2023). A standardised questionnaire refers to an assessment instrument developed for repeated use, featuring a predetermined set of questions arranged in a fixed order and format, accompanied by explicit guidelines for deriving quantitative metrics from respondents' answers (Lewis et al., 2015). Developers of such instruments routinely evaluate their psychometric properties through assessments of validity (the extent to which a questionnaire measures what it claims to measure), reliability (its consistency of measurement), and

sensitivity (the extent to which independent variables affect measurement) (Greene et al., 2016). In addition, standardised usability questionnaires are designed to avoid different types of bias (Lewis, 2014). However, standardised usability questionnaires have been criticised for their limited adaptability, as deviations from their original design, such as changes in wording, scale length, or intended context of use, can compromise their validity and reliability, potentially leading to invalid results (Graziotin et al., 2017; Felipe et al., 2023). Nevertheless, research has shown that robust standardised usability questionnaires are able to tolerate some deviation from their set guidelines (Lewis, 2014).

Hodrien and Fernando (2021) identified 14 standardised usability questionnaires that vary in their features, including the number of statements, scale, level of detail, and length. For example, the Computer System Usability Questionnaire (Lewis, 1995) is a lengthy, detailed questionnaire that consists of 16 statements evaluated on a 7-point Likert scale and is suitable for providing diagnostic insights into where exactly a system is failing. The System Usability Scale (SUS) (Brooke, 1996) is a moderately detailed questionnaire that contains 10 statements scored on a 5-point Likert scale and provides a reliable overview of a system's usability. The Usability Metric for User Experience (UMUX) (Finstad, 2010) and UMUX-Lite (Lewis, 2013) are short questionnaires of 4 and 2 statements, respectively, assessed on a 7-point Likert scale and ideal for agile environments, lightweight tracking, and quick pulse checks.

Despite the availability of numerous standardised usability questionnaires, the System Usability Scale (SUS) is one of the most widely adopted instruments due to its ability to assess a broad range of products, services, and systems (Gao et al., 2020). By 2009, the SUS was reported to have been used in 43% of unpublished industrial usability studies (Sauro & Lewis, 2009), while by 2017, the paper that introduced the SUS questionnaire had been cited 5,015 times, three times more than its closest rival (Lewis, 2018). Its popularity stems from its brevity, ease of completion, and balanced mix of positively and negatively worded items, which helps reduce response bias and improve objectivity (Rotaru et al., 2024), making it an effective measure for practitioners and researchers (Gao et al., 2020). In addition, the reliability and adaptability of the SUS have been tested, with results demonstrating excellent reliability, with coefficient alpha typically exceeding .90 (Lewis, 2018). It has also been shown to be insensitive to minor wording changes (Bangor et al., 2008), to the use of all-positive or mixed-tone statements (Sauro & Lewis, 2011), and to the omission of a single statement that does not fit the measurement context (Lewis & Sauro, 2017). This indicates that the SUS is a highly robust and versatile tool for usability assessment that can be applied with confidence to both large and small sample sizes (Lewis et al., 2015; Lewis, 2018).

Research has shown that there is a strong correlation between the SUS questionnaire and effectiveness at the system level, measured through task success rates (Kortum & Peres, 2014). However, using a usability questionnaire does not negate the need for comprehensive usability testing to ensure robust validation (Johnson et al., 2022), as the SUS questionnaire only provides quantitative data on the usability of a system and does not provide insights into which aspects of the system need improvement or how best to improve them (Klug, 2017). Therefore, it is important to combine a usability questionnaire such as the SUS, and the three ISO metrics of usability, with additional qualitative methods, such as interviews, or to incorporate open-ended questions to gain a better understanding of the performance and overall usability of the tested system (Fathiyyah et al., 2022; Johnson et al., 2022).

Research questions

The aim of this study is to evaluate the proposed CF methodology and tool for NOCs by assessing their usability and practicality. This will be achieved by answering the following research questions (RQs):

- RQ1: To what extent does the proposed CF standard methodology and tool meet accepted usability standards as measured by the System Usability Scale?
- RQ2: How does the proposed CF methodology and tool perform against the four usability measures of effectiveness, efficiency, satisfaction and learnability?
- RQ3: What refinements and improvements are required to enhance the usability, functionality, and practical adoption of the proposed methodology and tool?

METHODOLOGY

Most existing studies evaluating CF methodologies and tools have predominantly relied on either conceptual proposal without empirical testing, or case studies conducted by the methodology developers themselves. Case studies have been widely employed as a validation approach, offering insights into technical feasibility and calculation accuracy (Griffiths et al., 2024; Issa Zadeh et al., 2023). However, when testing is undertaken solely by developers, it may introduce bias and fail to capture the operational constraints and contextual challenges encountered by independent end-users in practice (Harangozo & Szigeti, 2017; Spandonidis, 2025). Furthermore, studies that do develop supporting tools rarely differentiate between testing the methodological soundness of the underlying approach and evaluating the usability of the tool itself (see Mateo Pla et al., 2021; Loyarte-López et al., 2020). This distinction is critical, as methodological rigour does not guarantee practical usability, nor does tool usability ensure methodological robustness (Xia et al., 2025).

The over-reliance on developer-led case studies and the conflation of methodology and tool testing have constrained a comprehensive understanding of how CF methodologies perform when applied by real users in authentic contexts. Harangozo & Szigeti (2017) and Xia et al. (2025) have therefore advocated diversifying evaluation approaches beyond developer-conducted testing, recommending the active involvement of intended end-users through both qualitative and quantitative methods. Such engagement enables systematic feedback on usability, clarity, and overall effectiveness. To address these gaps in the literature, this study employs a standardised usability questionnaire (the SUS), complemented by a qualitative method (semi-structured interviews), to ensure that the proposed CF methodology and tool are both technically robust and practically usable for their intended users.

The system usability scale

The SUS is a simple, ten-item standardised questionnaire widely employed to obtain a global, subjective assessment of perceived usability (Brooke, 1996). It is typically administered after respondents have interacted with the system under evaluation, but before any debriefing or discussion, with respondents instructed to record their immediate responses to each item without prolonged deliberation (Brooke, 1996). It is a 5-point Likert-scale instrument in which the statements have been carefully selected to ensure that respondents read each item and actively consider their level of agreement (Brooke, 1996). The SUS questionnaire employs alternating positively and negatively worded statements as a strategy to mitigate response biases, such as extreme response bias and acquiescence bias, the tendency for respondents to agree with statements regardless of content (Lewis, 2014). By alternating tone, the SUS requires respondents to make a deliberate cognitive effort for each item and may also help researchers identify inattentive respondents (Kortum et al., 2021). However, evidence suggests that including mixed-tone items can introduce greater problems than those it seeks to resolve (Chyung et al., 2018). This approach has been associated with reduced internal reliability, distorted factor structures, and increased complexity in cross-cultural interpretation (Lewis, 2018). Furthermore, respondents may find it difficult to adjust their response patterns when alternating between positively and negatively worded statements, increasing the likelihood of inadvertent errors, while researchers may overlook the necessary step of reverse-scoring negatively phrased

items, resulting in response miscoding. Both of these mistakes can lead to more detrimental results than the response biases such wording is intended to mitigate (Sauro & Lewis, 2011; Lewis, 2018).

The SUS has demonstrated flexibility to minor wording changes and statement tone changes. The original SUS statements have the word “system” and “cumbersome” in statement 8. Research has shown that replacing the words “system” with the product, service or the name of the product or service and replacing the word “cumbersome” with “awkward” to clarify the meaning for non-English speakers, had identical results to the original SUS version (Lewis & Sauro, 2009). Similarly, in the original SUS version, the five odd statements are worded with a positive tone and the five even statements are worded with a negative tone. Studies examining the effects of substituting the negatively phrased items on the SUS with positively phrased ones found that the different versions did not lead to variations in response bias or the strength of responses. However, it did reveal that the original version resulted in more errors from both participants and researchers, seemingly because of the shift in tone between items (Sauro & Lewis, 2011; Kortum et al., 2021). Furthermore, since the all-positive SUS relies on the same core psychometric properties as the original tool, the final numerical score behaves identically. Sauro and Lewis (2011) demonstrated that all-positive SUS scores yield mean values and distributional properties identical to those obtained with the original SUS, confirming the psychometric equivalence of the two forms. Therefore, once the final P-SUS score (0 to 100) is calculated, the canonical industry benchmarks can be directly applied.

Given the international nature of NOCs and the fact that many respondents are not native English speakers, as well as the need to avoid potential errors by both participants and researchers when using mixed-tone statements, an all-positive version with simple English was used to evaluate the usability of the CF methodology and tool (Table 1).

Table 1. The SUS statements adapted from Sauro and Lewis (2011).

The SUS statements
I think that I would like to use the CF methodology and tool frequently.
I found the CF methodology and tool to be simple.
I thought the CF methodology and tool were easy to use.
I think that I could use the CF methodology and tool without the support of a technical person.
I found the various functions in the CF methodology and tool were well integrated.
I thought there was a lot of consistency in the CF methodology and tool.
I would imagine that most people would learn to use the CF methodology and tool very quickly.
I found the CF methodology and tool very intuitive.
I felt very confident using the CF methodology and tool.
I could use the CF methodology and tool without having to learn anything new.

Each statement is designed to measure a perceived attribute of the system: statement 2 assesses perceived complexity, statement 3 perceived ease of use, statement 6 perceived consistency, statements 7 and 10 perceived learnability, and statement 9 confidence in use (Lewis & Sauro, 2018). However, the creator of the SUS has emphasised that it is a unidimensional measure of usability, as scores for individual statements are not meaningful in isolation (Brooke, 1996). Nevertheless, research has shown that the SUS can be divided into two subscales: usability and learnability, with statements 4 and 10 aligning with learnability, while the remaining eight statements relate to usability (Sharfina & Santoso, 2017). Moreover, the SUS has been used to evaluate five subfactors: effectiveness (statements 5 & 6), efficiency (statements 7 & 10), engagement (statements 1 & 8), error tolerance (statement 4), ease of use (statements 2 & 3), and self-efficacy (statement 9) (Wagner, 2023). Despite the ability of the SUS to be subdivided into different scales, the usability and

learnability subscales remain the most commonly applied by researchers and practitioners (Sagar & Saha, 2017; Maqbool & Herold, 2024).

The SUS has demonstrated its ability to produce reliable results with small sample sizes (Brooke, 2013). Nielsen (2012) argues that administering the SUS with as few as five participants is sufficient to identify as many usability issues as would be detected with a larger sample. Other studies have shown that the SUS yields reliable results with sample sizes ranging from 6 to 14 participants (Binyamin et al., 2016; Lewis, 2018), with no significant differences observed between the lower and upper ends of this range (Orfanou et al., 2015).

SUS scores range from 0 to 100, with each statement rated on a scale from 1 (strongly disagree) to 5 (strongly agree). The total score is calculated by subtracting 1 from the score of odd-numbered (positively worded) statements, subtracting the score of even-numbered (negatively worded) statements from 5, summing the adjusted values, and multiplying the total by 2.5 (Brooke, 1996). However, when all statements are positively worded and scored on the same 1–5 scale, it is only necessary to subtract 1 from each statement score, sum the values, and multiply by 2.5 (Sauro & Lewis, 2011). Higher SUS scores indicate better usability (Liang et al., 2018). However, the original version of the SUS does not provide guidelines for interpreting the scores. Bangor et al. (2008) developed an absolute grading scale with A: 90–100, B: 80–89, C: 70–79, D: 60–69, and F: <60 to help interpret SUS scores, while Sauro and Lewis (2016), using the same dataset, carried this work forward by developing a curved rather than an absolute grading scale. This curved grading scale produced an average SUS score of 68, which was intentionally positioned as the midpoint of the scale (grade C), and which would have corresponded to grade D under the absolute grading system (Mol et al., 2020). The scale includes 11 distinct grade bands (Table 2), offering greater granularity than the absolute grading scale, and has become widely adopted in industry practice for assigning letter grades to average SUS scores (Lewis, 2018).

Table 2. The curved grading scale for assigning the SUS scores (Sauro & Lewis, 2016).

SUS Score range	100-84.1	84.0-80.8	80.7-78.9	78.8-77.2	77.1-74.1	74.0-72.6	72.5-71.1	71-65.0	64.9-62.7	62.6-51.7	51.6-0.0
Grade	A+	A	A-	B+	B	B-	C+	C	C-	D	F

Based on SUS scores, system usability can be assigned grades ranging from F to A, with the following interpretations: F (inadequate), D (below average), C (OK), B (good), and A (excellent).

Semi-structured interviews

Within qualitative research, interviews are the most commonly used data collection method (Thelwall & Nevill, 2021), with semi-structured interviews (SSIs) being recognised as the most effective interview technique for generating in-depth, contextually grounded data (Eppich et al., 2019; Bearman, 2019). Longhurst & Johnston (2023), define SSIs as a verbal interchange (conversation) between two people, the interviewer and the interviewee, where the former seeks to elicit information from the latter through predetermined, yet flexible, questions.

SSIs can be positioned along a continuum between structured interviews, which are fully controlled with a rigid set of questions asked in a fixed order and limited interviewer flexibility, and unstructured interviews, which lack predetermined questions and are guided more freely by the interviewer (Clifford et al., 2016). SSIs offer the advantages of both interview types: the flexibility of unstructured interviews, where interviewers can

adjust questions, rephrase, introduce new topics, or pursue unexpected but valuable lines of inquiry based on the interviewee's responses, while at the same time following an interview protocol or guide to ensure that the conversation remains on track and covers all relevant topics (Mashuri et al., 2022; Naz et al., 2022).

Kallio et al. (2016) argue that the development of an SSI guide involves five stages: (1) identifying the prerequisites for using SSIs; (2) retrieving and utilising previous knowledge; (3) formulating the preliminary SSI guide; (4) pilot testing the interview guide; and (5) presenting the complete SSI guide. Similarly, Castillo-Montoya (2016) proposes a four-stage Interview Protocol Refinement Framework to systematically develop and refine interview protocols, which includes: (1) ensuring interview questions align with research questions; (2) constructing an inquiry-based conversation; (3) receiving feedback on the interview protocol; and (4) piloting the interview protocol. While both frameworks overlap in several stages, they also differ in others, offering complementary strengths that can enhance the overall development process.

The purpose of the first stage of identifying the prerequisites is to assess the suitability of SSIs as a rigorous data collection method for addressing the research questions (Kallio et al., 2016). SSIs are particularly well suited to exploring participants' independent thoughts, perceptions, and opinions through open-ended questions and probing techniques (Cridland et al., 2015; Newcomer et al., 2015). They are also appropriate in contexts where participants are direct beneficiaries of the service (Newcomer et al., 2015), and for research questions focused on how a service or tool is currently performing or how it can be improved (Adeoye-Olatunde & Olenik, 2021).

The aim of the second stage of retrieving and using previous knowledge is to develop a thorough understanding of the subject through the critical appraisal of existing knowledge. This is achieved through conducting a literature review, which forms the basis for the interview framework (Kallio et al., 2016). This allows the identification of relevant and broad areas to be explored during the interview, which are then refined into smaller, more manageable categories by grouping similar themes before proceeding to the formatting and refinement of the questions (Naz et al., 2022).

Formulating the preliminary SSI guide is the third stage. An interview guide is defined as a set of questions used to steer the conversation towards the research topic during the interview (Smulowitz, 2017). However, the structure of an SSI guide is designed to be free and flexible in order to facilitate dialogue, change the order of the questions, and move easily between questions (Mashuri et al., 2022). The quality of the interview guide affects both how the interview is conducted and how the collected data are analysed (Cridland et al., 2015). Therefore, the questions it contains should be participant-focused, non-leading, clearly phrased, single-barrelled, and open-ended (Manouchehri et al., 2022). This encourages participants to give answers that are spontaneous, detailed, distinctive, and vivid, helping to generate the richest possible dataset (Kallio et al., 2016). Nevertheless, ensuring that the interview questions align with the research questions is a key consideration, which can be verified by developing a matrix that maps interview questions to the corresponding research questions (Castillo-Montoya, 2016).

A well-designed SSI guide should be prepared in advance and organise questions around the main themes the interviewer aims to cover, with an appropriate balance of follow-up questions and probes (Mashuri et al., 2022). In addition, interview questions should be phrased differently from the research questions and structured in line with the social norms of everyday conversation (Castillo-Montoya, 2016). All participants should be asked about these core themes in a logical and progressive order, with follow-up questions used to clarify key points, maintain the flow of conversation, keep the discussion focused, and elicit accurate and comprehensive responses (Kallio et al., 2016). Verbal probing techniques such as echoing (repeating the

participant's point), verbal agreement (e.g. 'yes', 'okay', or 'uh-huh'), and baiting (suggesting prior awareness of certain information), alongside non-verbal silence, can be used to encourage participants to elaborate further and provide deeper insights (Whiting, 2008).

The fourth stage involves pilot testing the SSI guide. This is undertaken to assess the coverage and relevance of the interview questions and to allow for any necessary revisions, adjustments, or rewording prior to implementation, thereby ensuring the quality of the collected data (Kallio et al., 2016). By simulating an actual interview, this stage provides an opportunity to test the interview guide in terms of question order, interview duration, and whether participants are able to respond effectively, as well as to refine elements such as rapport building, the interview process, consent procedures, setting, recording, and timekeeping (Castillo-Montoya, 2016). Pilot testing can be performed in three different ways: internal testing, expert assessment, and field testing (Naz et al., 2022). After completing the previous stages, the final stage is presenting the completed SSI guide.

Although the limitations of SSIs include being time- and resource-intensive, the potential for interviewer bias, and the requirement for specific interviewer skills (Newcomer et al., 2015), these challenges can be mitigated through several approaches. Software that enables recording with automatic captions can reduce the time and resources required for transcribing interviews (Siedlecki, 2022), while interviewer bias can be minimised through reflexive practices, such as taking notes after each interview to reflect on how the interviewer's experiences, values, and position may have influenced the process (Cairns-Lee et al., 2022). In addition, interviewing skills such as active listening and rapport building can be developed through the use of an interview guide, as well as mock interviews and role play (Shea & Barney, 2015; Goodell et al., 2016).

Data collection

Recruitment

As the methodology and tool are aimed at NOCs, the main inclusion criterion was being a staff member of an IOC-recognised NOC, regardless of whether the organisation had previously engaged with CF assessment. All participants recruited for the study were active staff members from different NOCs, with some or no experience of CF assessment, and held sustainability or senior management roles within their organisations. The final number of participants was 14, which is more than adequate to yield reliable results for the SUS, as the recommended sample size ranges from 6 to 14 participants (Binyamin et al., 2016; Lewis, 2018), while for SSIs the recommended sample size to reach saturation, where no further new information is obtained, is between 10 and 15 (Mustaffa et al., 2022).

An initial email was sent to participants containing the purpose of the study and what was expected of them. After participants agreed to take part in the study, a second email was sent containing a breakdown of the study phases. In phase one, participants tested the CF methodology and tool in their organisations by completing a full or partial CF assessment over a period of 4–6 weeks. Phase two involved completing the SUS questionnaire and taking part in a 30–60 minute interview. The email also contained the following attachments and documents: the methodology manual, the Excel tool, the participant information sheet (PIS), and the consent form.

The PIS informed participants about the purpose of the study, the voluntary and anonymous nature of participation, the right to withdraw, and data protection. The consent form reiterated the voluntary nature of participation, withdrawal rights, and anonymity, and sought permission to record the interviews. Written consent was obtained through an online consent form, and verbal consent was sought before recording the interviews and reconfirmed after recording had begun.

The System Usability Scale

After participants completed the testing phase, they were asked to complete the SUS questionnaire before the interview took place, with respondents instructed to record their immediate responses to each item without prolonged deliberation (Brooke, 1996). Because participants were located in different time zones, and to facilitate dissemination, the questionnaire was hosted on an online platform. The first page of the questionnaire included the consent form, stating the anonymity of the study and the participants' right to withdraw before the data were anonymised, as well as instructions on how to complete the questionnaire. Participants could not move to the next page without signing the consent form and acknowledging their understanding of the instructions. The following page contained the 10 SUS statements on a 5-point Likert scale. The final page asked participants for the name of their organisation (for administrative use only), their job role, years of experience, and any previous experience with CF assessment.

Semi-structured interviews

The SSIs were conducted online, as face-to-face interviews were not possible due to participants residing in different geographical locations. The interviews were carried out using the Microsoft Teams communication platform, selected for its features, including video or audio recording, no time limits, live transcription, and data encryption (Azhari et al., 2022). The platform's recording and live transcription features were used to enable the interviewer to remain fully engaged during the interview and to support a natural flow of conversation, while also reducing the time and resources associated with the transcription process (Adeoye-Olatunde & Olenik, 2021). Immediately after each interview, the recording was reviewed and the transcription was checked for accuracy and consistency against the recording (Lu et al., 2024). During this process, all transcripts were anonymised by removing any identifying information related to participants (Adeoye-Olatunde & Olenik, 2021). In total, 14 interviews were conducted, ranging from 40 to 75 minutes, with an average duration of 55 minutes, in accordance with the interview guide.

The interview guide was developed based on the frameworks proposed by Kallio et al. (2016) and Castillo-Montoya (2016). As the aim of this study is to assess and improve the usability and practicality of the proposed CF methodology and tool from the perspectives of participants, who are the direct beneficiaries of the proposed CF methodology and tool, this aligns with the prerequisites for using SSIs as a data collection method (stage 1) to answer the research questions. A thorough understanding of the subject was developed through the literature review and insights gained from previous studies (stage 2). This helped in formulating the preliminary SSI guide (stage 3). In this stage, two important considerations were observed: ensuring that the interview questions aligned with the research questions (see Appendix A) and constructing an inquiry-based conversation that reflected the social norms of everyday conversation. After the preliminary SSI guide was developed, it was shared with the supervisory team for feedback on structure, length, writing style, and comprehension (Castillo-Montoya, 2016). After the review was completed and feedback was implemented, the SSI guide was pilot tested (stage 4) with three sport sustainability practitioners with characteristics similar to those of the participants in the actual study (Maxwell, 2013). Following the pilot interviews, no major changes were made. With the conclusion of this stage, the complete SSI guide was developed (stage 5) (see Appendix B).

Data analysis

The SUS questionnaire data were calculated based on positively worded statements by subtracting 1 from each statement score, summing the values, and multiplying by 2.5 (Sauro & Lewis, 2011), and the scores were interpreted using the Curved Grading Scale of Sauro and Lewis (2016). In addition, statistical analysis was conducted using a one-sample t-test to assess whether the mean SUS scores differed significantly from the midpoint value of three (on the 5-point Likert scale) for each individual SUS item, as well as for the

usability and learnability constructs. As this involved ten separate item-level comparisons in addition to the two subscale-level comparisons, a Holm-Bonferroni step-down correction was applied to the item-level tests to control for the increased risk of Type I error associated with repeated significance testing, with corrected and uncorrected p-values reported alongside one another for transparency (Holm, 1979). The internal consistency and reliability of both constructs were evaluated using Cronbach's alpha (Hamm et al., 2017).

The SSI data were analysed deductively, focusing on specific topics of interest guided by the research questions, using the framework for thematic analysis proposed by Braun and Clarke (2006). The process began with familiarisation with the data to develop an overall understanding. This was followed by coding according to the previously defined main themes, which was used to identify relevant extracts from the transcripts. Similar extracts were grouped to form sub-themes under the main themes. These themes were then reviewed to ensure that they accurately reflected both the coded extracts and the entire dataset. Finally, direct quotes from participants were used to support each theme. Supplementary Table (Appendix C) presents representative raw quotes alongside their assigned codes, sub-themes, and themes, illustrating this coding process in full.

RESULTS AND DISCUSSION

The System Usability Scale

Table 3 shows how each participant scored each SUS statement, while Table 4 presents the SUS score for each participant and the average overall SUS score across all respondents. The CF methodology and tool achieved a score of 68.4 on the unidimensional overall usability scale, placing it just above the SUS average score (Sauro & Lewis, 2016; Sauro, 2011). This score can be interpreted in four different ways (Figure 1) (Sauro, 2018). The CF methodology and tool scored “*marginal-high*” for acceptability range, “*good*” for descriptive adjectives (Bangor et al., 2009), and “*grade C*” on the curved grading scale (Sauro & Lewis, 2016), as well as “*passive*” on the Net Promoter Score (NPS), which measures loyalty and satisfaction by asking users whether they would recommend the system to other users (Sauro, 2018).

Table 3. Participants' individual SUS scores and years of experience.

Participant	SUS score	Years of experience
P1	65	5
P2	57.5	3
P3	80	12
P4	65	6
P5	82.5	14
P6	75	9
P7	45	1
P8	65	4
P9	85	15
P10	60	4
P11	77.5	10
P12	55	2
P13	75	8
P14	70	7
SUS Score	68.4	

Positioning the CF methodology and tool score of 68.4 within the existing literature is challenging, as there are no directly comparable applications. The closest comparisons include a usability evaluation of an

integrated Life Cycle Analysis tool (Qiang & Pa, 2024), which reported an SUS score of 50, and an assessment of MS Excel (Kortum & Bangor, 2013), which scored 54.4, both below the acceptable average. A conceptually similar tool, a mobile application for CF tracking, achieved a higher SUS score of 82.5; however, mobile applications typically attain higher usability scores than desktop-based tools (Vlachogianni & Tselios, 2022; Wetzlinger et al., 2014). More broadly, sustainability-related tools and applications report a wide range of SUS scores, ranging from 59.5 to 93.25 (e.g., Yonata & Purwarianti, 2017; McGlinn et al., 2017; Aresta et al., 2024; Oshidero & Coley, 2026). Lewis and Sauro (2021) further argue that, while achieving an SUS score of around 80 is feasible for many systems, it is less realistic for spreadsheet-based applications, such as the proposed CF methodology and tool.

Overall, the results indicate that the CF methodology and tool were perceived positively by participants, although there remains scope for improvement. To further examine usability issues with the CF methodology and tool, the individual SUS items were analysed against the midpoint value.

Table 4. Statistical analysis of SUS statements.

Item	Mean	t-value	df	p-value	Significant	Holm-adj. p	Significant
SUS 1	4.29	5.77	13	<.001	Yes	<.001	Yes
SUS 2	3.86	3.72	13	.002	Yes	.010	Yes
SUS 3	3.57	2.82	13	.014	Yes	.029	Yes
SUS 4	2.50	-3.58	13	.003	Yes	.010	Yes
SUS 5	4.50	10.80	13	<.001	Yes	<.001	Yes
SUS 6	4.50	10.80	13	<.001	Yes	<.001	Yes
SUS 7	3.14	0.62	13	.549	No	.546	No
SUS 8	4.07	4.81	13	<.001	Yes	.002	Yes
SUS 9	4.43	6.26	13	<.001	Yes	<.001	Yes
SUS 10	2.36	-3.83	13	.002	Yes	.010	Yes

Note. Holm-adj. p = Holm-Bonferroni adjusted p-value, calculated across the ten item-level comparisons to control for Type I error inflation (Holm, 1979).

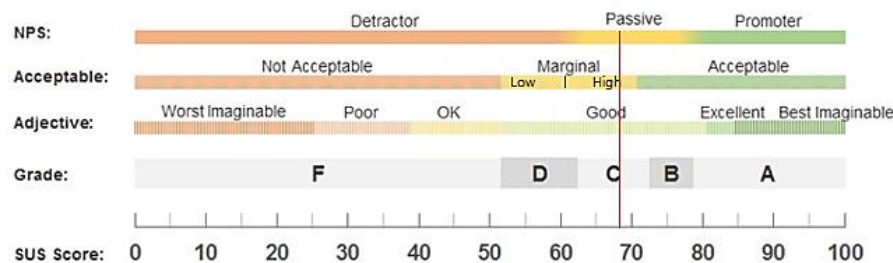


Figure 1. Scale for determining NPS, acceptability range, adjective rating, and letter grade (adapted from Sauro (2018) and Bangor et al. (2009)).

The statistical comparison between the SUS scores and the neutral midpoint (Table 4) shows that most SUS items were rated significantly above or below the midpoint benchmark, with the exception of statement 7, which remained neutral. To account for the increased risk of Type I error arising from repeated significance testing across the ten individual items, a Holm-Bonferroni correction was additionally applied (Holm, 1979). All items that were significant under the uncorrected threshold remained significant following correction, with statement 7 remaining the sole non-significant item under both the uncorrected and corrected analyses, indicating that the observed pattern of item-level results is robust to correction for multiple comparisons. Analysis of the SUS usability and learnability subscales (Table 5) indicates that both were statistically significant; however, the usability construct ($M = 4.05$, $p < .001$) was rated significantly positively, whereas

the learnability construct ($M = 2.43$, $p = .001$) was rated significantly negatively. Internal consistency was excellent for usability (Cronbach's $\alpha = .94$) and acceptable for learnability (Cronbach's $\alpha = .73$) (Hair et al., 2010), supporting the two-factor structure of the SUS instrument. These findings suggest that, although the CF methodology and tool are perceived as usable, well-integrated, and confidence-inspiring, users experience a notable learnability burden, particularly in terms of requiring technical support and learning new content. This is also supported by the results for statement 7, which was not significantly different from the neutral midpoint, a similar observation found in various SUS tests (Bangor et al., 2008), indicating that participants were undecided about whether the CF methodology and tool would be rapidly learnable by typical users, in contrast to their own positive ratings of usability and confidence.

Table 5. Statistical analysis of the SUS usability and learnability subscales.

Construct	Items	Mean (vs. 3)	t-test p-value	Cronbach's α	Interpretation
Usability	Q1,2,3,5,6,7,8,9	4.05	<.001	.94	Excellent reliability, positive rating
Learnability	Q4, Q10	2.43	.001	.73	Acceptable reliability, negative rating

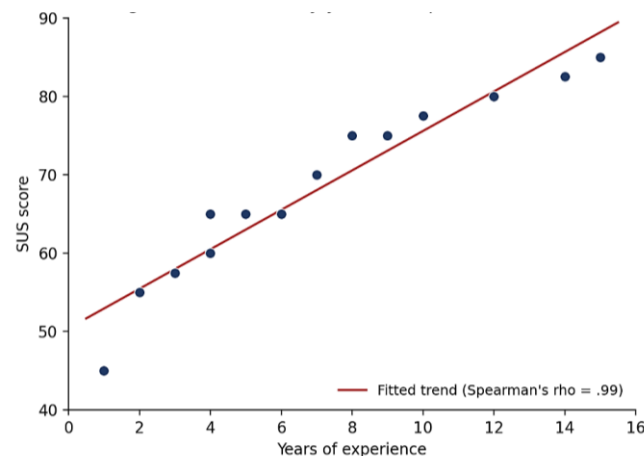


Figure 2. SUS score by years of experience ($n = 14$).

A Spearman rank-order correlation revealed a strong, significant positive relationship between years of experience and SUS scores, $r = .99$, 95% CI [.97, 1.00], $p < .001$ (Figure 2). A leave-one-out sensitivity analysis, in which the correlation was recalculated after excluding each participant in turn, showed rho ranging from .989 to .997, confirming that the association was not driven by a small number of influential observations. Participants with more years of experience gave significantly higher usability ratings for the CF methodology and tool, which aligns with the findings of Kortum and Bangor (2013).

The statistical comparison of the SUS scores with the midpoint results is also considered alongside the findings from the semi-structured interviews.

Semi-structured interviews

The main themes under which the interview data were coded are the four usability components: effectiveness, efficiency, satisfaction, and learnability, as well as refinements and improvements. The interview questions were structured around the definitions of the four usability components.

Effectiveness

The CF methodology and tool were broadly perceived as effective, as participants succeeded in completing the task accurately, completely, and appropriately. This is also confirmed by the statistical analysis of statements 5 and 6 ($M = 4.50$, $p < .001$), which relate to effectiveness (Wagner, 2023). However, across all three sub-themes, differences emerged between advanced and beginner users.

Task completion

This refers to participants' ability to complete the CF assessment for their respective NOCs. All participants were able to complete the CF assessment, although the extent of coverage varied due to experience and data availability. Lower-experience users struggled to complete all categories because of unclear inputs and missing data. For example, P7 was unable to complete *"staff commuting because I don't have that survey data."* Higher-experience users reported full completion and a logical structure, as P5 completed all tabs and highlighted the OA versus OG separation as a *"major effectiveness win."* Being unable to complete certain categories is not a major concern, as the proposed methodology does not adopt completeness as a guiding principle.

Accuracy and credibility of outputs

Participants expressed confidence that the calculated emissions accurately reflected reality. P13 stated that *"the kgCO₂e figures felt plausible compared to our previous consultant's report."* However, less experienced users did not show the same level of confidence, as they had nothing to compare the outputs against, with P12 stating *"I felt I got an output but was it the correct one?"* Participants with more experience showed greater confidence in the accuracy of the outputs, with P3 noting that *"the outputs felt broadly credible, though I still checked a few emission factors against our own records."* P5 stated that *"outputs felt well-structured and defensible"* and P6 asserted that *"I trusted the calculations once I got used to the formula structure."*

Application of methodological components

Some methodological elements were well received by participants, as they considered them useful for completing their CF assessment accurately. For example, P10 noted that *"the materiality thresholds (Beginner 20%) helped me justify excluding some smaller sources"*, while P5 praised the organisational boundary decision flowchart based on sustainability control as *"the most innovative part of the methodology"*. However, some of these methodological terms were too technical for intermediate and beginner participants. For example, P12 stated that *"Legal responsibility?" and 'Direct influence?' were too abstract"*, while P10 *"was confused by the 'Dynamic Materiality' concept"*. Although these terms are not explained in the tool, they are explained in detail in the methodology document. This indicates that the methodology and tool cannot be separated, as they complement each other.

Efficiency

Overall, participants perceived the CF methodology and tool as less efficient than effective. This is supported by the individual statistical scores for statements 2 ($M = 3.86$, $p = .002$), 3 ($M = 3.57$, $p = .014$), and 8 ($M = 4.07$, $p < .001$), which relate to efficiency (Hertzum, 2026). Under this main theme, the thematic analysis uncovered two sub-themes.

Time and effort required

This was highly dependent on prior experience and data readiness. Participants with less experience found that most of their effort was spent *"trying to understand what the sheet wanted"* (P7) or *"checking instructions rather than entering data"* (P12). For more experienced participants, the proportionality between the effort expended and the quality of outputs was acceptable. P5 stated that *"for a tool of this scope, the workload*

was reasonable” and P9 echoed this view, stating that “for the depth of reporting it requires, I found the process efficient”.

Participants found that most of the time was spent collecting the required data rather than using the CF methodology and tool. Beginner and intermediate users complained about the data collection burden, with P4 noting that “data gathering slowed everything down”, while P1 stated that “collecting data from different colleagues took longer than the spreadsheet itself” and P8 claimed that “the data collection side still took considerable effort”. In contrast, more experienced users found the data collection burden less problematic, as P14 stated that “we already track most of this. The tool didn’t create new data burdens”, while P9 noted “I have all our data in one place already”. This is consistent with Hertzum (2026), who suggests that users are more sensitive to perceived workload than to task completion time when evaluating system usability. Beginner users tend to associate longer task times with their own learning process rather than with usability issues, whereas experienced users are more likely to link usability directly to task completion time.

Navigation & input design

Participants agreed that the pre-filled emission factors and dropdown menus reduced input burden and increased efficiency. Three participants highlighted that “the design reduced unnecessary manual work” (P9), “the dropdowns and pre-set factors reduced a lot of manual effort” (P5) and “pre-filled factors and dropdowns reduced effort” (P13). However, other participants felt that “some steps felt repetitive” (P10) and “some tabs required more clicking around than expected” (P6). Current CF calculation tools suffer from low efficiency, as they require repetitive and complex manual input for multiple variables, which can be mitigated through design features such as dropdown menus (Abbott et al., 2019). While dropdown menus reduce the burden of user input (Quaglio et al., 2024), they can still be perceived as tedious and inefficient for frequent use (Paskaran et al., 2022) and may introduce classification errors when users lack technical knowledge (Rabbani & Wibowo, 2025). Nevertheless, given the wide variation in emission sources involved in CF calculations, the use of dropdown menus remains a practical approach to improving efficiency while minimising manual input (Espinoza Pérez et al., 2024).

On one hand, participants praised the methodological design of incorporating both activity-based and spend-based approaches, noting that “the dual-method approach is sensible” (P11) and “the dual-method approach is essential; it reduces the ‘all or nothing’ efficiency cost of a data gap” (P3). This aligns with studies suggesting that relying on a single approach may be insufficient for a comprehensive evaluation, and that combining both methods represents an effective strategy for robust carbon accounting while also providing a practical means of completing a timely inventory and addressing data gaps (Kennelly et al., 2019; Brook, 2025; Rosales & Dhakal, 2024; Xu & MacAskill, 2024). On the other hand, P13 highlighted a usability issue, stating that “having to flip to a separate (SB) sheet broke my workflow.” Although separating information across multiple tabs in carbon calculators can create usability challenges, it also addresses the broader issue of effectively communicating carbon emission data to users, which is the underlying purpose of separating complex methods (Sternberg, 2023).

Satisfaction

Participants demonstrated a high level of satisfaction with the CF methodology and tool. This is reflected in the statistical analysis of SUS statement 1 ($M = 4.29$, $p < .001$), indicating that participants would be willing to use the methodology and tool frequently, and SUS statement 9 ($M = 4.43$, $p < .001$), which showed that participants felt confident using them. However, responses were strongly polarised by experience. Overall, satisfaction with outputs was higher than satisfaction with the process of producing them. Two sub-themes emerged from the thematic analysis.

Satisfaction with CF methodology and tool

Participants found the experience of using the CF methodology and tool to be broadly positive, although less experienced users expressed frustration at times. Regarding the methodology, less experienced users did not criticise it, while more experienced users praised it as *“more robust than the ad hoc methods many organisations use”* (P14), as having a *“strong sense of methodological rigour”* (P11), and as *“credible, well thought through, and suitable for NOCs”* (P9). As for the tool, less experienced users found it *“a bit stressful at times”* (P2), *“frustrating and a bit intimidating”* (P7), and *“moderately frustrating because [he/she] wanted to use it independently but kept needing help”* (P12), while experienced users found it to be *“professional and appropriately detailed”* (P13), *“fit for purpose”* (P5), and *“credible, practical”* (P3).

This suggests that the underlying CF methodology is broadly accepted across different user groups; however, the implementation tool may impose a level of cognitive load or procedural complexity that exceeds the unaided capabilities of novice users. This aligns with cognitive load theory (Sweller, 2011), whereby experienced users possess automated schemas for structured analysis that free cognitive resources to recognise and utilise the tool's affordances, whereas novice users must devote greater effort to basic operations, increasing the likelihood of confusion and frustration.

Satisfaction with outputs

Participants were highly satisfied with the outputs for reporting purposes. More experienced users had complete confidence in using the outputs for both internal and external purposes, with P13 stating that *“the Summary sheet is board ready. The pie chart by scope is perfect. I would present this to our Executive Committee”*, while P14 noted that the outputs could confidently be used in *“both internal reports and when presenting to stakeholders for example the IOC or funders.”* However, beginner users were more cautious about using and sharing the outputs, although this was attributed more to input data quality than to a tool design limitation, with P8 stating that *“I wouldn't present these results to anyone. The problem is the quality of the underlying data, not the presentation”* and P10 noting that *“it was really a gap in my comfort with the data, not a fundamental flaw in the tool itself”*. This reflects a common understanding in carbon accounting: the quality of reporting is fundamentally linked to the quality of the underlying data sources used to complete emissions inventories (Baehr et al., 2024).

Learnability

Participants perceived learnability as the least positive aspect of the CF methodology and tool. This was reflected in the statistical analysis, where the two statements associated with the learnability construct received the lowest scores among all SUS items, with statement 4 ($M = 2.5$, $p = .003$) and statement 10 ($M = 2.36$, $p = .002$) both scoring below the midpoint benchmark, while statement 7 ($M = 3.14$, $p = .549$) received a neutral score. These findings indicate that participants did not perceive learning to use the CF methodology and tool as straightforward. The thematic analysis identified three sub-themes.

Complexity level

All participants felt that the CF methodology and tool were at the upper limit of complexity but deemed this appropriate given the seriousness and comprehensiveness of CF assessment. Beginner-to-intermediate participants found it more challenging, with P10 stating that the CF methodology and tool were *“designed by experts for experts”*, while P7 noted that *“it was actually a bit too advanced, especially for a beginner like me.”* Experienced participants appreciated the level of complexity but confirmed that novice users might struggle with it, as P3 mentioned, *“it's exactly what's missing from the GHG Protocol for a sector like sport but I have 12 years of experience. I worry for the beginners”*, and P11 reiterated that *“a beginner would struggle without significant support.”*

This consensus suggests that the complexity is not accidental but deliberate, reflecting the demanding nature of the assessment context (Du et al., 2024). However, the subsequent divergence by expertise level reveals a potential exclusion risk, indicating that novice users may face a steep learning curve.

Learning curve

The initial orientation to the CF methodology and tool differed between participants. Experienced participants quickly understood the structure and began data entry immediately, with participants stating, *"I picked it up fairly quickly"* (P6) and *"I was entering data within five minutes"* (P9), while less experienced participants required more time to interpret the methodology, understand the spreadsheet structure, familiarise themselves with the tabs and workflows, and consult the guidance materials. Reflecting this, participants made comments such as *"I needed time to get going, especially understanding the more technical methodological terms"* (P10), *"it took me a while of clicking through tabs to see the pattern"* (P12), and *"I spent considerable time reading the Index & Data Guide sheet"* (P7).

However, in-tool learning aids such as the dropdowns, colour-coded cells and the Index & Data Guide sheet were consistently valued by all participants and helped with improving the learning curve for some participants, as P12 explained that *"the dropdowns were very useful. The colour-coding was useful after I read the Introduction sheet"*, P3 highlighted that *"the dropdowns and the colour-coded cells are critical for learnability"*, and P9 noted that *"the guide and structure made the learning process easier."*

Participants believed that competency with the CF methodology and tool would improve substantially through repeated use and continued exposure. They felt that the consistent structure of the spreadsheet and repeated annual application would support retention and reduce future re-learning efforts. Participants indicated that *"a user would build competency over time"* (P5), *"the tool's consistent structure really does help reinforce familiarity"* (P14), and *"same tabs, same logic is the tool's best feature for long-term retention"* (P1).

Overall, the CF methodology and tool impose a steep initial learning curve for beginner users, as they are not entirely self-explanatory. As users build competency over time, the design and structure of the CF methodology and tool appear to be optimised for longitudinal use rather than initial encounter. It is common for the orientation phase of encountering a new system to be dominated by learnability concerns (Dirin et al., 2023; Karapanos et al., 2009); however, these initial concerns may have limited long-term significance, as user experience evolves over time through the ongoing relationship between the user and the system (Kujala et al., 2019). Nielsen (2010) distinguishes between walk-up-and-use systems (e.g. museum information systems) and professional systems, such as the proposed CF methodology and tool, arguing that professional systems may involve a steeper learning curve. However, once users overcome the initial learning effort, performance levels can become substantially higher, reflecting a trade-off between initial ease of use and the long-term performance potential offered by professional systems.

Prior knowledge requirements

Participants reported that successful use of the CF methodology and tool required some foundational prior knowledge, particularly in carbon accounting concepts and terminology, GHG Protocol scope classifications, and basic Excel functionality. Although the guidance materials supported users throughout the process, participants felt that complete beginners would struggle to use the system independently without initial training, technical guidance, or prior exposure to sustainability reporting. P8 explained that the CF methodology and tool *"assumes a certain level of familiarity with carbon accounting concepts"*, while, when asked about drawing on prior knowledge, P6 answered *"quite a lot, honestly. Without that foundation, the tool would be very confusing"*, and P2 similarly replied *"a lot. You need to know what Scope 1, 2, and 3 are,*

or what an emission factor is.” This was further reinforced by P13, who stated that *“this tool is not for a complete novice without support. It assumes you understand the GHG Protocol’s categorisation.”*

Since the CF methodology and tool were developed as tailored adaptations and improvements of existing methodologies and prior knowledge, the use of common terminology such as scopes, boundaries, and emission factors is necessary to maintain alignment with existing practice and preserve familiarity (Faisal & AlQouz, 2023). This reduces the need for users to learn entirely new terminology, which could otherwise further steepen the learnability curve for both novice and experienced users alike (Forsey et al., 2025). Eleftheriadis and Anagnostopoulou (2024) stated that they deliberately avoided using CF terminology in their tool because doing so would require additional user training and potentially discourage adoption. However, they did not provide evidence demonstrating that this approach improved the usability or learnability of the tool. In contrast, Rodriguez-Jimenez et al. (2025) incorporated common CF terminology into their tool for emergency departments and found that this did not negatively affect usability or learnability, as participants with no specific CF knowledge or experience were still able to use the tool effectively. Furthermore, the concern that novice users would struggle to use the CF methodology and tool was primarily raised by experienced participants, whereas the novice users themselves did not express the same view.

Refinements and improvements

Participants provided a range of recommendations aimed at improving the effectiveness, efficiency, satisfaction, and learnability of the CF methodology and tool. Although most participants viewed the methodology and tool positively overall, the suggested refinements primarily focused on reducing complexity for less experienced users, improving workflow efficiency, strengthening validation processes, and enhancing user guidance and feedback mechanisms.

Several participants proposed refinements to improve the quality and accuracy of the results produced by the methodology and tool. One suggestion was to allow the use of country-specific emission factors instead of pre-populated emission factors. For example, P3 recommended allowing users to *“input their own country-specific electricity grid factor”*, noting that the default UK electricity factor was not representative of NOCs operating in other countries. This is a valid refinement that enhances result accuracy; however, the tool already accommodates such customisation, as all emission factor cells are user-editable, allowing users to overwrite the default emission factor values with their own supplier- or country-specific factors. Participants also highlighted the need for stronger validation mechanisms to reduce user errors. P7 suggested that the tool should automatically flag missing or incomplete data inputs, explaining that *“if I enter a ‘0’ or leave a cell blank for a material source, the Summary sheet should flag it in red”*.

Beyond the specific case of editable emission factors, P3’s comment points to a broader consideration for the international applicability of the CF methodology and tool. As the methodology and tool are intended for deployment across a globally diverse population of NOCs, the pre-populated default values, largely derived from UK sources such as DESNZ and DEFRA, may not consistently reflect the electricity-grid characteristics, regulatory reporting requirements, or data-availability conditions of NOCs operating in other countries. Electricity grid emission factors in particular vary substantially between countries depending on the underlying generation mix, and organisations operating in jurisdictions with limited access to nationally published factors, or with less mature emissions-reporting infrastructure, may face a greater burden in sourcing appropriate replacement values than the customisation feature alone can resolve. Similarly, differences in national or regional regulatory environments may shape what data organisations are able or required to collect, which in turn affects the completeness and consistency of CF assessments across the NOC population. While allowing user-editable emission factors provides a mechanism for adaptation, it does not by itself guarantee

that all users have equal access to the country-specific data needed to exercise this customisation. Future refinements of the methodology and tool could therefore consider incorporating a curated library of regionally appropriate default factors, alongside guidance on sourcing nationally verified alternatives, to reduce this variation in implementation burden across different national contexts.

Participants also identified several opportunities to improve efficiency and reduce the burden associated with using the tool. One of the most common suggestions involved simplifying the spreadsheet structure and reducing the number of tabs. Participants felt that having separate spend-based sheets for each emission source created unnecessary complexity and interrupted workflow continuity. Reflecting this, P8 recommended consolidating spend-based methods within the main activity-based sheets by adding a “*small, collapsible spend-based fallback section at the bottom of the main tab*”.

Participants further highlighted several improvements that could enhance the overall user experience and satisfaction associated with the methodology and tool. Less experienced users, in particular, expressed a need for clearer feedback and reassurance during the completion process. P12 recommended introducing visual completion indicators, such as a “*green checkmark*” after finishing a tab, explaining that they were often “*never sure if [they were] finished*”. Participants also suggested improving the language used within the uncertainty assessment outputs. Although users understood the rationale behind labels such as “*Poor*”, some participants perceived the wording as discouraging. P9 suggested replacing terms such as “*Poor*” with more constructive wording, such as “*Needs Improvement*”, alongside guidance explaining how users could improve their data quality scores.

In addition, participants identified several refinements aimed at reducing the learning burden and improving accessibility for beginners. A common recommendation involved providing more interactive onboarding and training support. P2 proposed including a “*Demo Data*” function that would populate the tool with realistic example data, allowing users to understand how information flows throughout the spreadsheet before entering their own data. Similarly, P5 recommended developing a short introductory training video demonstrating the overall workflow of the methodology and tool, including navigation between tabs, the purpose of colour-coded cells, and the relationship between OA and OG reporting. Participants believed that these additional support materials would significantly reduce the initial learning curve, particularly for users with limited prior experience in carbon accounting.

Overall, the refinements suggested by participants indicate that while the CF methodology and tool were generally perceived as comprehensive and professionally designed, improvements are required to enhance accessibility, reduce user burden, and support less experienced users. Many of the proposed refinements focused not on changing the underlying methodology itself, but rather on improving usability, guidance, automation, and user confidence during implementation.

Limitations

A key limitation of this study is the relatively small sample size ($n = 14$), which restricts the extent to which the findings can be generalised beyond the study context. Although small samples are common in usability evaluation research, particularly where in-depth qualitative insights are prioritised (Hennink & Kaiser, 2022; Wutich et al., 2024), the limited number of participants may not fully capture the diversity of experiences, technical competencies, organisational structures, and carbon accounting practices present across NOCs. Therefore, the findings should be interpreted as context-specific rather than statistically representative of all NOCs. Nevertheless, the sample size was considered sufficient for exploratory usability testing and

qualitative inquiry, where the objective is often to identify usability issues and generate practical insights rather than achieve population-level generalisability (Vasileiou et al., 2018).

The use of the SUS also presents several limitations. While the SUS is widely recognised for its reliability and simplicity, it provides only a broad measure of perceived usability and does not identify the specific causes of usability problems (Brooke, 1996). The instrument is also based on subjective self-reported perceptions, which may be influenced by participants' prior experience, expectations, or response tendencies (Lewis, 2018).

The sample also varied considerably in the type, size, and resourcing of the NOCs represented. Participants were drawn from NOCs across Africa, the Americas, Asia, and Europe, ranging from small island and mountain-state committees with limited administrative capacity to larger, well-resourced NOCs with dedicated sustainability functions. Some participants worked within NOCs where one or two staff members covered sustainability alongside a broader remit, often on a partly voluntary basis, while others worked within NOCs that maintained specialist environmental roles and established data-management infrastructure. This range in organisational maturity, rather than participants' individual carbon accounting experience alone, is likely to have shaped perceptions of the CF methodology and tool. An NOC with limited administrative capacity may face practical barriers to sustained implementation regardless of the tool's usability, since even a highly usable tool still depends on staff time and available activity data to complete a CF assessment. This organisational dimension is examined further in the discussion of organisational data readiness in the following section and is a further reason why the present findings should be interpreted as context-specific rather than representative of all NOCs.

A related limitation is that the study relies on participant perceptions rather than objective usability metrics, such as task completion rates, error frequency, support requests, or completion time. Self-reported measures of effectiveness, efficiency, satisfaction, and learnability capture how participants experienced the CF methodology and tool, but they do not necessarily correspond to observed performance. A participant who reports low confidence in a completed task may still have produced an accurate output, while a participant who reports high satisfaction may have made undetected errors. Future evaluations of the CF methodology and tool should therefore incorporate objective performance measures, including timed task completion, error rates, and the frequency of support requests during use, to complement the perception-based findings reported in this study.

The semi-structured interviews are similarly subject to methodological limitations, particularly interviewer bias (Adams, 2015). The flexibility of semi-structured interviews allows probing and clarification; however, the interviewer may unintentionally influence responses through question phrasing, tone, prompts, or non-verbal cues. Given the researcher's familiarity with the proposed CF methodology and tool, there is also a risk that personal assumptions or expectations may have shaped the direction of discussions or interpretation of participant responses. To minimise these risks, reflexive practices were adopted throughout the research process, including maintaining awareness of potential biases, using a structured interview guide, and encouraging participants to provide open and critical feedback (Braun & Clarke, 2021).

Thematic deductive analysis also introduces potential limitations. Because the analysis was guided by predefined usability themes and concepts, there is a risk of confirmation bias (Braun & Clarke, 2006), whereby data may be interpreted in ways that reinforce existing assumptions or expected findings. Deductive approaches may also overlook unexpected insights or emerging themes that fall outside the predetermined

analytical framework (Terry et al., 2017). As a result, some novel or unanticipated participant perspectives may not have been explored in sufficient depth.

Toward a conceptual model of adoption

The usability findings reported above also carry a broader theoretical implication for how the CF methodology and tool might progress beyond this evaluation toward routine adoption across NOCs. This section proposes a conceptual model that links four constructs identified through this study, tool usability, prior carbon accounting knowledge, organisational data readiness, and perceived relative advantage, to adoption outcomes. The model draws on the Technology Acceptance Model (Davis, 1989) and the Diffusion of Innovations framework (Rogers, 2003), situating the present findings within the broader implementation-science literature on how new methods and tools move into routine organisational use.

The Technology Acceptance Model (TAM) holds that adoption intention is shaped by two beliefs: perceived usefulness and perceived ease of use (Davis, 1989). The present findings map closely onto this distinction. Perceived usefulness was high and consistent across the sample. Participants at every experience level described the methodology as more robust than the ad hoc practices their NOCs currently rely on, and the tool's outputs were repeatedly judged credible enough to present to an Executive Committee or external funder. Perceived ease of use followed a markedly different pattern. It was the lowest-scoring SUS construct overall, and the one most strongly moderated by prior experience: participants with limited carbon accounting background needed considerably longer to orient themselves and expressed less confidence that their entries were correct, while experienced participants reported beginning data entry within minutes. Read through TAM, this divergence indicates that resistance to adopting the CF methodology and tool is unlikely to stem from doubts about its value. It is more likely to stem from the effort required to use it without prior carbon accounting knowledge.

Diffusion of Innovations (DOI) offers a complementary account of the same divergence. Rogers (2003) identifies five attributes that determine how quickly an innovation is adopted: relative advantage, compatibility, complexity, trialability, and observability. The CF methodology exhibited high relative advantage over the ad hoc approaches it is designed to replace, and high compatibility with GHG Protocol terminology and existing reporting conventions already familiar to sustainability practitioners. Both attributes reinforce the strong perceived usefulness findings reported above. The methodology and tool also scored high on complexity; however, the one DOI attribute inversely related to adoption. The refinements participants proposed can be read as attempts to compensate for this complexity along the two remaining attributes. A "Demo Data" function to rehearse the tool before entering live figures would increase trialability. Automated flags for missing entries and clearer completion indicators would increase observability of progress during use. Neither refinement touches the underlying methodology, which participants regarded as sound; both are aimed at making a complex but valuable method easier to trial and track.

A fourth construct, organisational data readiness, sits outside both TAM and DOI but emerged clearly from the interview data as a determinant of adoption independent of the tool's design. Participants who described their NOC as already tracking most relevant activity data (P9, P14) treated data collection as a non-issue and located all remaining effort in the calculation itself. Participants without this infrastructure (P1, P4, P8) identified data collection, rather than the spreadsheet, as the main source of effort, and attributed their lower confidence in the outputs to this data quality question rather than to any flaw in the tool. This is consistent with organisational readiness for change more broadly (Weiner, 2009), where the availability of resources, in this case usable emissions data, determines whether an otherwise acceptable innovation can be implemented in practice.

Together, these four constructs point to a conceptual model in which adoption of the CF methodology and tool is jointly determined by the perceived relative advantage and compatibility of the methodology itself (largely fixed, and already favourable), the ease of use of the tool as experienced through the lens of the user's prior carbon accounting knowledge (variable, and the primary barrier for novice users), and the organisational data readiness of the adopting NOC (variable, and independent of both the methodology and the tool). Perceived usefulness alone is not a sufficient condition for adoption in this context, since the methodology was judged valuable by novice and experienced participants alike. Ease of use and organisational data readiness are the constructs most likely to determine whether that judgement translates into sustained use, and both point to the same practical implication. Increasing the trialability and observability of the tool and supporting NOCs in establishing baseline data collection practices at the organisational level, are the two levers most likely to convert a well-regarded methodology into a widely adopted standard.

CONCLUSION

This study evaluated the proposed CF methodology and tool for NOCs, addressing three research questions through a mixed-methods usability assessment combining the System Usability Scale and semi-structured interviews. In doing so, the study addressed a key gap identified within the literature, namely the limited empirical testing of CF methodologies and tools by independent end-users in real organisational settings. Existing studies have largely prioritised methodological development and calculation accuracy, while giving comparatively limited attention to usability, learnability, and practical implementation challenges. By involving intended end-users from different NOCs with varying levels of experience, this study provided a more comprehensive evaluation of both the methodological robustness and practical applicability of the proposed CF methodology and tool.

In response to RQ1, the CF methodology and tool achieved an overall SUS score of 68.4, which is only slightly above the accepted average benchmark of 68 (Sauro & Lewis, 2016). This score places the system within the “*marginal-high*” acceptability range, the “*good*” adjective rating (Bangor et al., 2009), and grade C on the curved grading scale (Sauro & Lewis, 2016), indicating a level of perceived usability that, while broadly satisfactory, remains below the threshold typically associated with high-performing systems. Importantly, a strong positive correlation was identified between years of experience and SUS scores ($r = .99$, $p < .001$), highlighting that usability perceptions were substantially shaped by participants' prior familiarity with carbon accounting concepts and tools. This finding underscores that the CF methodology and tool, as a professional-grade instrument, may not be equally accessible to all user profiles without additional support. The strength of this correlation, confirmed by a leave-one-out sensitivity analysis (ρ range = .989–.997) and unlikely to reflect a small number of influential observations, is consistent with the relatively narrow and evenly distributed range of experience levels in the sample (1 to 15 years), which increases the likelihood of detecting a smooth, near-monotonic relationship even at $n = 14$. A correlation of this magnitude should nonetheless be interpreted with the sample size in mind, and future research with a larger and more heterogeneous sample is needed to confirm whether the relationship holds at this strength.

With respect to RQ2, the four usability dimensions of effectiveness, efficiency, satisfaction, and learnability yielded divergent findings. Effectiveness emerged as the strongest dimension: all participants succeeded in completing, at least partially, a CF assessment for their respective NOC, and outputs were broadly regarded as credible and defensible, particularly by more experienced users. Efficiency was perceived as moderate overall, with the data collection burden identified as the primary source of effort rather than the tool's design itself. Navigation aids such as dropdown menus and pre-populated emission factors were consistently valued as efficiency-enhancing features, though the separation of activity-based and spend-based calculation sheets

was identified as an interruption to workflow continuity for some participants. Satisfaction was the second strongest dimension, with participants expressing a high level of confidence in the outputs, particularly for reporting and stakeholder communication purposes. However, satisfaction was notably polarised by experience level, with less experienced users reporting frustration during the completion process, which is consistent with cognitive load theory (Sweller, 2011). Learnability emerged as the most significant area of concern, with the two associated SUS statements scoring below the neutral midpoint and representing the lowest-rated items across the entire instrument. While in-tool learning aids, including colour-coded cells, dropdown menus, and the Index & Data Guide sheet, were valued by participants, the CF methodology and tool were not perceived as self-explanatory, particularly for novice users, and required a degree of prior knowledge in carbon accounting and GHG Protocol terminology to be used independently and effectively.

Addressing RQ3, participants provided a range of practical refinements aimed at improving the usability, functionality, and practicality of the CF methodology and tool. Key recommendations included introducing automated data validation and flagging mechanisms to reduce user error, consolidating spend-based calculation sheets to improve workflow continuity, developing interactive onboarding materials such as a “Demo Data” function and introductory training video, and adopting more constructive language within uncertainty assessment outputs. While many of these suggestions were directed at reducing the learnability burden for beginner users, it is noteworthy that most refinements did not call for changes to the underlying methodology itself, which was broadly accepted across user groups as rigorous, credible, and fit for purpose. This distinction between methodological robustness and tool usability reinforces the critical point, consistent with the literature (Xia et al., 2025; Harangozo & Szigeti, 2017), that methodological soundness alone does not guarantee practical usability.

Taken together, the findings of this study demonstrate that the proposed CF methodology and tool represent a meaningful contribution to carbon accounting practice within the sport sector, offering a structured, sector-specific approach that addresses gaps in current NOC reporting frameworks. The methodology was consistently regarded as more robust than the ad hoc approaches currently employed by many organisations, and the tool was considered broadly appropriate for its intended purpose. Nonetheless, the findings also indicate that improvements to onboarding support, user guidance, and navigational design are necessary to ensure that the methodology and tool can be adopted equitably across the full spectrum of user competencies present within the NOC community. Addressing these refinements would not only improve the immediate usability of the tool but would also support the longer-term legitimacy, consistency, and scalability of the proposed CF standard, which have been identified as essential conditions for sector-wide adoption (Marinelli et al., 2021; Strohmeier et al., 2025). Future iterations of the methodology and tool should therefore incorporate the refinements identified through this testing process, with subsequent evaluation cycles used to assess the impact of those changes on usability outcomes.

Future research

Building on the findings and limitations reported above, four directions are proposed for future research on the CF methodology and tool.

First, comparative evaluations of the CF methodology and tool against generic corporate carbon calculators would help establish whether the usability challenges identified here are specific to the design choices made in this tool or characteristic of carbon accounting more broadly.

Second, longitudinal studies tracking the same NOCs across multiple annual reporting cycles would help distinguish onboarding effects from durable usability barriers, addressing the possibility, raised in the

Discussion, that some of the learnability difficulties reported here reflect the short duration of testing rather than a persistent design flaw.

Third, validation of the CF methodology's calculated outputs against independently audited GHG inventories, where available, would test the accuracy of the tool's calculations directly, complementing the perception-based usability evidence gathered in this study and addressing the absence of objective performance metrics noted in the limitations.

Fourth, dedicated research into organisational capability, including staffing levels, data-management infrastructure, and sustainability governance structures across NOCs, would extend the organisational data readiness construct proposed above and clarify which organisational characteristics most strongly predict successful and sustained implementation of the CF methodology, independent of the tool's design.

Pursued together, these four directions would extend the present study from a single-cycle usability evaluation toward a fuller evidence base for standardising carbon accounting practice across the Olympic Movement.

FUNDING

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

CONFLICT OF INTEREST

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this manuscript.

AI USE DISCLOSURE

In accordance with current publishing ethics and transparency recommendations, artificial intelligence (AI) tools were used solely to assist with translation and language editing, with the aim of improving clarity and readability. No AI tools were used in the generation of scientific content, including the study design, data collection, analysis, interpretation of results, or the formulation of conclusions. The authors retain full responsibility for the content of the manuscript and confirm its originality, integrity, and accuracy.

REFERENCES

- Abbott, J., Gao, G., & Shih, P. (2019, March). Creen: a carbon footprint calculator designed for calculation in context. In *International Conference on Information* (pp. 769-776). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-15742-5_72
- Abran, A., Khelifi, A., Suryn, W., & Seffah, A. (2003). Usability meanings and interpretations in ISO standards. *Software Quality Journal*, 11(4), 325-338. <https://doi.org/10.1023/A:1025869312943>
- Adams, W. C. (2015). *Conducting semi-structured interviews*. In K. E. Newcomer, H. P. Hatry, & J. S. Wholey (Eds.), *Handbook of practical program evaluation* (4th ed., pp. 492-505). Wiley. <https://doi.org/10.1002/9781119171386.ch19>
- Adeoye-Olatunde, O. A., & Olenik, N. L. (2021). Research and scholarly methods: Semi-structured interviews. *Journal of the American College of Clinical Pharmacy*, 4(10), 1358-1367. <https://doi.org/10.1002/jac5.1441>

- Adhikari, N., Li, H., & Gopalakrishnan, B. (2025). A bibliometric and systematic review of carbon footprint tracking in cross-sector industries: emerging tools and technologies. *Sustainability*, 17(9), 4205. <https://doi.org/10.3390/su17094205>
- Aldoihi, S. (2020). *Product Usability Driven System Engineering*. The Centre for Direct Scientific Communication.
- Almazroi, A. A. (2021). A systematic mapping study of software usability studies. *International Journal of Advanced Computer Science and Applications*, 12(9). <https://doi.org/10.14569/IJACSA.2021.0120927>
- Aresta, M., Beça, P., Siso, D., & Quintero, E. (2024, November). Promoting sustainable behaviours in organizations: The IDEA System. In *International Conference on Design and Digital Communication* (pp. 163-174). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-77566-6_12
- Asam, D., Jürgen, E., & Friedl, G. (2025). How carbon accounting supports corporate decarbonization. *Foundations and Trends® in Accounting*, 19(3-4), 46-77. <https://doi.org/10.1561/14000000080-3>
- Augoye, O., Muyiwa-Ajayi, T. P., & Sobowale, A. (2024). The effectiveness of carbon accounting in reducing corporate carbon footprints. *International Journal of Multidisciplinary Research and Growth Evaluation*, 5(1), 1364-1371. <https://doi.org/10.54660/IJMRGE.2024.5.1.1364-1371>
- Azarkamand, S., Ferré, G., & RM, D. (2021, September). *Calculating Carbon Footprint in ports through a new Standard Tool: case study applications*. In 17th International Conference on Environmental Science and Technology. Athens, Greece.
- Azhari, A. E. A., Chai, J., & Anderson, C. (2022). reflections of conducting online interviews during the pandemic: benefits, challenges and practical solutions. *Malaysian Journal of Qualitative Research*, 8(1). <https://doi.org/10.61211/mjqr080103>
- Baehr, J., Zenglein, F., Sonnemann, G., Lederer, M., & Schebek, L. (2024). Back in the driver's seat: how new EU greenhouse-gas reporting schemes challenge corporate accounting. *Sustainability*, 16(9), 3693. <https://doi.org/10.3390/su16093693>
- Ball, K. R., Burke, I. C., Collins, D., Kruger, C. E., & Yorgey, G. (2023). Digging deeper: Assessing the predictive power of common greenhouse gas accounting tools for soil carbon sequestration under organic amendment. *Journal of Cleaner Production*, 429, 139448. <https://doi.org/10.1016/j.jclepro.2023.139448>
- Bangor, A., Kortum, P. T., & Miller, J. T. (2008). An empirical evaluation of the system usability scale. *Intl. Journal of Human-Computer Interaction*, 24(6), 574-594. <https://doi.org/10.1080/10447310802205776>
- Bangor, A., Kortum, P., & Miller, J. (2009). Determining what individual SUS scores mean: Adding an adjective rating scale. *Journal of Usability Studies*, 4(3), 114-123.
- Barendregt, W., Biørn-Hansen, A., & Andersson, D. (2020). Users' experiences with the use of transaction data to estimate consumption-based emissions in a carbon calculator. *Sustainability*, 12(18), 7777. <https://doi.org/10.3390/su12187777>
- Bearman, M. (2019). Eliciting rich data: A practical approach to writing semi-structured interview schedules. *Focus on Health Professional Education: A Multi-Professional Journal*, 20(3), 1-11. <https://doi.org/10.11157/fohpe.v20i3.387>
- Bekaroo, G., Roopowa, D., & Bokhoree, C. (2019, September). Mobile-based carbon footprint calculation: insights from a usability study. In *2019 Conference on Next Generation Computing Applications (NextComp)* (pp. 1-6). IEEE. <https://doi.org/10.1109/NEXTCOMP.2019.8883622>
- Bevan, N., Carter, J., & Harker, S. (2015, July). ISO 9241-11 revised: What have we learnt about usability since 1998?. In *International conference on human-computer interaction* (pp. 143-151). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-20901-2_13

- Binyamin, S., Rutter, M., & Smith, S. (2016). The utilization of system usability scale in learning management systems: a case study of Jeddah Community College. In *ICERI2016 Proceedings* (pp. 5314-5323). IATED. <https://doi.org/10.21125/iceri.2016.2290>
- Birnik, A. (2013). An evidence-based assessment of online carbon calculators. *International Journal of Greenhouse Gas Control*, 17, 280-293. <https://doi.org/10.1016/j.ijggc.2013.05.013>
- Blattgerste, J., Behrends, J., & Pfeiffer, T. (2022, June). A web-based analysis toolkit for the system usability scale. In *Proceedings of the 15th International Conference on Pervasive Technologies Related to Assistive Environments* (pp. 237-246). <https://doi.org/10.1145/3529190.3529216>
- Bowen, F., & Wittneben, B. (2011). Carbon accounting: Negotiating accuracy, consistency and certainty across organisational fields. *Accounting, Auditing & Accountability Journal*, 24(8), 1022-1036. <https://doi.org/10.1108/09513571111184742>
- Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2), 77-101. <https://doi.org/10.1191/1478088706qp0630a>
- Braun, V., & Clarke, V. (2021). Thematic analysis: A practical guide. Sage. <https://doi.org/10.4135/9781036232078.n9>
- Breton, M. L., & Aggeri, F. (2018). Counting before acting? The performativity of carbon accounting called into question - Calculation acts and dispositifs in a big French construction company. HAL (Le Centre Pour La Communication Scientifique Directe).
- Brook, H. (2025). Greenhouse gas (GHG) inventory development and management. In *Cases and Exercises on Sustainability Accounting* (pp. 129-140). Edward Elgar Publishing. <https://doi.org/10.4337/9781035330607.00020>
- Brooke, J. (1996). SUS: A "quick and dirty" usability scale. In P. W. Jordan, B. Thomas, I. L. McClelland, & B. Weerdmeester (Eds.), *Usability evaluation in industry* (pp. 189-194). Taylor & Francis.
- Brooke, J. (2013). SUS: a retrospective. *Journal of Usability Studies*, 8(2).
- Büchs, M., Bahaj, A. S., Blunden, L., Bourikas, L., Falkingham, J., James, P., ... & Wu, Y. (2018). Promoting low carbon behaviours through personalised information? Long-term evaluation of a carbon calculator interview. *Energy Policy*, 120, 284-293. <https://doi.org/10.1016/j.enpol.2018.05.030>
- Cairns-Lee, H., Lawley, J., & Tosey, P. (2022). Enhancing researcher reflexivity about the influence of leading questions in interviews. *The Journal of Applied Behavioral Science*, 58(1), 164-188. <https://doi.org/10.1177/00218863211037446>
- Castillo-Montoya, M. (2016). Preparing for interview research: The interview protocol refinement framework. *The Qualitative Report*, 21(5). <https://doi.org/10.46743/2160-3715/2016.2337>
- Chatty, T., Qu, Y., Ba-Sabaa, H. H., & Murnane, E. L. (2021). Examining the user experience of life cycle assessment tools and their ability to cater to Ecodesign in early-stage product development practice. *Proceedings of the Design Society*, 1, 1441. <https://doi.org/10.1017/pds.2021.405>
- Christy, A. P., Elnahass, M., Amézaga, J. M., Browne, A., & Heidrich, O. (2023). A dynamic framework to align company climate reporting and action with global climate targets. *Business Strategy and the Environment*, 33(4), 3103. <https://doi.org/10.1002/bse.3635>
- Chyung, S. Y., Barkin, J. R., & Shamsy, J. A. (2018). Evidence-based survey design: The use of negatively worded items in surveys. *Performance Improvement*, 57(3), 16-25. <https://doi.org/10.1002/pfi.21749>
- Ciocoi-Pop, D. R., Boros-Balint, I., & Baciú, M. A. (2025). Sport and Sustainability Reporting: Assessing the Readiness of Sports Clubs in the Balkans. *Amfiteatru Economic*, 27(70), 922-938. <https://doi.org/10.24818/EA/2025/70/921>
- Clifford, N., Cope, M., Gillespie, T., & French, S. (Eds.). (2016). Key methods in geography. Sage.
- Corsini, M., Scopigno, R., Calori, L., Graf, H., Pletinckx, D., Pescarin, S., ... & Denard, H. (2012). Analysis of Visualisation and Interaction Tools Authors. *The Centre for Direct Scientific Communication*.

- Cridland, E. K., Jones, S. C., Caputi, P., & Magee, C. A. (2015). Qualitative research with families living with autism spectrum disorder: Recommendations for conducting semistructured interviews. *Journal of Intellectual and Developmental Disability*, 40(1), 78-91. <https://doi.org/10.3109/13668250.2014.964191>
- Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, 13(3), 319-340. <https://doi.org/10.2307/249008>
- Dirin, A., Nieminen, M., & Laine, T. H. (2023). Feelings of being for mobile user experience design. *International Journal of Human-Computer Interaction*, 39(20), 4059-4079. <https://doi.org/10.1080/10447318.2022.2108964>
- Du, J., Zhong, Z., Shi, Q., Cao, Y., Yang, C., & Wang, L. (2024). A review of household carbon footprint: measurement methods, evolution and emissions assessment. *Carbon Footprints*, 3(3), N-A. <https://doi.org/10.20517/cf.2024.08>
- Eleftheriadis, I., & Anagnostopoulou, E. (2024). Developing a tool for calculating the carbon footprint in SMEs. *Sustainability*, 16(5), 1905. <https://doi.org/10.3390/su16051905>
- Eppich, W. J., Gormley, G. J., & Teunissen, P. W. (2019). In-depth interviews. In *Healthcare simulation research: A practical guide* (pp. 85-91). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-26837-4_12
- Espinoza Pérez, L., Mauricio, C., Araya, C., Contreras, R., Suazo, F., Velastín, N., ... & Vásquez, Ó. C. (2024, October). Advancing Food Sustainability: A Prototype Platform for Informed Decision-Making. In *International Conference on Advanced Research in Technologies, Information, Innovation and Sustainability* (pp. 250-263). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-83210-9_19
- Faisal, M., & AlQouz, E. H. (2023, July). Utilizing Crowdsourced Heuristic Evaluation in the Assessment of User Experience for Online Tools. In *International Conference on Human-Computer Interaction* (pp. 132-137). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-49212-9_18
- Fathiyyah, D., Diani, M. D. S., & Saputri, Z. A. (2022, November). Usability Evaluation on Life Insurance Application Using System Usability Scale and ISO 9241-11. In *2022 8th International HCI and UX Conference in Indonesia (CHLuXiD)* (Vol. 1, pp. 94-99). IEEE. <https://doi.org/10.1109/CHLuXiD57244.2022.10009774>
- Felipe, D. A., Kalinowski, M., Graziotin, D., & Natividade, J. C. (2023). Psychometric instruments in software engineering research on personality: Status quo after fifty years. *Journal of Systems and Software*, 203, 111740. <https://doi.org/10.1016/j.jss.2023.111740>
- Finstad, K. (2010). The usability metric for user experience. *Interacting with Computers*, 22(5), 323-327. <https://doi.org/10.1016/j.intcom.2010.04.004>
- Forsey, H., Leahy, D., Fields, B., Minocha, S., Attfield, S., & Snell, T. (2025). Designing for learnability: Improvement through layered interfaces. *Ergonomics in Design*, 33(3), 135-141. <https://doi.org/10.1177/10648046241273291>
- Gao, M., Kortum, P., & Oswald, F. L. (2020). Multi-language toolkit for the system usability scale. *International Journal of Human-Computer Interaction*, 36(20), 1883-1901. <https://doi.org/10.1080/10447318.2020.1801173>
- Golondrino, G. E. C., Alarcón, M. A. O., & Martínez, L. M. S. (2023). Determination of the satisfaction attribute in usability tests using sentiment analysis and fuzzy logic. *International Journal of Computers Communications & Control*, 18(3). <https://doi.org/10.15837/ijccc.2023.3.4901>
- Goodell, L. S., Stage, V. C., & Cooke, N. K. (2016). Practical qualitative research strategies: Training interviewers and coders. *Journal of Nutrition Education and Behavior*, 48(8), 578-585. <https://doi.org/10.1016/j.jneb.2016.06.001>

- Goundar, M. S., Kumar, B. A., & Ali, A. S. (2024). Development of usability guidelines: a systematic literature review. *International Journal of Human-Computer Interaction*, 40(5), 1298-1316. <https://doi.org/10.1080/10447318.2022.2141009>
- Graziotin, D., Fagerholm, F., Wang, X., & Abrahamsson, P. (2017, May). Consequences of unhappiness while developing software. In *2017 IEEE/ACM 2nd International Workshop on Emotion Awareness in Software Engineering (SEmotion)* (pp. 42-47). IEEE. <https://doi.org/10.1109/SEmotion.2017.5>
- Greene, K. K., Kelsey, J. M., & Franklin, J. M. (2016). Measuring the usability and security of permuted passwords on mobile platforms. *US Department of Commerce, National Institute of Standards and Technology*. <https://doi.org/10.6028/NIST.IR.8040>
- Hair Jr, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2010). Multivariate data analysis. In *Multivariate data analysis* (pp. 785-785).
- Hamm, J., Money, A., & Atwal, A. (2017). Fall prevention self-assessments via mobile 3D visualization technologies: community dwelling older adults' perceptions of opportunities and challenges. *JMIR Human Factors*, 4(2), e7161. <https://doi.org/10.2196/humanfactors.7161>
- Harangozo, G., & Szigeti, C. (2017). Corporate carbon footprint analysis in practice-With a special focus on validity and reliability issues. *Journal of Cleaner Production*, 167, 1177-1183. <https://doi.org/10.1016/j.jclepro.2017.07.237>
- Hennink, M., & Kaiser, B. N. (2022). Sample sizes for saturation in qualitative research: A systematic review of empirical tests. *Social Science & Medicine*, 292, 114523. <https://doi.org/10.1016/j.socscimed.2021.114523>
- Hertzum, M. (2026). System Usability Scale: A Meta-Analysis of How SUS Relates to Workload, Task Time, and Error Rate. *International Journal of Human-Computer Interaction*, 1-14. <https://doi.org/10.1080/10447318.2026.2625260>
- Hodrien, A., & Fernando, T. (2021). A Review of Post-Study and Post-Task Subjective Questionnaires to Guide Assessment of System Usability. *Journal of Usability Studies*, 16(3).
- Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2), 65-70.
- ISO, I. (2018). *Ergonomics of human-system interaction-Part 11: Usability: Definitions and concepts* (ISO 9241-11: 2018).
- ISO/IEC 25010. (2024). *Systems and Software Engineering: Systems and Software Quality Requirements and Evaluation (SQuaRE): Measurement of System and Software Product Quality*.
- Johnson, S. G., Potrebny, T., Larun, L., Ciliska, D., & Olsen, N. R. (2022). Usability methods and attributes reported in usability studies of mobile apps for health care education: scoping review. *JMIR Medical Education*, 8(2), e38259. <https://doi.org/10.2196/38259>
- Kallio, H., Pietilä, A. M., Johnson, M., & Kangasniemi, M. (2016). Systematic methodological review: developing a framework for a qualitative semi-structured interview guide. *Journal of Advanced Nursing*, 72(12), 2954-2965. <https://doi.org/10.1111/jan.13031>
- Karapanos, E., Zimmerman, J., Forlizzi, J., & Martens, J. B. (2009, April). User experience over time: an initial framework. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 729-738). <https://doi.org/10.1145/1518701.1518814>
- Kennelly, C., Berners-Lee, M., & Hewitt, C. N. (2019). Hybrid life-cycle assessment for robust, best-practice carbon accounting. *Journal of Cleaner Production*, 208, 35-43. <https://doi.org/10.1016/j.jclepro.2018.09.231>
- Klug, B. (2017). An overview of the system usability scale in library website and system usability testing. *Weave: Journal of Library User Experience*, 1(6). <https://doi.org/10.3998/weave.12535642.0001.602>

- Kong, M., Lee, M., Kang, H., & Hong, T. (2021). Development of a framework for evaluating the contents and usability of the building life cycle assessment tool. *Renewable and Sustainable Energy Reviews*, 150, 111475. <https://doi.org/10.1016/j.rser.2021.111475>
- Kortum, P. T., & Bangor, A. (2013). Usability ratings for everyday products measured with the system usability scale. *International Journal of Human-Computer Interaction*, 29(2), 67-76. <https://doi.org/10.1080/10447318.2012.681221>
- Kortum, P., & Peres, S. C. (2014). The relationship between system effectiveness and subjective usability scores using the System Usability Scale. *International Journal of Human-Computer Interaction*, 30(7), 575-584. <https://doi.org/10.1080/10447318.2014.904177>
- Kortum, P., Acemyan, C. Z., & Oswald, F. L. (2021). Is it time to go positive? Assessing the positively worded system usability scale (SUS). *Human Factors*, 63(6), 987-998. <https://doi.org/10.1177/0018720819881556>
- Kujala, S., Miron-Shatz, T., & Jokinen, J. J. (2019). The cross-sequential approach: a short-term method for studying long-term user experience. *Journal of Usability Studies*, 14(2), 105-116.
- Lee, S., & Sah, Y. J. (2020). Development of an approach to measuring learnability based on Ngoms from perspectives of extended learnability. *International Journal of Human-Computer Interaction*, 36(2), 199-209. <https://doi.org/10.1080/10447318.2019.1625569>
- Lewis, J. R. (1995). IBM computer usability satisfaction questionnaires: psychometric evaluation and instructions for use. *International Journal of Human-Computer Interaction*, 7(1), 57-78. <https://doi.org/10.1080/10447319509526110>
- Lewis, J. R. (2014). Usability: lessons learned... and yet to be learned. *International Journal of Human-Computer Interaction*, 30(9), 663-684. <https://doi.org/10.1080/10447318.2014.930311>
- Lewis, J. R. (2018). The System Usability Scale: Past, present, and future. *International Journal of Human-Computer Interaction*, 34(7), 577-590. <https://doi.org/10.1080/10447318.2018.1455307>
- Lewis, J. R., & Sauro, J. (2009, July). The factor structure of the system usability scale. In International conference on human centered design (pp. 94-103). Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-02806-9_12
- Lewis, J. R., & Sauro, J. (2017). Can I Leave This One Out? The Effect of Dropping an Item From the SUS. *Journal of Usability Studies*, 13(1).
- Lewis, J. R., & Sauro, J. (2021). Usability and user experience: Design and evaluation. *Handbook of human factors and ergonomics*, 972-1015. <https://doi.org/10.1002/9781119636113.ch38>
- Lewis, J. R., Utesch, B. S., & Maher, D. E. (2013, April). UMUX-LITE: when there's no time for the SUS. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 2099-2102). <https://doi.org/10.1145/2470654.2481287>
- Lewis, J. R., Utesch, B. S., & Maher, D. E. (2015). Measuring perceived usability: The SUS, UMUX-LITE, and AltUsability. *International Journal of Human-Computer Interaction*, 31(8), 496-505. <https://doi.org/10.1080/10447318.2015.1064654>
- Liang, J., Xian, D., Liu, X., Fu, J., Zhang, X., Tang, B., & Lei, J. (2018). Usability study of mainstream wearable fitness devices: feature analysis and system usability scale evaluation. *JMIR mHealth and uHealth*, 6(11), e11066. <https://doi.org/10.2196/11066>
- Longhurst, R., & Johnston, L. (2023). 10 semi-structured interviews and focus groups. *Key methods in geography*, 168. <https://doi.org/10.4135/9781036233679.n10>
- Lu, Q., Tao, L., Peng, X., & Chen, J. (2024). Signs of front-line healthcare professionals' information anxiety during the COVID-19 pandemic: Grounded theory study in a Wuhan hospital. *Nursing Open*, 11(6), e2207. <https://doi.org/10.1002/nop2.2207>
- Malakahmad, A., Albakri, N. A. B. H., & Shafiq, N. (2015, May). Assessment and Reduction of Carbon Footprint: An Approach via Best Management Practices in a Construction Site. In InCIEC 2014:

- Proceedings of the International Civil and Infrastructure Engineering Conference 2014 (pp. 1019-1030). Singapore: Springer Singapore. https://doi.org/10.1007/978-981-287-290-6_89
- Manouchehri, E., Taghipour, A., Ebadi, A., Homaei Shandiz, F., & Latifnejad Roudsari, R. (2022). How do I deal with breast cancer: a qualitative inquiry into the coping strategies of Iranian women survivors. *BMC Women's Health*, 22(1), 284. <https://doi.org/10.1186/s12905-022-01865-0>
- Maqbool, B., & Herold, S. (2024). Potential effectiveness and efficiency issues in usability evaluation within digital health: A systematic literature review. *Journal of Systems and Software*, 208, 111881. <https://doi.org/10.1016/j.jss.2023.111881>
- Marinelli, E., Radini, S., Foglia, A., Lancioni, N., Piasentin, A., Eusebi, A. L., & Fatone, F. (2021). Validation of an evidence-based methodology to support regional carbon footprint assessment and decarbonisation of wastewater treatment service in Italy. *Water Research*, 207, 117831. <https://doi.org/10.1016/j.watres.2021.117831>
- Marlowe, J., & Clarke, A. (2022). Carbon Accounting: A Systematic Literature Review and Directions for Future Research. *Green Finance*, 4(1), 71. <https://doi.org/10.3934/GF.2022004>
- Mashuri, S., Sarib, M., Rasak, A., Alhabsyi, F., & Syam, H. (2022). Semi-structured interview: A methodological reflection on the development of a qualitative research instrument in educational studies. *Journal of Research and Method in Education*, 12(1), 22-29.
- Maxwell, J. A. (2013). Qualitative research design: An interactive approach: An interactive approach. Sage.
- McGlinn, K., Yuce, B., Wicaksono, H., Howell, S., & Rezgui, Y. (2017). Usability evaluation of a web-based tool for supporting holistic building energy management. *Automation in Construction*, 84, 154-165. <https://doi.org/10.1016/j.autcon.2017.08.033>
- Melville, N. P., & Whisnant, R. (2014). Energy and carbon management systems: organizational implementation and application. *Journal of Industrial Ecology*, 18(6), 920-930. <https://doi.org/10.1111/jiec.12135>
- Mol, M., van Schaik, A., Dozeman, E., Ruwaard, J., Vis, C., Ebert, D. D., ... & Smit, J. H. (2020). Dimensionality of the system usability scale among professionals using internet-based interventions for depression: a confirmatory factor analysis. *BMC Psychiatry*, 20(1), 218. <https://doi.org/10.1186/s12888-020-02627-8>
- Mulrow, J., Machaj, K., Deanes, J., & Derrible, S. (2018). The state of carbon footprint calculators: An evaluation of calculator design and user interaction features. *Sustainable Production and Consumption*, 18, 33. <https://doi.org/10.1016/j.spc.2018.12.001>
- Mustaffa, N. K., Mat Isa, C. M., Ekundayo, D., & Joseph, V. R. A. (2022). Barriers and strategies for improving carbon emissions management approaches in Malaysian construction. *Construction Economics and Building*, 22(3), 99-123. <https://doi.org/10.5130/AJCEB.v22i3.7909>
- Naz, N., Gulab, F., & Aslam, M. (2022). Development of qualitative semi-structured interview guide for case study research. *Competitive Social Science Research Journal*. 3, no. 2 (2022): 42-52.
- Newcomer, K. E., Hatry, H. P., & Wholey, J. S. (Eds.). (2015). *Handbook of practical program evaluation*. <https://doi.org/10.1002/9781119171386>
- Nielsen, J. (1994). Usability engineering. Morgan Kaufmann. <https://doi.org/10.1016/B978-0-08-052029-2.50009-7>
- Nielsen, J. B. T.-U. E. R.-M. (2010). CHAPTER 1 - What Is Usability? (pp. 3–22). Morgan Kaufmann. <https://doi.org/10.1016/B978-0-12-375114-0.00004-9>
- Nielsen, J. (2012). How many test users in a usability study? Nielsen Norman Group. <https://doi.org/10.1042/CS20120633>
- Okon, D. U., & Marauche, O. (2025). The Drivers and Barriers of Corporate Carbon Accounting Adoption: A Systematic Review of Managerial, Organizational, and Policy Factors. *International Journal of Advances in Engineering and Management*, 7(7), 717. <https://doi.org/10.35629/5252-0707717727>

- Onat, N. C., Mandouri, J., Kucukvar, M., Kutty, A. A., & Al-Muftah, A. A. (2025). Driving sustainable business practices with carbon accounting and reporting: a holistic framework and empirical analysis. *Corporate Social Responsibility and Environmental Management*, 32(2), 2795-2814. <https://doi.org/10.1002/csr.3096>
- Orfanou, K., Tselios, N., & Katsanos, C. (2015). Perceived usability evaluation of learning management systems: Empirical evaluation of the System Usability Scale. *The International Review of Research in Open and Distributed Learning*, 16(2). <https://doi.org/10.19173/irrodl.v16i2.1955>
- Oshidero, D. F. O., & Coley, D. A. (2026). Talking carbon: a lexical approach to predictive embodied carbon analysis via machine learning. *Architectural Engineering and Design Management*, 1-34. <https://doi.org/10.1080/17452007.2026.2613773>
- Paskaran, S., Gamage, A., & Chandrasiri, S. (2022, November). A novel ranked emission-factor retrieval for emission calculation. In *2022 International Conference on Computer Engineering, Network, and Intelligent Multimedia (CENIM)* (pp. 1-8). IEEE. <https://doi.org/10.1109/CENIM56801.2022.10037450>
- Qiang, Y., & Pa, N. C. (2024). Sustainable software solutions: a tool integrating life cycle analysis and ISO quality models. *International Journal on Advanced Science, Engineering & Information Technology*, 14(5). <https://doi.org/10.18517/ijaseit.14.5.11268>
- Quaglio, B., Gallina, F., & Giordano, R. (2024, October). An Infrastructure's emissions assessment tool: AMICO-Account Method of Infrastructures embodied CarbOn. In *IOP Conference Series: Earth and Environmental Science* (Vol. 1402, No. 1, p. 012063). IOP Publishing. <https://doi.org/10.1088/1755-1315/1402/1/012063>
- Rabbani, R. M., & Wibowo, A. (2025, December). Implementating Mobile-Based Vehicle Carbon Footprint and Transportation Emission Calculation System to Enhance Users Environmental Awareness. In *8th International Conference on Applied Engineering (ICAE 2025)* (pp. 614-651). Atlantis Press. https://doi.org/10.2991/978-94-6463-982-7_37
- Rodriguez-Jimenez, L., Romero-Martín, M., & Gómez-Salgado, J. (2025). The emergency department carbon footprint calculator: Design and validation. *Medicine*, 104(22), e41652. <https://doi.org/10.1097/MD.00000000000041652>
- Rogers, E. M. (2003). *Diffusion of innovations* (5th ed.). Free Press.
- Rosales, M. V., & Dhakal, S. (2024, October). A Comparative Analysis of Carbon Accounting Methodologies and Tools. In *2024 International Conference on Sustainable Energy: Energy Transition and Net-Zero Climate Future (ICUE)* (pp. 1-7). IEEE. <https://doi.org/10.1109/ICUE63019.2024.10795543>
- Rotaru, O. A., Vert, S., Vasiu, R., & Andone, D. (2020, October). Standardised questionnaires in usability evaluation. applying standardised usability questionnaires in digital products evaluation. In *International Conference on Information and Software Technologies* (pp. 39-48). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-59506-7_4
- Sagar, K., & Saha, A. (2017). A systematic review of software usability studies. *International Journal of Information Technology*, 1-24. <https://doi.org/10.1007/s41870-017-0048-1>
- Sauro, J. (2011). SUSstified? Little-known system usability scale facts. *User Experience Magazine*, 10(3).
- Sauro, J. (2018). 5 Ways to Interpret a SUS Score - *MeasuringU*. *MeasuringU.Com*. <https://measuringu.com/interpret-sus-score/>
- Sauro, J., & Lewis, J. R. (2009, April). Correlations among prototypical usability metrics: Evidence for the construct of usability. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 1609-1618). <https://doi.org/10.1145/1518701.1518947>
- Sauro, J., & Lewis, J. R. (2011, May). When designing usability questionnaires, does it hurt to be positive?. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 2215-2224). <https://doi.org/10.1145/1978942.1979266>

- Sharfina, Z., & Santoso, H. B. (2016, October). An Indonesian adaptation of the system usability scale (SUS). In 2016 *International conference on advanced computer science and information systems (ICACSIS)* (pp. 145-148). IEEE. <https://doi.org/10.1109/ICACSIS.2016.7872776>
- Shea, S. C., & Barney, C. (2015). Teaching clinical interviewing skills using role-playing: conveying empathy to performing a suicide assessment: a primer for individual role-playing and scripted group role-playing. *Psychiatric Clinics*, 38(1), 147-183. <https://doi.org/10.1016/j.psc.2014.10.001>
- Siedlecki, S. L. (2022). Conducting interviews for qualitative research studies. *Clinical Nurse Specialist*, 36(2), 78-80. <https://doi.org/10.1097/NUR.0000000000000653>
- Smulowitz, S. (2017). Interview guide. *The International Encyclopedia of Communication Research Methods*, 1-2. <https://doi.org/10.1002/9781118901731.iecrm0124>
- Spandonidis, C. (2026). Designing for Impact: Evidence-Based Guidelines to Improve Inclusivity, Transparency, and Engagement in Carbon Footprint Tools. *Sustainable Development* 34(3), 3720–3731. <https://doi.org/10.1002/sd.70538>
- Sternberg, Z. (2023). Information Design Considerations for Effective Communication of Sustainability Metrics. *Massachusetts Institute of Technology*.
- Strohmeier, S., Becker, M., & Scheer-Weller, E. (2025). Beyond aversion-principles of appropriate algorithmic decision-making in human resource management. *Expert Systems with Applications*, 128954. <https://doi.org/10.1016/j.eswa.2025.128954>
- Sweller, J. (2011). Cognitive load theory. In *Psychology of learning and motivation* (Vol. 55, pp. 37-76). Academic Press. <https://doi.org/10.1016/B978-0-12-387691-1.00002-8>
- Terry, G., Hayfield, N., Clarke, V., & Braun, V. (2017). Thematic analysis. In C. Willig & W. Stainton Rogers (Eds.), *The Sage handbook of qualitative research in psychology* (2nd ed., pp. 17-37). Sage. <https://doi.org/10.4135/9781526405555.n2>
- Thelwall, M., & Nevill, T. (2021). Is research with qualitative data more prevalent and impactful now? Interviews, case studies, focus groups and ethnographies. *Library & Information Science Research*, 43(2), 101094. <https://doi.org/10.1016/j.lisr.2021.101094>
- Tuomisto, H. L., De Camillis, C., Leip, A., Nisini, L., Pelletier, N., & Haastrop, P. (2015). Development and testing of a European Union-wide farm-level carbon calculator. *Integrated environmental assessment and management*, 11(3), 404-416. <https://doi.org/10.1002/ieam.1629>
- Vasileiou, K., Barnett, J., Thorpe, S., & Young, T. (2018). Characterising and justifying sample size sufficiency in interview-based studies: Systematic analysis of qualitative health research over a 15-year period. *BMC Medical Research Methodology*, 18, 148. <https://doi.org/10.1186/s12874-018-0594-7>
- Vlachogianni, P., & Tselios, N. (2022). Perceived usability evaluation of educational technology using the System Usability Scale (SUS): A systematic review. *Journal of Research on Technology in Education*, 54(3), 392-409. <https://doi.org/10.1080/15391523.2020.1867938>
- Wagner, A. M. (2023). Perceived Usability Evaluation of Hands-On and Virtual Science Laboratories: Using the System Usability Scale (SUS) to Determine Adult Learners' Preferred At-Home Laboratory Experience (Doctoral dissertation, Franklin University).
- Wang, Y., Zhang, Y., & Chen, Y. (2023). A longitudinal study of children's perceived usability: Methodology and measurements. *International Journal of Human-Computer Interaction*, 39(18), 3604-3619. <https://doi.org/10.1080/10447318.2022.2100065>
- Weiner, B. J. (2009). A theory of organizational readiness for change. *Implementation Science*, 4(1), 67. <https://doi.org/10.1186/1748-5908-4-67>
- Wetzlinger, W., Auinger, A., & Dörflinger, M. (2014, June). Comparing effectiveness, efficiency, ease of use, usability and user experience when using tablets and laptops. In *International Conference of Design, User Experience, and Usability* (pp. 402-412). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-07668-3_39

- Whiting, L. S. (2008). Semi-structured interviews: guidance for novice researchers. *Nursing Standard (through 2013)*, 22(23), 35. <https://doi.org/10.7748/ns2008.02.22.23.35.c6420>
- Wutich, A., Beresford, M., & Bernard, H. R. (2024). Sample Sizes for 10 Types of Qualitative Data Analysis: An Integrative Review, Empirical Guidance, and Next Steps. *International Journal of Qualitative Methods*, 23, 16094069241296206. <https://doi.org/10.1177/16094069241296206>
- Xia, Z., Liu, K., Li, J., Gu, Y., Hu, B., Zhong, Y., ... & Lam, K. L. (2025). Evaluating carbon footprint calculators: a comprehensive assessment framework. *Clean Technologies and Environmental Policy*, 1-12. <https://doi.org/10.1007/s10098-025-03169-x>
- Xu, J., & MacAskill, K. (2024). Carbon data and its requirements in infrastructure-related GHG standards. *Environmental Science & Policy*, 162, 103935. <https://doi.org/10.1016/j.envsci.2024.103935>
- Yonata, R., & Purwianti, A. (2017, November). Inveo, a management information system for emissions inventory e-administration in Indonesia. In *2017 2nd International conferences on Information Technology, Information Systems and Electrical Engineering (ICITISEE)* (pp. 78-83). IEEE. <https://doi.org/10.1109/ICITISEE.2017.8285564>
- Zapata, B. C., Fernández-Alemán, J. L., Idri, A., & Toval, A. (2015). Empirical studies on usability of mHealth apps: a systematic literature review. *Journal of medical systems*, 39(2), 1. <https://doi.org/10.1007/s10916-014-0182-2>



This work is licensed under a [Attribution-NonCommercial-ShareAlike 4.0 International](https://creativecommons.org/licenses/by-nc-sa/4.0/) (CC BY-NC-SA 4.0).

APPENDIX A. Interview guide matrix.

Interview Question (Q#)	Section	RQ1 (SUS)	RQ2 (Effectiveness)	RQ2 (Efficiency)	RQ2 (Satisfaction)	RQ2 (Learnability)	RQ3 (Refinements & Adoption)
SUSQ Check	Introduction	x					
Q1	A: Warm-Up		x				
Q2	A: Warm-Up		x				
Q3	A: Warm-Up					x	
Q4	B: First Impressions				x	x	
Q5	B: First Impressions					x	
Q6	B: First Impressions					x	
Q7	C: Data Navigation		x	x			
Q8	C: Data Navigation		x	x			
Q9	C: Methodology Concepts		x			x	
Q10	C: Methodology Concepts		x				
Q11	C: Calculation Logic		x		x		
Q12	C: Calculation Logic		x				
Q13	C: Summarising Results		x	x			
Q14	C: Efficiency			x			
Q15	D: Satisfaction				x		
Q16	D: Satisfaction				x		
Q17	D: Challenges		x				x
Q18	D: Challenges						x
Q19	E: Practicality		x				x
Q20	E: Practicality						x
Q21	F: Refinements						x
Q22	F: Refinements						x
Q23	F: Refinements						x
Q24	G: Final Reflections				x		x
Q25	G: Final Reflections						x

APPENDIX B. Semi-structured interview guide.

Semi-Structured Interview Guide

Evaluation of the CF Methodology and Tool for NOCs:

A Usability Assessment Centred on Effectiveness, Efficiency, Satisfaction, and Learnability

Study Title	<i>Evaluation of the Proposed Carbon Footprint Methodology and Tool for National Olympic Committees (NOCs)</i>
Interviewer	<i>[Your Name / Title]</i>
Participant Role	<i>[e.g., NOC Sustainability Officer, CEO, Facilities Manager]</i>
Tool Version	<i>CF_Tool.xlsx (Standardised Methodology)</i>
Date & Time	<i>[Date] [Start Time]</i>
Interview ID	<i>[Unique participant reference — anonymised]</i>

This interview is structured around four validated usability constructs. Each section of the guide corresponds directly to one of these definitions:

EFFECTIVENESS *The extent to which users can successfully complete tasks accurately, completely, and appropriately, reflecting the quality and correctness of the system's outputs.*
Bevan et al. (2015); Kortum & Peres (2014)

EFFICIENCY *The number of resources required to complete a task, including time, human effort, costs, and materials. Together with effectiveness, efficiency assesses the overall performance of the system.*
ISO 9241-11 (2018); Bevan et al. (2015)

SATISFACTION *Users' subjective experience, reflecting their comfort and acceptability when interacting with the system.*
Corsini et al. (2012)

LEARNABILITY *The time required or the ease with which typical users can learn a system and achieve a certain level of competency.*
Lee & Sah (2020); Abran et al. (2003)

Research Questions underpinning this interview:

RQ1: To what extent does the proposed CF methodology and tool meet accepted usability standards as measured by the System Usability Scale (SUS)?

RQ2: How does the proposed CF methodology and tool perform against effectiveness, efficiency, and satisfaction?

RQ3: What refinements and improvements are required to enhance usability, functionality, and practical adoption?

Introduction (5–7 minutes)*Consent · Context ·
SUS check*

Thank the participant for their time and for engaging with the carbon footprint tool prior to this interview.

Introduce the research:

- This study evaluates a new carbon footprint (CF) methodology and Excel-based tool (CF_Tool.xlsx) developed specifically for National Olympic Committees.
- The methodology is built on sustainability control principles — responsibility and influence rather than financial ownership — and mandates Scope 1, 2, and 3 reporting across both routine operational activities and Olympic Games attendance activities.
- We are not evaluating you. We are evaluating the tool and the methodology.

Explain the three usability constructs (in plain language):

To assess the tool, I will ask about three things drawn directly from usability science:

- **Effectiveness** — whether you could complete tasks correctly and get accurate, complete results from the tool.
- **Efficiency** — how much time, effort, and resource was needed to use the tool — were the demands proportionate to what you got back?
- **Satisfaction** — how you felt using the tool — was it comfortable, acceptable, and something you could see yourself using again?
- **Learnability** — how quickly and easily you were able to learn how to use the tool and reach a working level of competency.

Formalities:

- Confirm permission to record. Recording is for transcription accuracy only and will not be shared externally.
- Assure confidentiality. You will be referred to only by role type (e.g., 'a sustainability officer from a European NOC') — never by name or NOC.
- Participation is voluntary. Responses may be withdrawn at any point without consequence.

SUS Questionnaire — CHECK before proceeding:

If completed: Thank you. This captures your initial, unbiased impression and gives us a baseline measure for RQ1.

If NOT completed: Please take 2–3 minutes to complete it now — before we talk in detail, as our conversation could influence your answers.

Structure: approximately 20 questions across five sections; 45–60 minutes total. There are no right or wrong answers — your honest experience is what matters.

Do you have any questions before we begin?

Section A: Background and Context (5 minutes)*Warm-up · RQ2
baseline*

Purpose: Establish the participant's baseline experience with carbon accounting and their NOC context, which frames how effectiveness, efficiency, and satisfaction are interpreted for their situation.

Q1
RQ2

Can you briefly describe your NOC and its current approach to environmental sustainability and carbon footprinting?

Methodological rationale: *Understanding baseline capability contextualises judgements about effectiveness (can the tool produce results users trust?) and efficiency (is the data burden proportionate to capability?).*

Probing questions:

- a. Does your NOC currently measure its carbon footprint, and if so, which scopes are included?
- b. Is carbon accounting managed in-house or by external consultants?
- c. Which frameworks or standards, if any, does your NOC currently follow?

Q2
RQ2

What is your role, and to what extent are you personally involved in carbon accounting or sustainability reporting?

Methodological rationale: *Role and prior exposure to CF tools shapes both efficiency perceptions (what feels burdensome) and satisfaction (what feels appropriate for someone in this position).*

Probing questions:

- d. Have you used other carbon accounting tools or Excel-based models before?
- e. Would you describe yourself as a beginner, intermediate, or advanced user in carbon footprinting?
- f. How much of your working time is currently dedicated to sustainability activities?

Section B: Effectiveness (15–20 minutes)

*Task accuracy ·
Completeness ·
Output quality |
RQ1, RQ2*

Definition in use: Effectiveness is the extent to which users can successfully complete tasks accurately, completely, and appropriately, reflecting the quality and correctness of the system's outputs.

Bevan et al. (2015); Kortum & Peres (2014)

Interview focus: Did users complete the right tasks? Did the tool produce correct, credible, and sufficiently complete emissions outputs?

Q3
RQ1/2

When you worked through the tool, were you able to complete each emission category from start through to a calculated output — without having to stop or abandon a section?

Methodological rationale: *Task completion is the primary indicator of effectiveness (Bevan et al., 2015). Identifying which sections could and could not be completed reveals where the tool succeeds or fails as a carbon accounting instrument.*

Probing questions:

- g. Were there any tabs or emission sources where you could not produce a result? What caused the blockage?
- h. Did you feel the tool was guiding you towards a correct output, or were you unsure at any point whether you had done it right?

- i. Were there emission sources in the methodology — from either the operational activity list or the Olympic Games list — that the tool did not adequately capture?

Q4
RQ1/2

How accurately does the tool capture your NOC's actual carbon footprint? Did the outputs — the kgCO₂e figures produced — feel credible and correctly derived?

Methodological rationale: Effectiveness requires not just task completion but accuracy and correctness of outputs (Kortum & Peres, 2014). This question probes whether the methodology's calculation logic produces results users can defend.

Probing questions:

- j. Was it clear which emission factor was being applied — for example, DESNZ activity-based factors or DEFRA spend-based factors — and why that factor was appropriate?
- k. Did the automatic calculation of Well-to-Tank (WTT) and Transmission & Distribution (T&D) losses produce results you understood and trusted?
- l. Were there any outputs that surprised you, or that you felt may have over- or underestimated your emissions?

Q5
RQ1/2

The methodology requires all three emission scopes — including the full list of Scope 3 categories — to be assessed. To what extent were you able to address all mandatory emission sources completely?

Methodological rationale: Completeness is an explicit component of effectiveness. The methodology mandates Scope 3 as non-optional, making complete task coverage a direct test of the tool's effectiveness for NOC contexts.

Probing questions:

- m. Were there mandatory emission categories — for example, staff commuting, athlete flights, official merchandise, cloud services, or Olympic Games-specific sources — where you could not produce a result due to data unavailability?
- n. Did the tiered materiality thresholds (Beginner 20%, Intermediate 10–20%, Advanced 5–10%) help you make appropriate decisions about which sources to include, or did they add uncertainty?
- o. Did the tool support you in documenting and justifying any sources that were excluded?

Q6
RQ1/2

The methodology introduces sustainability control principles — based on responsibility and influence rather than financial ownership — to define organisational boundaries. Were you able to correctly and confidently apply the Organisational Boundary Decision Flowchart?

Methodological rationale: Appropriateness is the third dimension of effectiveness: tasks must be completed in the right way. Correct application of the boundary framework is foundational to producing a methodologically appropriate carbon footprint.

Probing questions:

- p. Did the four criteria — legal responsibility, causation responsibility, benefit responsibility, and direct influence — make practical sense for your NOC's operational context?
- q. Were there activities — for example, outsourced event management, athlete training facilities, or shared NOC House operations — where determining boundary inclusion was genuinely ambiguous?
- r. Did the flowchart lead you to decisions you felt confident defending to an external auditor or governing body?

Q7
RQ1/2

After completing data entry, how well did the Summary sheet communicate the quality and completeness of your carbon footprint? Did the breakdown by

scope, method, and activity type support an accurate understanding of your outputs?

Methodological rationale: *Effectiveness extends to the quality of outputs presented to the user. The Summary sheet is the primary interface through which accuracy and completeness become visible — or invisible.*

Probing questions:

- s. Was the separation of emissions by Scope 1, 2, and 3 — and by activity-based versus spend-based methods — meaningful and useful?
- t. Were the uncertainty quality scores (Very Good to Poor) in the Summary helpful for identifying where data collection needs to improve?
- u. Did the visualisations — charts and graphs — accurately represent the emissions profile in a way you could use for reporting or decision-making?

Section C: Efficiency (15–20 minutes)

*Time · Effort ·
Resource demands
| RQ1, RQ2*

Definition in use: Efficiency relates to the number of resources required to complete a task, including time, human effort, costs, and materials. Together with effectiveness, efficiency assesses the overall performance of the system.

ISO 9241-11 (2018); Bevan et al. (2015)

Interview focus: Were the time, effort, and resource demands of using the tool proportionate to the outputs received?

Q8
RQ1/2

Overall, how would you describe the time and effort required to work through the tool from start to a completed Summary? Was that investment proportionate to the value of the outputs produced?

Methodological rationale: *ISO 9241-11 (2018) positions efficiency as the ratio of resources expended to outputs achieved. This opening question establishes the overall cost-benefit perception before drilling into specific resource types.*

Probing questions:

- v. Approximately how long did it take you to work through the tool, including reading guidance, entering data, and reviewing results?
- w. Did that time feel justified given what you got back — or was the effort disproportionate to the output?
- x. Would the time burden be acceptable as an annual process within your NOC's current workload?

Q9
RQ1/2

How efficient was the process of navigating between sheets, locating the correct input fields, and entering data? Did the tool's structure minimise or add to the effort required?

Methodological rationale: *Human effort — including cognitive load and navigation friction — is a core efficiency resource. The tool's colour-coding, tab structure, and dropdown menus are all designed to reduce effort; this question tests whether they do so in practice.*

Probing questions:

- y. Were the colour-coded cells — blue for manual entry, grey for auto-calculated, green for dropdowns — immediately clear, or did they require learning time?
- z. Did dropdown menus and pre-populated emission factors reduce your input burden, or did you find yourself overriding or questioning the defaults?

- aa. Were there any repetitive or duplicated steps that could have been automated or streamlined?

Q10
RQ1/2

How much effort was required to collect the data the tool needs — before you even opened the spreadsheet? Which data inputs were readily to hand, and which required significant additional work to obtain?

Methodological rationale: Efficiency includes the upstream resource cost of data collection, not just within-tool navigation. For NOCs, data collection burden is a primary barrier to adoption; this question surfaces where the real-world effort lies.

Probing questions:

- bb. Which emission sources — for example, utility bills for Heating, travel logs for Business Travel, consignment notes for Freight, or staff surveys for Commuting — had data immediately available?
- cc. For which sources did you need to contact other departments, external suppliers, or make estimates? How long did that take?
- dd. Were there sources where you had to use the spend-based fallback method because activity data was simply not accessible? Did that feel like an acceptable workaround?

Q11
RQ1/2

The tool offers both activity-based and spend-based calculation methods. How efficiently could you switch between these methods when needed — and did the dual-method approach reduce or increase the overall effort of completing the tool?

Methodological rationale: Providing two calculation pathways is a deliberate design choice to reduce the efficiency cost of data gaps. This question tests whether the dual-method architecture achieves that goal or introduces its own friction.

Probing questions:

- ee. Was it immediately clear, within each tab, when the spend-based method was available and how to activate it?
- ff. Did having the option to use spend-based data for some sources and activity-based for others make the process feel more or less manageable?
- gg. Were there sources where neither method felt proportionate — either too data-intensive or too imprecise to be worth the effort?

Q12
RQ1/2

The Index & Data Guide provides detailed guidance for every emission source. How efficiently did this resource support your use of the tool — did it reduce the effort of making decisions, or did consulting it add to your workload?

Methodological rationale: Guidance materials are a resource with their own efficiency cost. If the Index requires extensive consultation, it may reduce overall system efficiency even while supporting effectiveness. The ideal is guidance that is accessed quickly and purposefully.

Probing questions:

- hh. Did you need to consult the Index & Data Guide often, occasionally, or rarely?
- ii. When you did consult it, did it answer your question quickly and clearly — or did it require significant reading before you could proceed?
- jj. What format of guidance — for example, in-cell tooltips, worked examples, short videos — would have reduced the effort of using the tool?

Section D: Satisfaction (10–15 minutes)

Comfort ·
Acceptability ·
Subjective

experience | RQ1,
RQ2

Definition in use: Satisfaction concerns users' subjective experience, reflecting their comfort and acceptability when interacting with the system.

Corsini et al. (2012)

Interview focus: How did using the tool feel? Was the experience comfortable, acceptable, and appropriate for users in this role?

Q13
RQ1/2

Setting aside what the tool can do, how did it feel to use? Would you describe the experience as comfortable, manageable, stressful, or something else?

Methodological rationale: *Satisfaction is a subjective construct — how the experience felt, independently of task outcomes (Corsini et al., 2012). This opening question invites an unguided affective response before more targeted probes.*

Probing questions:

- kk. Were there moments where you felt confident and in control — or moments of frustration, confusion, or anxiety about whether you were doing things correctly?
- ll. Did the tool feel like it was designed for someone with your level of experience, or did it feel pitched too high or too low?
- mm. If a colleague asked you what it was like to use, what would you say in one or two sentences?

Q14
RQ1/2

How acceptable does this tool feel as an ongoing part of your NOC's annual sustainability workflow — something you could see being used year after year?

Methodological rationale: *Acceptability is the forward-looking dimension of satisfaction: not just how the tool felt today, but whether users would willingly integrate it into their practice (Corsini et al., 2012). This is directly linked to adoption potential.*

Probing questions:

- nn. Is there anything about the experience that would put you off using the tool again in future?
- oo. Would you feel comfortable delegating use of the tool to a colleague with less sustainability experience — or does it require someone with a specific background?
- pp. On a scale of 0 to 10, how likely would you be to recommend this tool to another NOC, and what is the primary reason for that score?

Q15
RQ1/2

How comfortable were you with the level of methodological complexity embedded in the tool — for example, the sustainability control boundary principles, the materiality framework, and the mandatory Scope 3 requirements?

Methodological rationale: *Comfort with complexity is a satisfaction-specific concern. A methodologically rigorous tool may produce discomfort even when it is effective — and this discomfort is a legitimate usability finding that affects acceptability.*

Probing questions:

- qq. Did the sophistication of the methodology feel appropriate, or did it create discomfort — a sense of uncertainty about whether you were applying it correctly?
- rr. Were there concepts — for example, double materiality, dynamic materiality, or sustainability control — that felt unfamiliar or unsettling in a practical context?
- ss. Would simplified language, a glossary, or plain-English summaries of the key concepts improve your comfort with the tool?

Q16
RQ1/2

How satisfied are you with the outputs the tool produces — both their content and their presentation? Do the results feel like something you could present to your board, a funder, or the IOC with confidence?

Methodological rationale: *Satisfaction with outputs bridges effectiveness (accuracy) and satisfaction (acceptability). Users may produce correct results but remain dissatisfied if outputs feel insufficient for their reporting context.*

Probing questions:

- tt. Does the Summary sheet provide results in a format suitable for internal reporting — for example, to a board or sustainability committee?
- uu. Would you feel confident sharing the results with an external audience — for example, as part of an IOC sustainability report, a public disclosure, or a third-party audit?
- vv. Is there any aspect of the outputs — format, level of detail, visualisation style — that does not match what your NOC would actually need to report?

Q17
RQ1/2

What would you say was the single aspect of using this tool that generated the most positive experience — and the single aspect that generated the most discomfort or dissatisfaction?

Methodological rationale: *This contrastive question surfaces the affective peaks and troughs of the experience (Corsini et al., 2012), which often predict adoption behaviour more reliably than average satisfaction ratings.*

Probing questions:

- ww. Was the best aspect related to a specific feature, a point of clarity, a result that felt meaningful, or something else?
- xx. Was the most uncomfortable moment related to confusion, a sense of inadequacy, concern about accuracy, or something structural about the tool?
- yy. If you could change only one thing to improve how the tool feels to use — not what it does, but how it feels — what would that be?

Section E: Learnability (10–15 minutes)

*Learning curve ·
Competency ·
Support needs |
RQ1, RQ2*

Definition in use: Learnability is the time required or the ease with which typical users can learn a system and achieve a certain level of competency.

Lee & Sah (2020); Abran et al. (2003)

Interview focus: How quickly and easily could users orient themselves, understand the tool's logic, and reach independent working competency — with or without support?

Q18
RQ1/2

When you first opened the tool, how quickly were you able to understand its structure and feel ready to begin entering data? Can you describe the learning curve you experienced?

Methodological rationale: *Learnability begins at first contact with a system (Lee & Sah, 2020). The speed of initial orientation — before any guidance is consulted — reveals how intuitively the tool communicates its own logic and purpose.*

Probing questions:

- zz. Did the Introduction sheet and Index & Data Guide give you enough context to start independently, or did you feel uncertain about where to begin?

- aaa. How many tabs or sections did you need to explore before the overall structure of the tool made sense to you?
- bbb. At what point — if at all — did you feel you had reached a working level of competency with the tool?

Q19
RQ1/2

How much prior knowledge — of carbon accounting, Excel, or the GHG Protocol — did you need to draw on in order to use this tool competently? Is this tool accessible to a user with limited background in any of these areas?

Methodological rationale: *Learnability is assessed relative to 'typical users' (Abran et al., 2003). For NOCs, the typical user may have limited specialist knowledge; understanding the prerequisite knowledge burden reveals whether the tool is genuinely accessible or implicitly gatekept.*

Probing questions:

- ccc. Were there concepts — such as Scope 1/2/3 categorisation, well-to-tank losses, emission factors, or materiality thresholds — that you needed to look up or already understand before the tool made sense?
- ddd. Could a colleague with no prior carbon footprinting experience realistically learn to use this tool independently within a reasonable timeframe — and if so, what timeframe do you estimate?
- eee. Did any aspect of the tool assume a level of technical knowledge that a non-specialist sustainability officer would be unlikely to have?

Q20
RQ1/2

How useful were the in-tool learning aids — the Introduction sheet, the Index & Data Guide, the colour-coding system, and the dropdown menus — in helping you learn how to use the tool? Did they reduce or extend the time needed to reach competency?

Methodological rationale: *In-tool guidance is a primary mechanism for accelerating learnability (Lee & Sah, 2020). This question tests whether the tool's embedded learning support is sufficient, well-placed, and practically used — or whether users bypass it and learn by trial and error.*

Probing questions:

- fff. Did you use the Index & Data Guide as a learning resource — reading it before entering data — or as a reference you consulted reactively when stuck?
- ggg. Were the colour-coded cells and dropdown menus intuitive enough to reduce the amount of reading required to get started?
- hhh. Were there sections where in-tool guidance was insufficient and you had to consult external resources, ask for help, or make assumptions?

Q21
RQ1/2

Thinking about the second time you would use this tool — for next year's carbon footprint — how much re-learning do you think would be required? Does the tool support the retention of competency over time?

Methodological rationale: *Sustained learnability — the ability to return to a system after a gap and regain competency quickly — is a distinct dimension of learnability (Abran et al., 2003). For an annual-use tool, relearning cost is a critical adoption consideration.*

Probing questions:

- iii. Would you need to re-read the guidance materials before your second use, or does the tool's design make it easy to pick up again after a period of absence?
- jjj. Are there aspects of the tool's logic — for example, the boundary framework, the materiality thresholds, or the OA/OG separation — that you would need to relearn each time?
- kkk. What features — for example, annotated example data, a 'start here' checklist, or a brief summary of decisions made in the previous year — would most help a returning user regain competency quickly?

Section F: Refinements and Improvements (5–10 minutes)

RQ3 · Usability ·
Functionality ·
Adoption

Purpose: Identify specific, actionable changes to the methodology and tool that would improve effectiveness, reduce efficiency costs, enhance satisfaction, and accelerate learnability. Questions are anchored explicitly to each construct.

Q22
RQ3

Thinking specifically about effectiveness — the accuracy, completeness, and appropriateness of outputs — what changes to the methodology or tool would most improve the quality of results NOCs can produce?

Methodological rationale: Improvements to effectiveness may involve calculation logic, emission factor accuracy, boundary guidance, or scope coverage. Grounding suggestions in the effectiveness construct focuses responses on output quality rather than interface preferences.

Probing questions:

- lll. Are there emission sources in the methodology — operational or Olympic Games — where the calculation formula, the emission factor defaults, or the data input structure is insufficient for producing accurate results?
- mmm. Would additional validation checks — for example, warnings when inputs look implausible, or automatic flags for missing mandatory sources — improve output quality?
- nnn. Does the methodology need additional sector-specific guidance for NOC-specific activities not currently well-covered, such as digital marketing, national training centres, or athlete wellbeing programmes?

Q23
RQ3

Thinking specifically about efficiency — the time, effort, and resources required — what changes to the tool's design, structure, or supporting materials would most reduce the burden of using it?

Methodological rationale: Efficiency improvements are the most common target of usability refinement but must be framed in terms of the ISO definition: reducing resource expenditure (time, effort, cost, materials) without reducing output quality.

Probing questions:

- ooo. Are there steps in the data entry process that could be automated — for example, pre-population of standard distances, integration with travel management systems, or import of utility data directly from bills?
- ppp. Would a simplified 'quick-start' version of the tool — covering only the highest-priority emission sources — reduce the initial effort burden enough to encourage first-time adoption?
- qqq. What supporting resources — training materials, onboarding guides, or worked-example datasets — would reduce the effort of preparing to use the tool?

Q24
RQ3

Thinking specifically about satisfaction — comfort and acceptability — what changes would make the experience of using this tool more positive, more confidence-inspiring, or more appropriate for users in your role?

Methodological rationale: Satisfaction improvements often target interface design, language, feedback mechanisms, and the emotional experience of use. Anchoring to the Corsini et al. (2012) definition keeps responses focused on subjective experience rather than technical functionality.

Probing questions:

- rrr. Are there aspects of the interface — layout, colour-coding, terminology, or structural logic — that create discomfort or reduce confidence?
- sss. Would more frequent in-tool feedback — for example, confirmation messages, progress indicators, or contextual help text — make the experience feel more reassuring?

ttt. What would need to change for a user in your position to move from 'I could use this' to 'I want to use this'?

Q25
RQ3

Thinking specifically about learnability — the ease and speed with which a new user can reach competency — what changes to the tool's design or supporting resources would most reduce the learning burden?

Methodological rationale: *Learnability improvements directly affect first-time adoption rates (Lee & Sah, 2020; Abran et al., 2003). For a tool aimed at NOCs with widely varying sustainability expertise, reducing the learning cost is a prerequisite for broad uptake.*

Probing questions:

- uuu. Would a structured onboarding sequence — for example, a guided walkthrough, an annotated first-use checklist, or a short introductory video — significantly reduce the time needed to reach independent competency?
- vvv. Are there aspects of the tool's terminology, conceptual framework, or structural logic that should be simplified or more explicitly explained to lower the learning threshold for first-time users?
- www. Would a 'beginner mode' — limiting visible tabs and options to only the most essential emission sources — make the initial learning experience more manageable without compromising the methodology's integrity?

Close (3–5 minutes)

That concludes the main questions. Before we finish, is there anything important about the tool or methodology that we have not covered — any observation, concern, or suggestion you would like to add?

Permission for follow-up:

- Would you be willing to be contacted again in the future — for clarification of today's responses, to review a revised version of the tool, or to participate in a follow-up validation study?

Thank you sincerely for your time and for the quality of your responses. Your input is essential for developing a carbon accounting tool that genuinely works for NOCs at every stage of their sustainability journey.

I will now stop the recording.

Appendix: Question Alignment Matrix

Each question is mapped to its primary usability construct, research question, and SUS item(s) to support systematic analysis.

Q#	Core focus	Construct	RQ	ISO 9241-11	SUS items
1	NOC baseline & CF maturity	Context	RQ2	—	—
2	Participant role & prior CF experience	Context	RQ2	—	—
3	Task completion across emission tabs	Effectiveness	RQ1/2	Completeness	2, 4, 6
4	Accuracy & credibility of outputs	Effectiveness	RQ1/2	Accuracy	1, 2
5	Completeness of mandatory Scope 3	Effectiveness	RQ1/2	Completeness	2, 4
6	Boundary flowchart — appropriate use	Effectiveness	RQ1/2	Appropriateness	2, 6
7	Summary sheet — output quality	Effectiveness	RQ1/2	Accuracy	1, 9
8	Overall time & effort investment	Efficiency	RQ1/2	Time/resources	3, 5
9	Navigation & data entry effort	Efficiency	RQ1/2	Human effort	3, 7, 8
10	Upstream data collection burden	Efficiency	RQ1/2	Cost/effort	3, 5
11	Dual-method switching efficiency	Efficiency	RQ1/2	Human effort	3, 5
12	Index & Data Guide — effort to consult	Efficiency	RQ1/2	Human effort	4, 10
13	Affective experience — comfort in use	Satisfaction	RQ1/2	Comfort	1, 3, 9
14	Acceptability & adoption willingness	Satisfaction	RQ1/2	Acceptability	1, 9
15	Comfort with methodological complexity	Satisfaction	RQ1/2	Comfort	4, 6, 10
16	Satisfaction with outputs for reporting	Satisfaction	RQ1/2	Acceptability	1, 9
17	Peak/trough satisfaction moments	Satisfaction	RQ1/2	Comfort	1–10
18	Initial orientation & learning curve	Learnability	RQ1/2	Ease of learning	7, 10
19	Prerequisite knowledge burden	Learnability	RQ1/2	Ease of learning	4, 7
20	Effectiveness of in-tool learning aids	Learnability	RQ1/2	Ease of learning	4, 7, 10
21	Retention of competency over time	Learnability	RQ1/2	Competency	7, 10
22	Improvements targeting effectiveness	Effectiveness	RQ3	All	—
23	Improvements targeting efficiency	Efficiency	RQ3	All	—
24	Improvements targeting satisfaction	Satisfaction	RQ3	All	—
25	Improvements targeting learnability	Learnability	RQ3	All	—

SUS items: (1) use frequently; (2) unnecessarily complex; (3) easy to use; (4) need support; (5) well-integrated; (6) inconsistency; (7) quick to learn; (8) cumbersome; (9) confident; (10) needed to learn a lot. Shading: green = Effectiveness, amber = Efficiency, purple = Satisfaction, crimson = Learnability.

APPENDIX C. Deductive coding framework for the semi-structured interview data.

This table shows representative raw quotes alongside their assigned codes, sub-themes, and main themes, illustrating the transformation from raw data to theme across the deductive thematic analysis reported in the Semi-structured interviews section.

Theme	Sub-theme	Participant	Raw quote	Code
Effectiveness	Task completion	P7	<i>“unable to complete staff commuting because I don’t have that survey data”</i>	Data availability barrier
Effectiveness	Task completion	P5	<i>“completed all tabs” and described the OA vs OG separation as a “major effectiveness win”</i>	Boundary structure supports completion
Effectiveness	Accuracy and credibility of outputs	P13	<i>“the kgCO_{2e} figures felt plausible compared to our previous consultant’s report”</i>	Output plausibility benchmarking
Effectiveness	Accuracy and credibility of outputs	P12	<i>“I felt I got an output but was it the correct one?”</i>	Output uncertainty (novice)
Effectiveness	Accuracy and credibility of outputs	P3	<i>“the outputs felt broadly credible, though I still checked a few emission factors against our own records”</i>	Independent verification behaviour
Effectiveness	Application of methodological components	P10	<i>materiality thresholds “helped me justify excluding some smaller sources”</i>	Materiality threshold utility
Effectiveness	Application of methodological components	P12	<i>‘Legal responsibility?’ and ‘Direct influence?’ were “too abstract”</i>	Terminology abstraction barrier
Efficiency	Time and effort required	P7	<i>time “was spent trying to understand what the sheet wanted”</i>	Interface comprehension effort
Efficiency	Time and effort required	P1	<i>“collecting data from different colleagues took longer than the spreadsheet itself”</i>	Data-collection burden (external)
Efficiency	Time and effort required	P9	<i>“I have all our data in one place already”</i>	Prior data readiness (experienced)
Efficiency	Navigation and input design	P5	<i>“the dropdowns and pre-set factors reduced a lot of manual effort”</i>	Pre-filled inputs reduce burden
Efficiency	Navigation and input design	P6	<i>“some tabs required more clicking around than expected”</i>	Navigation friction
Efficiency	Navigation and input design	P13	<i>“having to flip to a separate (SB) sheet broke my workflow”</i>	Workflow interruption across tabs
Satisfaction	Satisfaction with methodology and tool	P2	<i>the process felt “a bit stressful at times”</i>	Process-related frustration (novice)
Satisfaction	Satisfaction with methodology and tool	P14	<i>“more robust than the ad hoc methods many organisations use”</i>	Perceived methodological credibility
Satisfaction	Satisfaction with outputs	P13	<i>“the Summary sheet is board ready. The pie chart by scope is perfect”</i>	Output presentation readiness

Theme	Sub-theme	Participant	Raw quote	Code
Satisfaction	Satisfaction with outputs	P8	<i>"I wouldn't present these results to anyone. The problem is the quality of the underlying data, not the presentation"</i>	Data-quality attribution, not tool design
Learnability	Complexity level	P10	<i>"designed by experts for experts"</i>	Perceived expert-oriented design
Learnability	Complexity level	P7	<i>"actually a bit too advanced, especially for a beginner like me"</i>	Beginner-complexity mismatch
Learnability	Learning curve	P9	<i>"I was entering data within five minutes"</i>	Rapid orientation (experienced)
Learnability	Learning curve	P12	<i>"it took me a while of clicking through tabs to see the pattern"</i>	Exploratory orientation delay (novice)
Learnability	Learning curve	P3	<i>"the dropdowns and the colour-coded cells are critical for learnability"</i>	In-tool learning aids valued
Learnability	Prior knowledge requirements	P8	<i>the tool "assumes a certain level of familiarity with carbon accounting concepts"</i>	Assumed domain knowledge
Learnability	Prior knowledge requirements	P2	<i>"you need to know what Scope 1, 2, and 3 are, or what an emission factor is"</i>	Terminology as prerequisite
Refinements and improvements	Validation and feedback	P7	<i>the Summary sheet "should flag it in red" if a cell is left blank</i>	Automated error-flagging suggestion
Refinements and improvements	Validation and feedback	P12	<i>recommended a "green checkmark" after finishing a tab</i>	Completion-feedback suggestion
Refinements and improvements	Onboarding support	P2	<i>proposed a "Demo Data" function to populate the tool with example data</i>	Guided onboarding suggestion
Refinements and improvements	Onboarding support	P5	<i>recommended "a short introductory training video" demonstrating the overall workflow</i>	Video-based onboarding suggestion



This work is licensed under a [Attribution-NonCommercial-ShareAlike 4.0 International](https://creativecommons.org/licenses/by-nc-sa/4.0/) (CC BY-NC-SA 4.0).