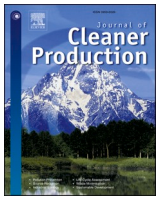
Journal of Cleaner Production 529 (2025) 146762

Contents lists available at ScienceDirect

Journal of Cleaner Production

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



Towards valid and reliable measurement of sustainability knowledge

Talia Stough [a,d,*], Alexander Brewer [b], Aurélien Decamps [b,c], Scott Blair [b], Wim Lambrechts [a,e], Estela Castelli Florino Pilz [b], Marjolein C.J. Caniëls [a], Jean-Christophe Carteron [b]

[a] Open Universiteit, Faculty of Management, Heerlen, Netherlands
[b] Sulitest Impact, Marseille, France
[c] Kedge Business School, Marseille, France
[d] KU Leuven, Faculty of Economics and Business, Leuven, Belgium
[e] Hasselt University, Faculty of Business Economics, Hasselt, Belgium

HIGHLIGHTS

- Introduces a new tool (TASK) to assess sustainability knowledge in higher education.
- Multidimensional Item Response Theory in TASK ensures strong psychometric properties.
- Sustainability knowledge assessment informs curriculum (re)design.

ARTICLE INFO

Keywords:
Education for sustainability
Sustainability knowledge assessment
Epistemological assessments
Multidimensional item response theory

ABSTRACT

As sustainability is increasingly integrated into higher education, being able to assess the level of learners' sustainability-related knowledge is critical to understand where potential gaps are and how curricula can be (re) designed to foster higher levels of attainment. Research on measuring knowledge of sustainability is sparse due to the contested nature of the construct and the lack of valid and reliable measurement tools. In this research, we aim to address these barriers. We consider how different conceptualizations of sustainability could lead to different manifestations of the latent construct and thematic structure of measurement tools. We introduce The Assessment of Sustainability Knowledge (TASK) which employs an "embedded" conceptualization of sustainability to measure the knowledge of the interrelatedness of ecological and social systems (of which economic systems are embedded). Regarding the reliability of sustainability knowledge measurement tools, we posit that the assumption of unidimensionality should be rejected, given the interrelatedness of sustainability as a concept. We describe the use of Multidimensional Item Response Theory employed in TASK and demonstrate the strong psychometric properties such an approach offers. We contribute novel insights regarding sustainability knowledge assessments garnished through developing and piloting TASK to further theoretical and practical discussions of sustainability knowledge assessments.

1. Introduction

As the global discourse of "sustainability" becomes operationalized in goals and strategic frameworks, sustainability has become one of the most prevalent themes in the context of higher education as found in the thematic review of higher education by Daenekindt and Huisman (2020). In response to the United Nations Sustainable Development Goal (SDG) 4 (Education), Target 7, stating that by 2030, all learners should acquire the "knowledge and skills needed to promote sustainable development" (UN, 2016), educators the world over are trying to respond to this call by integrating sustainability into higher education (Leal Filho et al., 2023). As sustainability becomes increasingly thematically integrated, higher education institutions (HEIs) are taking formal steps to assess the progress of these efforts (Berzosa et al., 2017; Gutiérrez-Mijares et al., 2023), including assessing attainment of Education for Sustainability (EfS)-related learning outcomes (Gutierrez-Bucheli et al., 2022).

Assessment of learning outcome attainment is useful to indicate the effectiveness of efforts to integrate themes into curriculum and inform curricular (re)design (Mendoza et al., 2022). In addition to this, since

* Corresponding author. Open Universiteit, Valkenburgerweg 177, 6419 AT, Heerlen, Netherlands.
E-mail address: talia.stough@ou.nl (T. Stough).

https://doi.org/10.1016/j.jclepro.2025.146762
Received 2 February 2024; Received in revised form 21 August 2025; Accepted 28 September 2025
Available online 10 October 2025

EfS learning outcomes often include behavioral elements (for example, SDG 4.7 includes the behavior of "promote sustainable development"), assessing the levels of attainment of learning outcomes that are presumed to positively influence behavior (e.g., knowledge, skills, attitudes) could better inform the predictive validity of such learning outcomes as behavioral determinants in the context of EfS (Stough et al., 2025). However, progress in this field has been slow, which we posit is partly due to a lack of appropriate tools with strong psychometric properties suitable for such diagnostic purposes (Kuehl et al., 2021). Adding to this, defining sustainability as the latent construct in assessment activities needs special consideration given the contested nature of the concept (Stough et al., 2018).

In this research, we aim to address these barriers with the hope that doing so will foster more empirical research on sustainability knowledge attainment, which could ultimately better inform the effectiveness and (re)design of sustainability-related education. By developing a robust measurement tool we contribute to theoretical discussions of, and practical approaches to, sustainability knowledge assessment. First, by delineating sustainability as a contested concept and exploring the possible manifestations of such when defining the latent construct in a measurement tool, we respond to the concern of Stough et al. (2018) about validity issues that can arise when assessing sustainability given the contested nature of the concept. In line with the work of Raworth (2017), we propose the use of an embedded conceptualization of sustainability as a latent construct in the new The Assessment of Sustainability Knowledge (TASK) tool. Via data garnished through piloting TASK, we offer novel insights on sustainability knowledge assessments from a user perspective. Responding to the call of Yavuz Temel et al. (2022) that epistemological measurement tools should demonstrate strong psychometric properties, as well as the concern of Kuehl et al. (2021) about the psychometric properties of current tools in the EfS landscape, we employ Multidimensional Item Response Theory in the scoring model (Hartig and Höhler, 2009). Further, we reflect on the assumption of unidimensionality for the context of sustainability knowledge. Throughout, we consider how measurement of knowledge attainment can help inform the field of EfS both practically and theoretically.

In Section 2, we offer the reader background context. Given the contested nature of the concept (Reid and Petocz, 2006; Connelly, 2007; Kurucz et al., 2014), we explore how "sustainability" could manifest when defining the latent construct in a measurement tool. We offer a background on measuring sustainability knowledge and field explore shortcomings with existing tools and approaches of this in the context of EfS. We also consider knowledge as one learning outcome among many in the field of EfS. In Section 3, we introduce the TASK tool, explain the motivation for its thematic hierarchy, discuss insights garnished while piloting the tool, describe how items are developed, present the scoring model for the tool, and explore its psychometric properties. In Section 4, we offer a discussion about the implications of our research on sustainability knowledge assessments in higher education. In Section 5 we offer concluding remarks.

2. Background

SDG 4.7 (stated above) encompasses some major tensions that the field of EfS must constantly grapple with. The first tension is that "knowledge and skills" are in function of "promoting" sustainable development—hence there is an assumption that the development of certain knowledge and skills (along with other desired learning outcomes discussed in the EfS literature—e.g., attitudes) will influence the future behavior of graduates (i.e., the role of these learning outcomes as behavioral determinants). The second tension deals with the contested nature as to which exact "knowledge and skills" are required (i.e., which behavioral determinants are most effective to develop in function of which exact behaviors EfS would like to influence). The last tension deals with contested nature of sustainability itself—i.e., how the end goal of "sustainable development" is defined. In this section, we offer the reader a short background to these defining tensions.

2.1. Sustainability-related learning outcomes

In the literature, different types of desired EfS-related outcomes are described, referring to ways to integrate and assess individual sustainability competencies or capabilities (see Sandri et al., 2018; Redman et al., 2021 for review). Proposed conceptualisations often refer to competencies as the overarching concept that encompasses knowledge, skills, attitudes, and behavior related to sustainability in rather general terms, influencing the integration of such outcomes in curricula (cf. Lambrechts et al., 2013). Proposed assessment methods thereby often refer to student self-assessment of competencies, e.g. through reflective writing, without specifying the type of knowledge, skills or attitudes targeted. This poses challenges regarding the validity of such approaches. Building on the work of Besong and Holland (2015), Stough et al. (2024a) proposed distilling EfS-related learning outcomes into four *foci*: 1) disposition focused; 2) knowledge focused 3) skills focused; and; 4) behavior focused. Within these foci are discrete constructs that could be assessed with specific measurement instruments, for example: attitudes towards sustainability, knowledge about sustainability, sustainability-related skills (e.g., cognitive skills like systems thinking; as well as intrapersonal skills like working in a diverse team), and sustainability-related behaviors. Some constructs might be directly observable (e.g., certain types of behavior), while others would require the use of a psychometric tool to measure, as they are not directly observable (e.g., knowledge, attitude).

There is an assumption that by developing these learning outcomes, certain future behaviors of graduates (e.g., "promote sustainable development") will be influenced. Hence, the desired learning outcomes articulated in the EfS literature can be thought of as behavioral determinants. To shortly illustrate the theoretical role of these EfS-related foci as behavioral determinants, we refer to the Integrated Behavioral Model of Montaño and Kasprzyk (2015), wherein behavior is influenced by: 1) knowledge and skills, 2) salience of the behavior; 3) intention (influenced by attitudes, norms, and agency), 4) environmental constraints, and 5) habits. Research exploring the effect of different sustainability-related learning outcomes as behavioral determinants is in a nascent phase (Redman et al., 2021). For an expanded discussion on how behavioral change theories can better inform desired sustainability-related learning outcomes see Stough et al. (2025).

2.2. Ensuring reliability of sustainability knowledge assessments

Common methods for estimating overall and subscores within psychometric modeling frameworks include: Classic Test Theory (CTT), Item Response Theory (IRT), Cognitive Diagnostic Models (CDM), and factor analysis (FA) (Sijtsma and van der Ark, 2020). If the goal of a measurement instrument is to diagnose a test takers' knowledge of sustainability (or gaps therein), the overall and subscores should have high psychometric properties (Yavuz Temel et al., 2022). As opposed to a raw score (i.e., percentage of correct answers), transformed scores from IRT methods give items different weights depending on the discriminating power of the item. IRT is increasingly utilized in the context of epistemological assessments (e.g., Shaw et al., 2020). Unidimensional IRT models, such as the Rasch model, take into account a person's ability, an item's score, and the difficulty level of the item. Internal consistency is concerned with the homogeneity of the items within a measurement tool (Devellis, 2017). Unidimensional IRT assumes "unidimensionality" of the phenomenon that is the latent construct. For example, assuming the three domains of ecological-, social-, and economic-sustainability knowledge represent three separate latent constructs, and that knowledge thereof is not interrelated. Multidimensional IRT does not assume that subcontructs are separate (Hartig and Hohler, 2009). The benefits of MIRT in knowledge

assessment have been increasingly recognized (Hartig and Höhler, 2009). For example, Yavuz Temel et al. (2022) demonstrated that while unidimensional IRT models could yield good results, multidimensional IRT models provided better fit for epistemological tests.

2.3. Measuring knowledge of sustainability

While various tools exist to assess sustainability in higher education institutions in general (see Berzosa et al., 2017 for review), there are scare tools to designed to assess sustainability knowledge specifically. In their guidance for Sustainability Tracking and Rating System (STARS), the Association for the Advancement of Sustainability in Higher Education (AASHE) offers a list of tools that reporting organizations can use to assess Academic Credit 6: the "sustainability literacy" of learners (AASHE, 2025). While AASHE's standards state that "sustainability literacy" assessment should focus on knowledge of sustainability topics and challenges (AASHE, 2024), the tools they suggested to measure literacy span various constructs. Some tools included in the list intend to measure attitudes, such as the Sustainability Attitude Scale (SAS) developed by the North American Environmental Education (evaluation-archive.naaee.org/tools/sustainability-attitudes-scale), which other tools intend to measure consciousness, such as the Environmental Sustainability Consciousness Questionnaire (ESCQ) (Gericke et al., 2019). Among these tools are two focused on knowledge attainment (i.e., conceptual awareness): ASK and Sulitest. In 2014, the Assessment of Sustainability Knowledge (ASK) tool was developed (Zwickle et al., 2014) and a revised version was launched in 2018 (Zwickle and Jones, 2018). The ASK tool organizes the latent construct into three subconstructs (social, environmental, and economic-sustainability), is composed of 12 questions, and applies Item Response Theory (IRT) for scoring (Zwickle et al., 2014; Zwickle and Jones, 2018). In 2016 the Sulitest launched the Sulitest Core Module, which is composed of 30 questions and is scored using the raw score (Décamps et al., 2017).

Both tools have also been used in research examining the level of sustainability-related knowledge students have. For example, The ASK tool has been employed in research exploring the level of sustainability knowledge attainment by discipline (Zwickle et al., 2014) as well as the effect of knowledge attainment in relation to where students get their information from (Michel and Zwickle, 2021). While the ASK tool reports good psychometric properties for test takers with lower ability levels, it is less reliable for high ability ranges (Zwickle et al., 2014). Zizka and Varga (2020) used Sulitest's Core Module tool to identify gaps in students' sustainability knowledge, which they propose can be used to inform curricular (re)design. However, The Sulitest's Core Module was designed as an awareness-raising tool (Décamps et al., 2017), and as such, its psychometric properties are not proven (Keuhl et al., 2021).

2.4. Dealing with sustainability as a contested concept when defining the latent construct

At its core, sustainability is a contested concept, that can vary significantly in how it manifests in the higher education context (Reid and Petocz, 2006), and how learning outcomes are defined and measured (cf. Jickling and Wals, 2008). For example, the views of Deep Ecology prioritize the environmental system (Næss, 1997). The "intertwined" view sees the environmental, social, and economic systems as an overlapping Venn diagram with a win-win-win sweet sport in the middle (Marcus et al., 2010; Kurucz et al., 2014). The "embedded" view of sustainability, views the economic system as embedded in the social system, which is in turn is embedded in the environmental system (Marcus et al., 2010; Kurucz et al., 2014). This view has been visualized as tiered like a wedding cake (in the context of the SDGs, cf. Folke et al., 2016) or as embedded rings, like a doughnut (Raworth, 2017) to illustrate the embedded nature of human and ecological systems.

Depending on how sustainability is viewed, the latent construct (and thematic hierarchy) of a measurement tool could manifest with significant variation, which can in turn lead to significant variation in the results of assessments (Stough et al., 2018). Therefore, it is important to carefully consider the contested nature of sustainability when defining the latent construct. For example, if the views of Deep Ecology were to be used when defining the latent construct, the dominate logic would be that the ecologic system is most important. To reflect this logic, the main dimensions of the assessment tool might only include the main dimensions of the environmental system and planetary boundaries framework (e.g., Rockström et al., 2009). Given the roots of EfS in the environmental education, there is a tendency to resort back to environmental-focused rhetoric (Monroe, 2012; Lindstone et al., 2014), which can be seen in the emphasis of environmental topics in higher education sustainability assessment tools (Shriberg, 2002; Yarime and Tanaka, 2012; Kamal and Asmuss, 2013).

If an "intertwined view" were to be chosen, the dominant logic would be that ecological, social, and economic systems are of equal importance. To reflect this logic, the main dimensions of a measurement tool might align with the triple bottom line narrative, giving equal weight to the social system (people), the environmental system (planet), and the economic system (profit, later replaced with prosperity). This approach was chosen by Zwickle et al. (2014) in their tool to assess sustainability knowledge. However, such approaches, focusing on equal weight of systems and the importance of "balance", have been criticized in recent years for their lack of critical inquiry (Elkington, 2018).

Viewing sustainability as "embedded" implies that the economic (and other socially-constructed systems) are embedded within social systems. This view is less focused on win-win-win solutions, but rather on the embeddedness and interrelatedness of systems. For example, building on the UN's Millennium Development Goals (UNESCO, 2017) and the subsequent Sustainable Development Goals (UN, 2016), Raworth (2017) proposed the metaphor of a Doughnut with an inner circle to illustrate a social minimum that should be ensured and an outer circle to illustrate an environmental maximum that should not be transgressed. The space in between represents our ability to organize to ensure that these two boundaries are respected.

3. The assessment of sustainability knowledge tool (TASK)

The desire for a sustainability knowledge measurement tool with strong psychometric properties that employs an embedded conceptualization of sustainability lead the Sulitest organization to begin developing The Assessment of Sustainability Knowledge (TASK) tool. In Section 3.1, the definition of the latent construct and subsequent thematic (hierarchal) structure of the TASK is described. An early version of the tool was piloted during October–November 2022 within networks on higher education for sustainability (e.g., Students Organizing for Sustainability; Climate Students; SDSN Youth). The primary goal of the first pilot was to collect data regarding user perceptions as well as begin training the scoring model (1378 participants completed the test, resulting in 132,288 item responses). User perceptions about TASK garnished in Pilot 1 (described in Section 3.3.) also helped to inform the desired design of subsequent TASK items (described in Section 3.2). A second pilot was conducted was conducted in February 2023 with the primary objective to further train the scoring model (described in Section 3.4). To ensure that TASK is appropriate as a diagnostic tool across a broad spectrum of abilities (i.e., levels of sustainability knowledge), participants were recruited using SurveySwap's paid participant-matching service, which facilitates survey distribution among individuals seeking to complete academic surveys. Eligibility was restricted to English-speaking adults (C1 and C2 levels), aged 18–30, and a balanced gender among respondents. A total of 600 participants completed the survey (resulting in 57,600 item responses). All participants provided informed consent prior to participation. The tool was launched in March 2023. The data generated between March–June 2023, wherein 4349 TASK sessions were completed, are used in Section 3.5 to report the psychometric properties of the tool.

3.1. Structure of TASK

Aligning with an embedded view, sustainability is defined in the TASK tool as: meeting human welfare while not exceeding planetary boundaries and employing socially-constructed systems to do so (i.e., economic organization, governance, etc.). TASK is organized to reflect the thematic (hierarchal) structure of an embedded conceptualization of sustainability (Fig. 1). The structure of TASK has 3 dimensions at the highest level; 9 second-level dimensions; and 28 third-level dimensions. At the highest level (i.e., the 1st level of abstraction), are the three dimensions of: *Earth System*—the "environmental ceiling" (Rockström, 2009; Raworth, 2017); *Human Welfare*—the "social foundation" of a minimum human welfare (UN, 2016; UN, 2016; Raworth, 2017); and *Levers of Opportunity*—the socially-constructed mechanisms that enable (or hinder) the ability to meet the needs of human welfare without exceeding planetary boundaries (UN, 2019).

The planetary boundaries framework (Rockström, 2009) offers a quantitative perspective of the critical planetary thresholds which should not be crossed if humanity is to continue to develop and thrive. These boundaries include: climate change; biosphere integrity; freshwater use; land-system change; ocean acidification; novel entities; biogeochemical flows; atmospheric aerosol loading; and stratospheric ozone depletion. These have been clustered in the TASK structure into two domains that inform the 2nd level of abstraction for Earth Systems: *core* planetary boundaries and *regulating* planetary boundaries.

The *Social Foundation* is comprised of the minimum social needs derived from the social priorities specified in the UN's Millennium Development Goals and Sustainable Development Goals (UN, 2000, 2015), which include: nutrition; health; access to water and sanitation; housing and human settlements; access to energy; basic income; social equity; gender equity; education and culture; peace, justice and political voice; and access to networks/social interactions (Raworth, 2017). Within the TASK structure, these social needs have been clustered into three domains that inform the 2nd level of abstraction for Social Foundation: *safety and basic needs*, *social equity*, and *human flourishing*.

Levers of Opportunity are the mechanisms that enable (or hinder) the ability to meet human welfare foundations without exceeding planetary ceilings. They require both individual and collective human actions for building a sustainable future. In alignment with the UN (2019), these levers have been clustered into four domains that inform the 2nd level of abstraction for Levels of Opportunity in the TASK structure: *governance*; *economy and finance*; *science and technology*; and *individual and collection action.*

3.2. TASK format and item development

Given how sustainability is defined as a latent construct for the tool, items are developed to measure a test takers knowledge (i.e., conceptual understanding) of a boundary. To add dimensionality about the type of knowledge measured in TASK, for each theme, test takers descriptive knowledge (definitions and key concepts), contextualized knowledge (current state and trends), causal knowledge (major causes), and integrated knowledge (systemic impacts) are measured. Definition questions deal with descriptive knowledge and require that test takers define the construct (via principles, functions, or thresholds). Questions about current states and trends deal with contextualized knowledge regarding the current status and trajectories of either meeting or not meeting basic thresholds. Questions about major causes deal with causal knowledge about the underlying and systemic causes of boundary transgressions. Questions about systemic impacts deal with the effects of boundary transgression on other concepts. For Levers of Opportunity, test takers are required to demonstrate knowledge about these mechanisms for addressing the challenge of sustainability, therefore questions are organized around the dual perspectives of definition and trends. This results in 96 discrete item pools, for which items are developed and evaluated on a regular basis. While items can differ for each test taker in the same session (i.e., to prevent sharing answers), the overall difficulty for each test taker should remain similar (as ensured by the scoring model described in Section 3.4).

For each item pool, general learning outcomes were formulated in consultation with the literature on EfS (e.g., UNESCO, 2017; Bianchi et al., 2022). TASK items are formulated in a multiple-choice format, with one correct response and three incorrect responses. Based on Evans (1984), during the development of TASK items, attention is given to

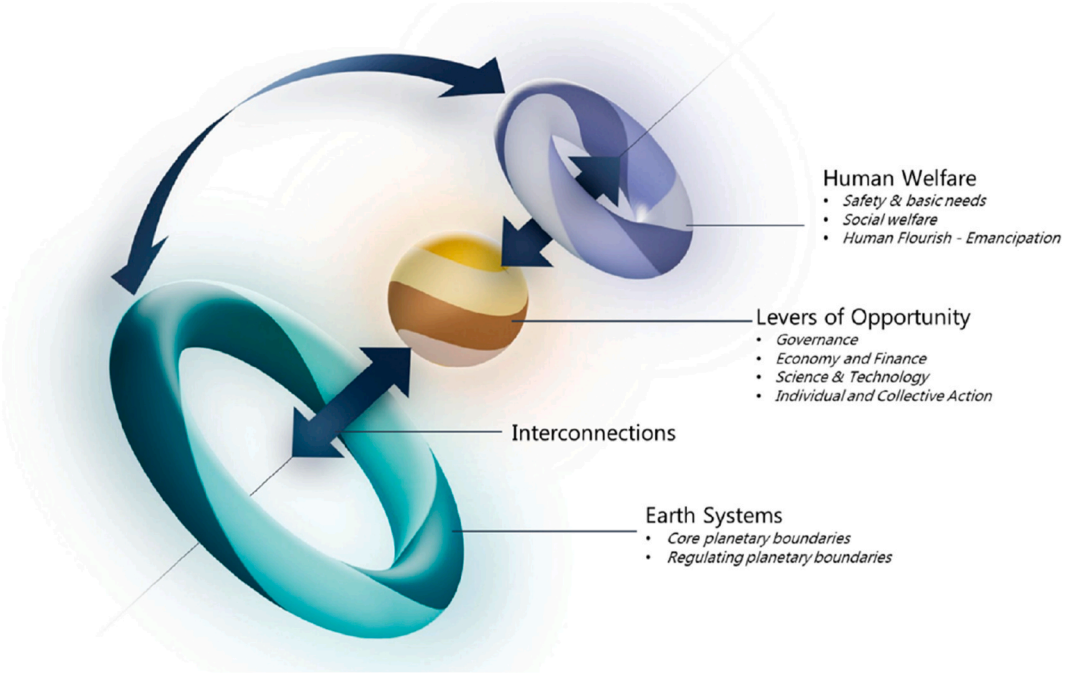


Fig. 1. Conceptualization of sustainability adopted in TASK with *Social Foundation* (UN, 2016) embedded in *Environmental Ceiling* (Rockström, 2009) as proposed by (Raworth, 2017) with the addition of *Levers of Opportunity* (UN, 2019).

ensure that they are: written clearly and consciously; avoid ambiguous wording; avoid multiple negatives; avoid difficult/technical, or culturally-specific language; that all incorrect answers are plausible; that response options are similar in wording; that response options are similar in length; that "always", "never", and "only" in incorrect answers are avoided; and that "all of the above" and "none of the above" in response options are avoided. Given user feedback in Pilot 1, special consideration is given to the use of technical/academic language. When a specific term is used in an item (e.g., "novel entities", "greenwashing"), but the item is not intended to measure knowledge of this term, a short definition is embedded in the item.

In response to user feedback from Pilot 1, items that intend to measure discrete facts are avoided in favor of phenomenon-based items. Sometimes, a specific factoid has been identified as part of the relevant learning outcomes for an item pool. For example, in the context of climate change, the goal of not exceeding 1.5 degrees of warming (United Nations Framework Convention on Climate Change, 2015). In such cases, attention is given to ensure that the range of response options are spread enough to avoid the item being too precise. To illustrate the operationalization of these considerations, Table 1 provides examples of TASK items for two themes: 1.1.1 Climate Change and 3.1.1 Laws, policies, institutions.

To support the organization of items, the software Notion is used to record information for each item, including: question title, matrix subject, type of knowledge, linked matrix subjects, related SDGs, corresponding learning outcome, bibliographical source, date of source publication, source region, source type, source hyperlink, date of question creation, question expiration date, author, reviewers (at least two), status in the validation process, and fields for reviewer and author comments. Everything logged into Notion is automatically saved and all changes are tracked.

3.3. User perceptions of TASK

Pilot participants reported a general alignment with the embedded conceptualization of sustainability, as well as subthemes included in the TASK tool. Participants found the lack of an environmentally-dominated conceptualization of sustainability a very positive attribute of the test. In the group sessions, many students noted that social sustainability themes were equally, if not more, meaningful to them, but they felt that environmental themes tend to take precedence in the context of sustainability courses.

In general, there was a perception that TASK was a difficult test (34 % found it "very difficult"; 62 % found it "slightly difficult"). This was coupled with participants finding it highly relevant to have a measurement of their sustainability knowledge. Participants of the group sessions reported having their curiosity about certain topics activated during the test, and some even reported looking up topics after the test for further understanding. When an items' link to sustainability was made explicit, it increased the relevancy of item and the test as a whole.

3.4. Scoring TASK

In the context of TASK, the latent construct of sustainability is defined as interrelated (e.g., causes of boundary transgression; impacts of boundary transgression on other boundaries; levels of opportunities to ensure boundaries are not transgressed; etc.) (Fig. 1). This interrelated interpretation of "sustainability" translates into an interrelated interpretation of "sustainability knowledge", wherein causes and effects of boundary transgressed are explicitly part of sustainability knowledge (Fig. 2). Therefore the assumption of unidimensionality for sustainability knowledge in the case of TASK was rejected. Instead, multidimensional IRT (MIRT) was employed for TASK.

The basis of the IRT model deployed is a Bayesian statistical framework modeled by exchangeability, which permits the investigation of prior information of group parameters (i.e., a vector) instead of individual parameters (i.e., components of a vector) (Lindley and Smith, 1972). Given that the matrix of TASK is organized in a thematic hierarchy, content of the tool was analyzed to reflect raised levels of abstraction. While investigating item parameters and respondents'

Table 1
Example of TASK item pool categorized into the four perspectives of: definition, trends, causes, and impacts.

	Definition	Trends	Causes	Impacts
Dimension: 1.1.1 **Climate change**	Which of the following indicators is used as the primary measure of the impact of human activity on the Earth's climate? a) the concentration of carbon dioxide (CO_2) in the atmosphere b) the aggregate amount of carbon emissions from industrial processes c) the combined average change in temperature of regional climates d) the average change in ocean temperature	To limit global warming to 1.5 °C above pre-industrial levels, global greenhouse gas (GHG) emissions must peak before 2025, and then decrease by 43 % by 2030 compared to 2019 levels. After adding up all the national emission-reduction commitments made by states in 2021, GHG emissions are currently projected to: a) increase by 14 % by 2030 b) stagnate by 2030 c) decrease by 14 % by 2030 d) decrease by 43 % by 2030	The primary driver of human-induced climate change is: a) greenhouse gas emissions b) aerosols, dust, smoke, and soot c) deforestation d) land-use change	The absorption of atmosphere carbon dioxide (CO2) emissions through dissolution in ocean water: a) reduces climate change but increases ocean acidification b) increases climate change but reduces ocean acidification c) reduces marine calcification but increases marine biodiversity d) increases marine calcification but decreases marine biodiversity
Dimension: 3.1.1 **Laws, policies, institutions**	The Sustainable Development Goals (SDGs) are an urgent call for action by all countries, developed and developing, in a global partnership. The SDGs build on decades of work by countries and which intergovernmental agency? a) the United Nations b) the Intergovernmental Panel on Climate Change c) the Intergovernmental Panel on Sustainable Development d) the Organisation for Economic Co-operation and Development	Official Development Assistance (ODA) is governmental aid from developed countries to developing countries with the aim of promoting "development" (economic, social, as well as climate-related targets). Which of the following statements best describes the trend in ODA over the last 50 years? a) While total ODA has increased from USD 40 billion (1960) to USD 160 billion (2020), the share of ODA as a percentage of donor countries' gross national income (GNI) has steadily declined over this period b) …		

TASK™ Matrix by Sulitest			x.1. Knowing and Understanding				x.2. Interlinkages			
			x.1.1 Definitions and Key Concepts *Descriptive Knowledge* What are we talking about? How does this work?		x.1.2 Current State and Trends *Contextualized Knowledge* Where are we now? How are things changing?		x.2.1 Major Causes *Causal Knowledge* Why is this happening? Who is doing what and why?		x.2.2 Systemic Impacts *Integrated Knowledge* What are the related effects? How is this affecting the larger system?	
1. Earth Systems The Environmental Ceiling	1.1 Core Planetary Boundaries	1.1.1 Climate Change	A	1.1.1.1.1	C	1.1.1.1.2	D	1.1.1.2.1	E	1.1.1.2.2
		1.1.2 Biosphere Integrity		1.1.2.1.1		1.1.2.1.2		1.1.2.2.1		1.1.2.2.2
	1.2 Regulating Planetary Boundaries	1.2.1 Freshwater Use		1.2.1.1.1		1.2.1.1.2		1.2.1.2.1		1.2.1.2.2
		1.2.2 Land-System Change		1.2.2.1.1		1.2.2.1.2		1.2.2.2.1		1.2.2.2.2
		1.2.3 Ocean Acidification		1.2.3.1.1		1.2.3.1.2		1.2.3.2.1		1.2.3.2.2
		1.2.4 Novel Entities		1.2.4.1.1		1.2.4.1.2		1.2.4.2.1		1.2.4.2.2
		1.2.5 Biogeochemical Flows		1.2.5.1.1		1.2.5.1.2		1.2.5.2.1		1.2.5.2.2
		1.2.6 Atmospheric Aerosols Loading		1.2.6.1.1		1.2.6.1.2		1.2.6.2.1		1.2.6.2.2
		1.2.7 Stratospheric Ozone Depletion		1.2.7.1.1		1.2.7.1.2		1.2.7.2.1		1.2.7.2.2
2. Human Welfare The Social Foundation	2.1 Safety and Basic Needs	2.1.1 Nutrition		2.1.1.1.1		2.1.1.1.2		2.1.1.2.1		2.1.1.2.2
		2.1.2 Health		2.1.2.1.1		2.1.2.1.2		2.1.2.2.1		2.1.2.2.2
		2.1.3 Access to Water and Sanitation		2.1.3.1.1		2.1.3.1.2		2.1.3.2.1		2.1.3.2.2
		2.1.4 Housing and Human Settlements		2.1.4.1.1		2.1.4.1.2		2.1.4.2.1		2.1.4.2.2
		2.1.5 Access to Energy		2.1.5.1.1		2.1.5.1.2		2.1.5.2.1		2.1.5.2.2
	2.2 Social Welfare	2.2.1 Basic Income and Decent Work		2.2.1.1.1		2.2.1.1.2		2.2.1.2.1		2.2.1.2.2
		2.2.2 Social Equity		2.2.2.1.1		2.2.2.1.2		2.2.2.2.1		2.2.2.2.2
		2.2.3 Gender Equality		2.2.3.1.1		2.2.3.1.2		2.2.3.2.1		2.2.3.2.2
	2.3 Human Flourishing	2.3.1 Education and Culture		2.3.1.1.1		2.3.1.1.2		2.3.1.2.1		2.3.1.2.2
		2.3.2 Peace, Justice, and Political Voice		2.3.2.1.1		2.3.2.1.2		2.3.2.2.1		2.3.2.2.2
		2.3.3 Access to Networks and Social Interaction		2.3.3.1.1		2.3.3.1.2		2.3.3.2.1		2.3.3.2.2
3. Levers of Opportunity That Make Sustainability Possible	3.1 Governance	3.1.1 Laws, Policies, and Institutions	B	3.1.1.1.1		3.1.1.1.2				
		3.1.2 Infrastructure, Planning, and Natural Resource Management		3.1.2.1.1		3.1.2.1.2				
	3.2 Economy and Finance	3.2.1 Macroeconomic Considerations and Finance		3.2.1.1.1		3.2.1.1.2				
		3.2.2 Microeconomic Considerations, Business, and Industry		3.2.2.1.1		3.2.2.1.2				
	3.3 Science and Technology	3.3.1 Sustainability Science		3.3.1.1.1		3.3.1.1.2				
		3.3.2 Technology and Innovation		3.3.2.1.1		3.3.2.1.2				
	3.4 Individual and Collective Action	3.4.1 Transformative Change		3.4.1.1.1		3.4.1.1.2				
		3.4.2 Cognitive Capacity for Sustainable Development		3.4.2.1.1		3.4.2.1.2				

N.B. Letters indicate the order in which TASK questions appear in the assessment.

Within each lettered category, TASK questions are randomized.

Fig. 2. Dimensions of the TASK (*based* Rockström et al., 2009; Raworth, 2017; UN, 2019).

abilities, the hierarchy (i.e., levels of abstraction) allows for the treatment of an item's location in the hierarchal structure of the test (Fig. 2) as prior information of test item parameters. To represent the relationship between a respondent and test items, alongside the relationship between the hierarchal structure of the test and items, multidimensional indexing was applied in the analysis. The dataset's indices are no longer represented by a vector of integers, rather the index now contains a multitude of levels expressing: respondents' user ID, the location in TASK's structure where a question is located, and the item ID (i.e., the question's unique identifying number) simultaneously as index values.

The model operates as a two-stage procedure. The initial phase of the model focuses primarily on the estimation of item parameters, specifically discrimination (denoted by a) and difficulty (denoted by b). The large and diverse calibration sample collected in Pilot 1 was used during this stage to train the model and compute these parameters for each item. In the second stage of the model, the focus was shifted towards estimating the abilities or traits (denoted by θ) of new individuals engaging with the test. This two-stage procedure leverages predetermined item parameters to estimate new test-takers' abilities, eliminating the need for perpetual re-estimation of item parameters. Figs. 3 and 4 show the respective first and second stages of the applied model.

Incorporating item and subject-level discrimination and difficulty parameters into the model allows for a comprehension of how individuals with varied ability levels, interact with distinct sustainability questions and topics. The parameters from both stages converge into a Bernoulli distribution to form a probability function. This function determines the probability of a respondent correctly answering an item, considering their ability level and the known parameters of the test. Bayesian sampling was achieved via Markov Chain Monte Carlo (MCMC) methods. Parameters were estimated using the No-U-Turn Sampler, a Hamiltonian Monte Carlo (HMC) algorithm as proposed by Hoffman and Gelaman (2014).

3.5. Psychometric properties of TASK

In the evaluation of the model sampling performance, the principles for diagnostic tools proposed by Gelman et al. (2021) were employed. Throughout the development of TASK, both the Rhat and ESS scores have consistently stayed within the acceptable ranges as proposed by

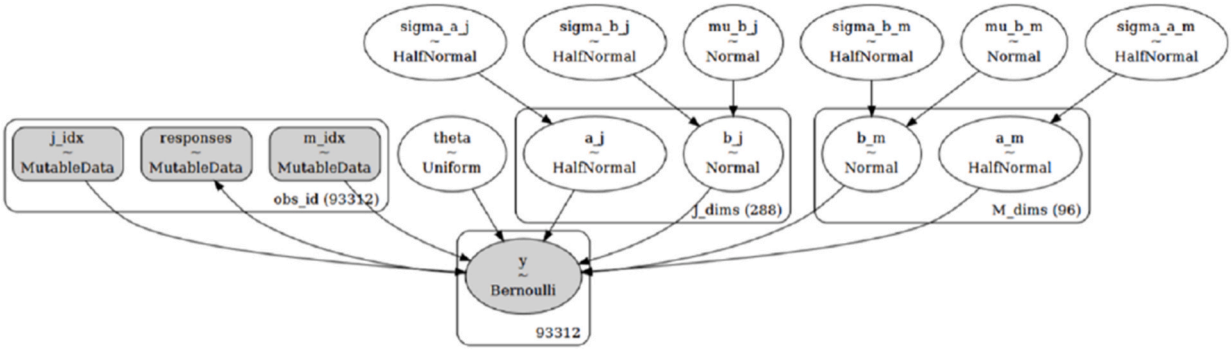


Fig. 3. First stage item parameter estimation.

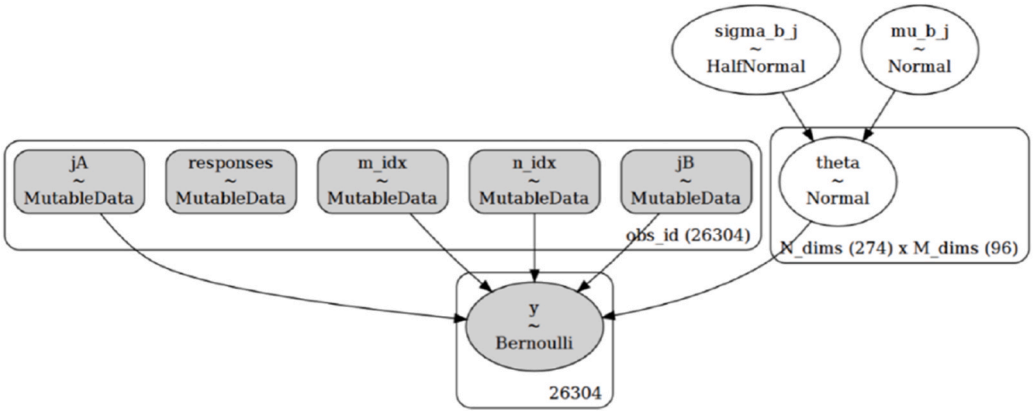


Fig. 4. Second stage ability estimation.

Gelman et al. (2021), implying satisfactory chain convergence. In assessing the reliability of a test under the lens of IRT, we focused our analysis on the Fisher Information statistic as a measure that provides an illustration of the association between the quantity of information (i.e., the precision) and an examinee's ability across the entire ability spectrum (Baker and Kim, 2004). Coefficients of psychological/educational measurements are often represented by a value between 0 and 1, with higher values indicating higher reliability (Nunnally, 1978; Reckase, 2009). In the early stages of research, a reliability coefficient of .70 might suffice, however, for applied settings, a coefficient of .80 is more desirable, and in the context of high-stakes testing, a value of .90 or higher is typically considered the standard (Nunnally, 1978). The reliability of the TASK tool is high (>.90) for the *entire spectrum* of ability; and very high (>.95) for ability levels .4–2.9; and extremely high (>.99) for ability levels .75–2.6. These results illustrate the strong psychometric properties of the TASK tool across ability levels, verifying that TASK is an appropriate tool regardless of test taker's level of sustainability knowledge.

Integrating this into the context of the Test Information Function (TIF)—a graphical representation within the IRT framework—provides an intuitive way to visually interpret reliability across different levels of the latent trait. In a TIF graph, the x-axis signifies varying levels of the latent trait, while the y-axis symbolizes the amount of test information (a direct reflection of measurement precision) at each trait level (Fig. 5).

Item Information Functions (IIFs) offer another layer of granularity in understanding of test reliability. The IIFs represent the amount of information, or measurement precision, that each item contributes at various levels of the latent trait. Visualizing these in Fig. 6, illustrates the distributed nature of TASK items across most of the latent trait spectrum.

The IIFs are optimally concentrated around a high-middle level of ability, indicating that TASK items are most reliable and informative for individuals in this range. Having examined the reliability of the TASK test through the lens of MIRT, it can be confidently concluded that the instrument is well-calibrated to provide accurate and meaningful insights into respondents' abilities.

4. Discussion

The new TASK tool fills a gap in the landscape of measuring sustainability knowledge by: 1) employing an embedded conceptualization of sustainability and 2) ensuring strong psychometric properties. During the development of TASK, insights were garnished that can further inform the field of measuring sustainability knowledge. A robust measurement of sustainability knowledge can be helpful for curricular (re) design. However, while a tool can have strong psychometric properties and thus demonstrate strong reliability, given the contested nature of the sustainability as a construct and the lack of empirical insights about the strength of various sustainability-related learning outcomes, there continues to be open debate about what EfS should develop in learners.

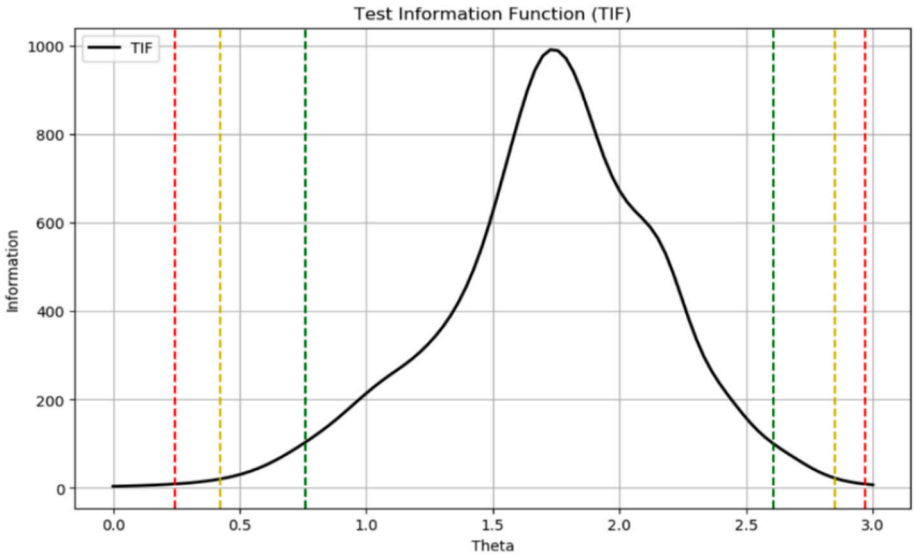


Fig. 5. Test Information Function (TIF) of data since March 2023.

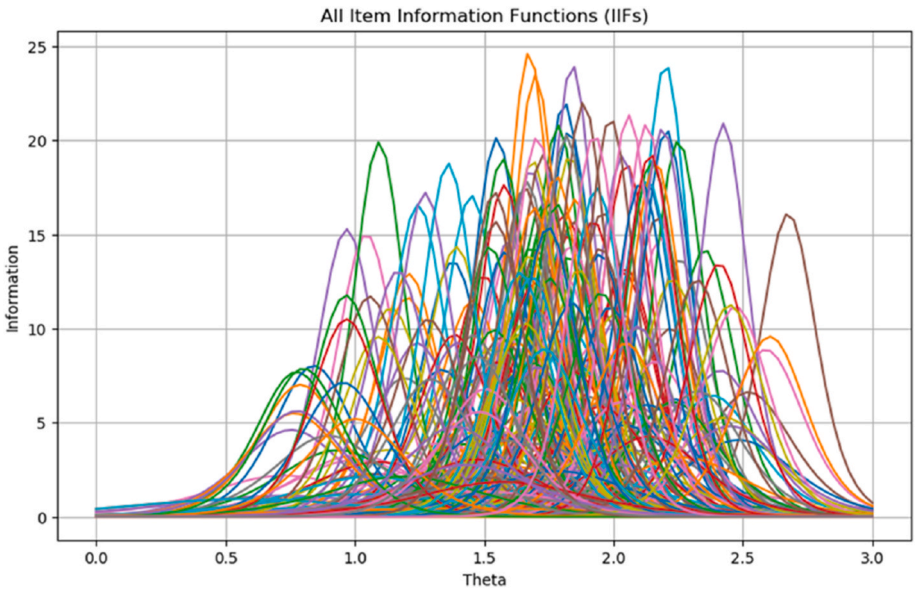


Fig. 6. Item Information Function for TASK items (March–June 2023).

4.1. Rejecting the assumption of unidimensionality for sustainability knowledge

For a diagnostic tool (i.e., used to measure gaps in knowledge), the psychometric properties of the overall score, as well as the subscores, is important for the reliability of the measurement (Yavuz Temel et al., 2022), which has been an issue for some tools intended to measure sustainability knowledge (Kuehl et al., 2021). In addition to these concerns, we posit that the assumption of unidimensionality is unlikely to hold in the context of sustainability knowledge given the inter-relatedness of sustainability themes. Indeed, while investigating a sample of TASK respondents from business schools, Stough et al. (2024b) found the ability level of test takers are highly correlated for many topics in TASK. As such, the assumption of unidimensionality is unlikely to hold in the context of sustainability knowledge.

4.2. Measuring knowledge attainment to inform curricular (Re)Design

The attainment of learning outcomes could be used to inform curricular (re)design (Mendoza et al., 2022). As proposed by Zizka and Varga (2020), measuring learner's level of sustainability knowledge could inform where gaps are (for example, when entering into a study program), which would be valuable to inform educators about the design of curricular inputs (e.g., themes included in the course). In this vein, some universities have committed themselves to using TASK to track the attainment of sustainability knowledge of their student population in reflection with curricular design. For example, the Toulouse Business School (TBS) uses TASK in pre- and post-assessments for their Master students, as part of Assurance of Learning (AoL) processes (Sulitest, 2024). Stough et al. (2024b) found that business students' TASK scores increased with the frequency in which they took standalone courses explicitly on sustainability-related themes (which was not the case for engineering students), suggesting that integrating explicit sustainability-related courses is an effective pathway of sustainability integration in business programs.

4.3. Increasing relevancy of the test taking experience

Through developing and piloting TASK, important considerations arose that should be taken into account during the assessment of sustainability knowledge. First, items that intend to measure discrete facts were noted as problematic by pilot participants. Such questions contributed to the perception of difficulty, as well as the propensity to guess, and reduced the overall sense of meaning that taking an assessment has. These questions should be avoided when developing items to measure sustainability knowledge. When a specific factoid is part of desired learning outcomes for a sustainability-related topic, spreading response options widely can help mitigate the negative perception of such questions.

In addition to avoiding discrete factual questions, lengthy questions and technical terms should be avoided to improve the understandability of items. When technical terms are used in an item, embedding definitions can be used to clarify these for the test taker. Explanations can be embedded to clarify the relevance of specific items. TASK pilot participants this noted improved their perception about the relevancy of the test-taking experience, and some pilot participants even reported looking up information after the test because their interest was sparked. While not investigated in this research, increased relevancy could help foster the subjectification of learning (Keys and Heck, 2023).

4.4. General versus disciplinary sustainability knowledge

The TASK tool employs an embedded view of sustainability—recognizing the embeddedness of the economic and other socially-constructed systems within the social system, which is in turn embedded in the environmental system. In doing so, TASK moves beyond the intertwined view (e.g., balance of environmental, social, and economic themes) found in some sustainability knowledge assessments (e.g., Zwickle et al., 2014), while still avoiding an environmentally-dominated conceptualization of sustainability found in other assessment tools in the higher education context (e.g., Kopnina, 2013), which was appreciated by pilot participants.

However, an embedded conceptualization of sustainability could also equate to less prevalence of questions relating to certain disciplines (e.g., business and economics), as socially-constructed systems are situated within Levers of Opportunity. Indeed, pilot participants from business and economics faculties particularly noted that they felt their disciplinary knowledge of sustainability was not being fully measured by TASK. Two opposing reflexes arise from this. The first reflex is that learners should attain *disciplinary-focused* sustainability knowledge. Following this logic, specific measurement tools for various disciplinary-specific sustainability knowledge could be developed (e.g., Aichele et al., 2021; Muff et al., 2022). The second reflex a reflection about a *minimum level of general knowledge* that all learners should have about

sustainability, regardless of their disciplinary background. In the case of business and economics students, while understanding the intricacies of the carbon, nitrogen, or phosphorus cycles might seem too technical, knowledge about these cycles (i.e., benefits, causes and effects to threshold transgression, etc.) could very well be needed to sense, seize and capture potential shared value creation (Bocken and Geradts, 2020). Hence, what constitutes disciplinary-specific literacy is changing as the boundaries of social, environmental, and economic systems become increasing blurred.

4.5. Limitations and future research pathways

While this research takes important steps to fill critical gaps in the EfS landscape, there are limitations. As discussed above, TASK is intended to measure an embedded knowledge of sustainability as defined in Table 1. As presented in Section 2.1, other intended learning outcomes (e.g., skills, attitudes, norms, etc.) are also likely important behavioral determinants (Montaño and Kasprzyk, 2015) that are not captured in TASK scores.

Understanding the influence of educational interventions on learning outcomes has long been a "blackbox" in the EfS landscape. Moving forward, TASK can be used (in addition with other measurement instruments) to understand relationships between: 1) educational interventions (e.g., pedagogies, program design, accreditation status of a university or faculty, involvement in voluntary initiatives, etc.), 2) learners (e.g., age, gender, experience, personality, culture, etc.), and learning outcomes (e.g., knowledge, skills, attitudes, behavior). Building on the work of Dyehouse et al. (2017), Keys and Heck (2023), and others, TASK scores could be used to understand the influence of a qualification like knowledge attainment, on socialization processes, and ultimately the process of subjectification—i.e., how learners transform knowledge, skills, and values into action.

As research begins to illustrate the usefulness of TASK scores for curricular design (e.g., Stough et al., 2024b), insights about the predictive validity of sustainability knowledge attainment (as measured by TASK or other instruments) could help inform EfS. While higher levels of environmental knowledge have been shown to be positive predictors of pro-environmental behavior (e.g., Zelezny et al., 2000), little is known about the effect of higher levels of sustainability knowledge attainment on the future behavior of graduates. For example, while mental health problems from environmental issues (i.e., eco-anxiety) of young people are of increasing concern (McCunn et al., 2024), emerging insights suggest that increased (environmental) knowledge could be associated with lower levels of (climate change) anxiety (Zacher and Rudolph, 2023). Such an effect would widen the benefit of sustainability-related knowledge attainment, and would motivate the inclusion of this as a intended learning outcome in HEIs beyond its role as a behavioral determinants in the context of EfS.

5. Conclusions

The contested nature of sustainability and the lack of measurement instruments with strong psychometric properties has hindered empirical research on sustainability knowledge attainment, which has left the field of EfS stifled from many relevant insights. The TASK tool has been developed to foster more measurement of sustainability knowledge. Through developing and piloting TASK, we offer novel insights to further theoretical and practical discussion on measuring sustainability knowledge. However, the development and validation of a measurement tool is merely a first step in unlocking the blackbox of EfS. We readily invite more research and discussion on sustainability knowledge attainment, with the hopes of informing the (re)design of EfS so that future decision makers can indeed respect social and planetary boundaries, thus meeting the aim of SDG 4.7.

CRediT authorship contribution statement

Talia Stough: Writing – review & editing, Writing – original draft, Project administration, Methodology, Formal analysis, Data curation, Conceptualization. **Alexander Brewer:** Methodology, Formal analysis, Data curation. **Aurélien Decamps:** Supervision, Project administration, Conceptualization. **Scott Blair:** Writing – review & editing, Conceptualization. **Wim Lambrechts:** Writing – review & editing, Supervision, Funding acquisition. **Estela Castelli Florino Pilz:** Writing – review & editing, Formal analysis, Data curation, Conceptualization. **Marjolein C.J. Caniëls:** Writing – review & editing, Supervision. **Jean-Christophe Carteron:** Writing – review & editing, Supervision, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Talia Stough reports financial support was provided by Dutch Research Council. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

A special thanks to Evelyne Gross, Fabrice Galia, Emeric Fortin, Benoit Martimort-Asso, Gerard Vidal, Kathleen Ng, Aude Serrano, Narciso Antunes, Pierre Schulz, Iain Patton, and Julia Solans Rossi for their contribution to the development of the TASK tool. A special thanks to all the participants of the TASK pilot studies, and to the students and instructors of the qualitative group sessions, for informing the analysis of the TASK tool. This research was funded in part by the Dutch Research Council (grant number 055.19.201).

Data availability

The data that has been used is confidential.

References

Association for the Advancement of Sustainability in Higher Education (AASHE), 2024. STARS Technical Manual v3.0: AC 6 – Sustainability Literacy Assessment.

Association for the Advancement of Sustainability in Higher Education (AASHE), 2025. Sustainability Literacy/Culture assessment tools. https://docs.google.com/spreadsheets/d/1bJtXOa1_bJFF5uAGcpdWpAu82oXMXeCnQa6EenjJJRM/edit?gid=0#gid=0.

Aichele, C., Hartig, J., Michaelis, C., 2021. Assessing learning progress: validating a test score interpretation in the domain of sustainability management. Stud. High Educ. 46 (10).

Baker, F.B., Kim, S.H. (Eds.), 2004. Item Response Theory: Parameter Estimation Techniques. CRC press.

Berzosa, A., Bernaldo, M.O., Fernández-Sanchez, G., 2017. Sustainability assessment tools for higher education: an empirical comparative analysis. J. Clean. Prod. 161, 812–820.

Besong, F., Holland, C., 2015. The dispositions, abilities and behaviours (Dab) framework for profiling learners' sustainability competencies in higher education. J. Teach. Educ. Sustain. 17 (1), 5–22.

Bianchi, G., Pisiotis, U., Cabrera, M., 2022. GreenComp, the European Sustainability Competence Framework. European Commission, Luxembourg, EU.

Bocken, N., Geradts, T.H.J., 2020. Barriers and drivers to sustainable business model innovation: organization design and dynamic capabilities. Long. Range Plan. 53 (4).

Connelly, S., 2007. Mapping sustainable development as a contested concept. Local Environ. 12 (3), 259–278.

Daenekindt, S., Huisman, J., 2020. Mapping the scattered field of research on higher education. A correlated topic model of 17,000 articles, 1991–2018. High. Educ. 80 (3), 571–587.

Décamps, Aurélien, Barbat, G., Carteron, J.C., Hands, V., Parkes, C., 2017. Sulitest: a collaborative initiative to support and assess sustainability literacy in higher education. Int. J. Manag. Educ. 15 (2), 138–152.

DeVellis, R., 2017. Scale Development Theory and Applications. Sage Publications, Thousand Oaks, CA.

Dyehouse, M., Weber, N., Fang, J., Harris, C., David, R., Hua, I., Strobel, J., 2017. Examining the relationship between resistance to change and undergraduate engineering students' environmental knowledge and attitudes. Stud. High Educ. 42 (2), 390–409.

Elkington, J., 2018. 25 years ago I coined the phrase "triple bottom line." here's why it's time to rethink it. Harv. Bus. Rev. 25, 2–5.
Evans, J., 1984. Heuristic and analytic processes in reasoning. Br. J. Psychol. 75 (4), 451–468.
Folke, C., Biggs, R., Norström, A.V., Reyers, B., Rockström, J., 2016. Social-ecological resilience and biosphere-based sustainability science. Ecol. Soc. 21 (3).
Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., Rubin, D., 2021. Bayesian Data Analysis. CRC press.
Gericke, N., Boeve-de Pauw, J., Berglund, T., Olsson, D., 2019. The Sustainability Consciousness Questionnaire: the theoretical development and empirical validation of an evaluation instrument for stakeholders working with sustainable development. Sustain. Dev. 27, 35–49.
Gutierrez-Bucheli, L., Kidman, G., Reid, A., 2022. Sustainability in engineering education: a review of learning outcomes. J. Clean. Prod. 330, 129734.
Gutiérrez-Mijares, M.E., Josa, I., Casanovas-Rubio, M. del M., Aguado, A., 2023. Methods for assessing sustainability performance at higher education institutions: a review. Stud. High Educ. 48 (8), 1137–1158.
Hartig, J., Hohler, J., 2009. Multidimensional IRT models for the assessment of competencies. Stud. Educ. Eval. 35 (2–3), 57–63.
Hoffman, M., Gelaman, A., 2014. The No-U-Turn sampler: adaptively setting path lengths in hamiltonian Monte Carlo. J. Mach. Learn. Res. 15, 1351–1381.
Jickling, B., Wals, A.E.J., 2008. Globalization and environmental education: looking beyondsustainable development. J. Curric. Stud. 40 (1), 1–21.
Kamal, A., Asmuss, M., 2013. Benchmarking tools for assessing and tracking sustainability in higher educational institutions: identifying an effective tool for the university of Saskatchewan. Int. J. Sustain. High Educ. 14 (4), 449–465.
Keys, N., Heck, D., 2023. Positioning and repositioning in higher education: first year students engaging with the world. Stud. High Educ. 1–14.
Kopnina, H., 2013. Evaluating education for sustainable development (ESD): using Ecocentric and Anthropocentric Attitudes toward the Sustainable Development (EAATSD) scale. Environ. Dev. Sustain. 15, 607–623.
Kuehl, C., Sparks, A., Hodges, H., Smith, E., 2021. The incoherence of sustainability literacy assessed with the Sulitest. Nat. Sustain. 4, 555–560.
Kurucz, E., Colbert, B., Marcus, J., 2014. Sustainability as a provocation to rethink management education: building a progressive educative practice. Manag. Learn. 45 (4), 437–457.
Lambrechts, W., Mulà, I., Ceulemans, K., Molderez, I., Gaeremynck, V., 2013. The integration of competences for sustainable development in higher education: an analysis of bachelor programs in management. J. Clean. Prod. 48, 65–73.
Leal Filho, W., Lange Salvia, A., Pires Eustachio, J.E., 2023. An overview of the engagement of higher education institutions in the implementation of the UN sustainable development goals. J. Clean. Prod. 386, 135694-135694.
Lindley, D., Smith, A., 1972. Bayes estimates for the linear model. J. Roy. Stat. Soc. B 34 (1), 1–41.
Lindstone, L., Wright, T., Sherren, K., 2014. Canadian STARS-Rated campus sustainability plans: priorities, plan creation and design. Sustainability 7, 725–746.
Marcus, J., Kurucz, E., Colbert, B., 2010. Conceptions of the business-society-nature interface: implications for management scholarship. Bus. Soc. 49 (3), 402–438.
McCunn, L.J., Osborne, B., Wister, A.V., 2024. Eco-depression and eco-anxiety among youth: a sex and gender analysis. Can. J. Psychiatr. 69 (5), 315–323.
Mendoza, W., Ramirez, G., González, C., Moreira, F., 2022. Assessment of curriculum design by Learning Outcomes (LO). Educ. Sci. 12. 1.
Michel, J.O., Zwickle, A., 2021. The effect of information source on higher education students' sustainability knowledge. Environ. Educ. Res. 27 (7), 1080–1098.
Montaño, D.E., Kasprzyk, D., 2015. Theory of reasoned action, theory of planned behavior, and the integrated behavioral model. In: Glanz, K., Rimer, B.K., Viswanath, K. (Eds.), Health Behavior: Theory, Research, and Practice, fifth ed. Jossey-Bass, pp. 95–124.
Monroe, M., 2012. The Co-evoluation of ESD and EE. Journal of Education for Sustainable Development 6 (1), 43–47.
Muff, K., Delacoste, C., Dyllick, T., 2022. Responsible Leadership Competencies in leaders around the world: assessing stakeholder engagement, ethics and values, systems thinking and innovation competencies in leaders around the world. Corp. Soc. Responsib. Environ. Manag. 29 (1), 273–292.
Næss, A., 1997. Sustainable development and the deep ecology movement. In: Baker, S., Kousis, M., Richardson, D., Young, S. (Eds.), The Politics of Sustainable Development. Routledge, London.
Nunnally, J.C., 1978. Psychometric Theory, second ed. McGraw-Hill, New York.
Raworth, K., 2017. Doughnut Economics, Seven Ways to Think like a 21st-Century Economist. Random House Business Books, London.
Reckase, M.D., 2009. Multidimensional Item Response Theory. Springer, New York, USA.
Rockström, J., Steffen, W., Noone, K., Persson, Å., Chapin, S., Lambin, E., Lenton, T., Scheffer, M., Folke, C., Schellnhuber, H.J., Nykvist, B., de Wit, C.A., Hughes, T., van der Leeuw, S., Rodhe, H., Sörlin, S., Snyder, P.K., Costanza, R., Svedin, U., Falkenmark, M., Karlberg, L., Corell, R.W., Fabry, V.J., Hansen, J., Walker, B., Liverman, D., Richardson, K., Crutzen, P., Foley, J., 2009. Planetary boundaries: exploring the safe operating space for humanity. Ecol. Soc. 14 (2), 32.
Redman, A., Wiek, A., Barth, M., 2021. Current practice of assessing students' sustainability competencies: a review of tools. Sustain. Sci. 16 (1), 117–135.
Reid, A., Petocz, P., 2006. University lecturers' understanding of sustainability. High. Educ. 51 (1), 105–123.
Sandri, O., Holdsworth, S., Thomas, I., 2018. Assessing graduate sustainability capability post-degree completion: why is it important and what are the challenges? Int. J. Sustain. High Educ. 19 (1), 2–14.
Shaw, A., Liu, O.L., Gu, L., Kardonova, E., Chirikov, I., Li, G., Hu, S., Yu, N., Ma, L., Guo, F., Su, Q., Shi, J., Shi, H., Loyalka, P., 2020. Thinking critically about critical thinking: validating the Russian HEIghten® critical thinking assessment. Stud. High Educ. 45 (9), 1933–1948.
Shriberg, M., 2002. Institutional assessment tools for sustainability in higher education: strengths, weaknesses, and implications for practice and theory. High Educ. Pol. 15, 153–167.
Sijtsma, K., van der Ark, L.A., 2020. Measurement Models for Psychological Attributes: Classical Test Theory, Factor Analysis, Item Response Theory, and Latent Class Models. Chapman and Hall/CRC.
Stough, T., Ceulemans, K., Lambrechts, W., Cappuyns, V., 2018. Assessing sustainability in higher education curricula: a critical reflection on validity issues. J. Clean. Prod. 172, 4456–4466.
Stough, T., Gross, E., Blair, S., Lambrechts, W., Francisco Carías Álvarez, J., 2024a. Learning Outcomes in the context of Education for Sustainability: foci, articulations, and assessments. In: Federico Rotondo, F., Giovanelli, L., Lozano, R. (Eds.), Sustainability in Higher Education: Strategies, Performance, and Future Challenges. Springer International Publishing.
Stough, T., Lambrechts, W., Brewer, A., Kourula, A., Moosmayer, D., Ceulemans, C., Castillo, M., Decamps, A., Caniels, M., 2024b. Ready to tackle sustainability issues? Business students' knowledge of sustainability. Proceedings of the 84th Annual Meeting of the Academy of Management (AOM). Chicago, USA, August 9-13, 2024.
Stough, T., Lambrechts, W., Desmet, L., Gültzow, 2025. Moving responsible management education beyond the rhetoric. *Proceedings of the 85th Annual Meeting of the Academy of Management (AOM)*. Copenhagen, Denmark, July 25-29, 2025.
Sulitest, 2024. Featuring our Change Leaders on the global stage: the stories our community shared in key education events. https://www.sulitest.org/news/change-leaders-the-stories-our-community-shared-in-key-education-events.
United Nations, 2000. United Nations Millennium Declaration (A/RES/55/2). https://www.un.org/millennium/declaration/ares552e.htm.
United Nations Educational, Scientific and Cultural Organization (UNESCO), 2017. Education for sustainable development goals: learning objectives. https://unesdoc.unesco.org/images/0024/002474/247444e.pdf.
United Nations Framework Convention on Climate Change, 2015. Paris agreement. https://unfccc.int/sites/default/files/english_paris_agreement.pdf.
UN, 2016. United Nations Sustainable Development Goals. http://www.un.org/sustainabledevelopment/sustainable-development-goals/.
UN, 2019. The future is now. https://sdgs.un.org/sites/default/files/2020-07/24797GSDR_report_2019.pdf.
Yarime, M., Tanaka, Y., 2012. The issues and methodologies in sustainability assessment tools for higher education institutions: a review of recent trends and future challenges. Journal of Education for Sustainable Development 6 (1), 63–77.
Yavuz Temel, G., Machunsky, M., Rietz, C., Okropiridze, D., 2022. Investigating subscores of VERA 3 German Test based on item response Theory/Multidimensional item response theory models. Front. Educ. 7.
Zacher, H., Rudolph, C.W., 2023. Environmental knowledge is inversely associated with climate change anxiety. Clim. Change 176 (32).
Zelezny, L.C., Chua, P.P., Aldrich, C., 2000. Elaborating on gender differences in environmentalism. J. Soc. Issues 56 (3), 443–457.
Zizka, Laura, Varga, Peter, 2020. Teaching sustainability in higher education institutions: assessing hospitality students' sustainability literacy. J. Hosp. Tour. Educ. 1–16.
Zwickle, A., Koontz, T., Slagle, K., Bruskotter, J., 2014. Assessing sustainability knowledge of a student population. Int. J. Sustain. High Educ. 15 (4), 375–389.
Zwickle, Adam, Jones, Keith, 2018. Sustainability knowledge and Attitudes—Assessing latent constructs. In: Handbook of Sustainability and Social Science Research. Springer, Cham, pp. 435–451.