

Development and validation of a multi-tier diagnostic test on the water footprint concept

International
Journal of
Sustainability in
Higher Education

729

Stergios Gkitsas^{id} and Georgios Malandrakis^{id}

*Department of Primary Education, Aristotle University of Thessaloniki,
Thessaloniki, Greece*

Nicolaos Theodossiou^{id}

*Department of Civil Engineering, Aristotle University of Thessaloniki,
Thessaloniki, Greece, and United Nations Sustainable Development Solutions
Network SDSN Black Sea, Thessaloniki, Greece, and*

Nikos Lambrinos

*Department of Primary Education, Aristotle University of Thessaloniki,
Thessaloniki, Greece*

Received 17 March 2026
Revised 7 July 2026
Accepted 4 September 2026

Abstract

Purpose – Conceptual understanding is essential for effective sustainability education, yet students often hold alternative conceptions about key environmental issues. The water footprint (WF), a multifaceted indicator of water use, is increasingly included in sustainability curricula but remains conceptually challenging. This study aims to describe the development and validation of the Water Footprint Diagnostic Instrument (WFDI), a multi-tier diagnostic test designed to assess undergraduate primary education students' scientific and alternative conceptions of the WF.

Design/methodology/approach – The construction of the instrument followed Treagust's principles for developing multi-tier tests and involved three phases: defining propositional knowledge statements, identifying students' alternative conceptions through two qualitative pilot studies ($n = 64$; $n = 94$) and developing the final multi-tier tool. The WFDI was piloted with 56 students, after which minor wording revisions were made before its final administration to 238 students from three Greek universities.

Findings – The reliability of the WFDI was acceptable (Cronbach's $\alpha = 0.866$), and item analysis showed an appropriate spread of difficulty and discrimination values. The results indicate that the tool successfully exposes a sharp misalignment between student accuracy and confidence, revealing that pre-service teachers' knowledge is often context-dependent and fragmented rather than structurally unified.

Practical implications – The WFDI provides Higher Education Institutions (HEIs) with a data-driven pathway to reform teacher-training curricula and to target specific sustainability competencies. Specifically, it is valuable for identifying undergraduate students' initial ideas, developing and evaluating targeted instructional interventions, informing curriculum design and supporting future research on WF-related sustainability topics.

Originality/value – This study addresses a significant gap in sustainability education by developing the first validated multi-tier diagnostic instrument designed to assess a core and complex sustainability concept, that of



© Stergios Gkitsas, Georgios Malandrakis, Nicolaos Theodossiou and Nikos Lambrinos. Published by Emerald Publishing Limited. This article is published under the Creative Commons Attribution (CC BY 4.0) licence. Anyone may reproduce, distribute, translate and create derivative works of this article (for both commercial and non-commercial purposes), subject to full attribution to the original publication and authors. The full terms of this licence may be seen at <http://creativecommons.org/licenses/by/4.0/>

International Journal of
Sustainability in Higher Education
Vol. 27 No. 10, 2026
pp. 729-752
Emerald Publishing Limited
1467-6370
DOI 10.1108/IJSHE-03-2026-0486

Water Footprint. By explicitly linking psychometric data to conceptual change learning theories, the instrument moves beyond purely descriptive testing to reveal structural flaws in future educators' systemic thinking about complex, global water-use systems.

Keywords Sustainability in higher education, Sustainability literacy, Education for sustainable development (ESD)

Paper type Research paper

Introduction

Water management is a critical issue towards sustainability and requires an understanding of how water use is connected to environmental integrity, social well-being and responsible patterns of production and consumption (Gleick and Cooley, 2021; Hoekstra *et al.*, 2011). The development of a water footprint diagnostic tool offers a valuable pedagogical pathway for strengthening sustainability education on the topic. By rendering visible the often-hidden water demands embedded in production and consumption processes, such a tool can promote more informed, evidence-based and systems-oriented learning (Chaney and Doukopoulos, 2018). It may also contribute to the quality and depth of students' understanding by linking individual choices with broader environmental, social and ethical dimensions of sustainable resource use. Thus, the need for quality conceptual understanding is particularly evident in sustainability education, where new and complex concepts and indicators, such as the water footprint (WF), are increasingly used to describe and communicate environmental impacts; yet, students' understanding of these concepts remains insufficiently examined (Barth *et al.*, 2007; Cebrián and Junyent, 2015; Lozano *et al.*, 2017; Ma *et al.*, 2025).

This gap highlights a broader institutional challenge, as Higher Education Institutions (HEIs) often struggle to move beyond generic environmental awareness towards deep, competency-based sustainability literacy (Leal Filho *et al.*, 2021; Rieckmann, 2012). Despite this institutional need, no research has rigorously assessed the actual conceptual understanding of the WF, and a reliable diagnostic instrument has been lacking. Existing quantitative tools emphasise the assessment of behaviour or awareness rather than the quality of learning. Therefore, the present study aims to address this gap by constructing and validating a multi-tier diagnostic instrument specifically for the WF. This tool would provide a multi-dimensional, quality, and reliable assessment of undergraduate primary education students' understanding of the concept.

In particular, the targeted group of undergraduate students play a crucial role in promoting sustainability education in primary schools. Gaps or misconceptions in their own understanding compromise the effectiveness of their teaching and diminish the quality of their future students' learning. Thus, reliable tools for assessing their learning and conceptual progression towards specific scientific concepts are essential. In the context of environmental conceptual understanding, "misconceptions" refer to stable, intuitive, but scientifically inaccurate mental models. Specifically, regarding the WF, these flawed frameworks arise when individuals conflate a product or activity's total water impact with only visible, direct water consumption (e.g. domestic use). In doing so, they systematically ignore the invisible, indirect "virtual water" embedded across global supply chains.

In addition, the WF, along with the Ecological and Carbon Footprints, is included in the new Greek education for sustainability curricula for primary schools (IEP, 2021). Hence, teachers must align their teaching not only with international trends in sustainability education, where environmental footprints have a prominent role, but also comply with the national curricula, where water-related topics are included.

Based on the above, the present study aims to address the following two research questions:

- RQ1. What are the psychometric characteristics (reliability, difficulty and discrimination) of the developed Water Footprint Diagnostic Instrument (WFDI) when applied to undergraduate primary education students?
- RQ2. How does students' performance on Water Footprint understanding change when the assessment moves from a single, content-based tier to a combined multi-tier framework that in addition includes reasoning and confidence?

Theoretical background

Sustainability literacy and conceptual understanding of the water footprint

Sustainability literacy is increasingly recognised as a foundational pillar for tertiary education institutions, as it aims to equip future citizens and policymakers with the knowledge and skills needed to address contemporary global challenges (Rieckmann, 2012). According to the literature, genuine sustainability literacy goes beyond passive memorisation of environmental facts or fragmented ecological information (e.g. Holst *et al.*, 2025). Instead, it requires developing key sustainability competencies, with systems thinking being particularly important, as it enables individuals to understand the complex interconnections between human activities and natural ecosystems (Wiek *et al.*, 2011). However, efforts to integrate sustainability into higher education institutions often remain fragmented and disconnected from the daily structures and routines of university culture (Holst *et al.*, 2025). In this context, conceptual understanding of modern sustainability metrics such as the WF is a vital component of sustainability literacy, as it enables students to critically assess the indirect environmental consequences of their daily consumer choices.

Regarding the importance of the topic at hand, it is widely recognised that humans consume natural resources faster than they are regenerated, leading to many undesirable consequences (Kılıç, 2020; Martins *et al.*, 2019; Soe and Yeo-Chang, 2019). Especially in the case of water, human activities consume and pollute large amounts of water that originates from natural ecosystems (Akhtar *et al.*, 2021). Globally, water is used primarily by crops, but freshwater ecosystems also provide other services to society; they regulate water flows, purify wastewater and detoxify industrial and domestic wastes, mitigate soil erosion and provide cultural benefits, including aesthetic, educational and spiritual benefits (Chapagain and Orr, 2008; Hoekstra *et al.*, 2011; WWAP, 2009). The draining of freshwater from ecosystems in quantities and at rates that exceed nature's "renewable" capacity is widely documented in many parts of the world (Gleick and Cooley, 2021).

Thus, the adoption of the principles of sustainability, a term officially introduced in 1987 (UN, 1987), is the first proposed solution for mitigating water-related problems. Water is central to the United Nations 2030 Agenda (UN, 2015), as several of the 17 Sustainable Development Goals are directly or indirectly related to freshwater (e.g. Goal #6, clean water and sanitation; Goal #14, life below water; Goal #15, life on land). The WF, an indicator of water use based on consumption, provides useful information about water use by individuals and communities, which can lead to more sustainable utilisation of water resources (Hoekstra, 2003). WF is increasingly popular among scientists, policymakers and the public because it effectively communicates water-related issues and highlights sustainable alternatives (Liu *et al.*, 2020). Given the importance of the issue, younger generations need to understand the concept of WF and learn to make appropriate, environmentally conscious decisions regarding the consumption of products and services, aiming to reduce it, as they will be most affected by environmental problems in the future (Venckute *et al.*, 2017).

To date, research on WF in education remains in its early stages and is heavily focused on documenting superficial trends. Most existing quantitative studies use traditional questionnaires (e.g. Likert-type scales) to assess students' consumption behaviour, general sensitivity, values, or attitudes towards water sustainability (e.g. [Çamur et al., 2020](#); [Elmaslar Özbaş et al., 2021](#); [Panatsa and Malandrakis, 2024](#)). As [Godfrey and Feng \(2017\)](#) have pointed out, simply providing information about the WF is not sufficient to change consumers' daily behaviour. Extending this argument to educational assessment, documenting quantitative data or positive attitudes is also insufficient to evaluate deep conceptual understanding. The main flaw of these self-report tools is that they cannot assess the quality of learning. They assume that ecological sensitivity equates to scientific understanding, leaving students' deeper cognitive structures unexplored.

On the other hand, researchers who sought to explore students' understanding of water literacy issues used qualitative assessment methods, such as drawing analysis and interviews, focusing on systems thinking and the hydro-social cycle (e.g. [Cole et al., 2024](#); [Horne et al., 2025](#); [Pozo-Muñoz et al., 2023](#)). Although these methods (e.g., student drawings) reveal more systemic connections than written responses ([Horne et al., 2025](#)), they have inherent limitations: they require very time-consuming qualitative analysis, involve subjective interpretation by the researcher and, most importantly, are typically restricted to small samples. This makes it impossible to generalise findings to larger populations.

This methodological dichotomy in the literature leaves the most critical dimensions of WF, such as the understanding of virtual water, unexplored. WF is a complex and multifaceted concept (blue, green and grey components) that relies on an understanding of global interdependencies ([Hoekstra et al., 2011](#)). Research shows that students have difficulty with these concepts ([Owens et al., 2020](#)), while [Gkitsas et al. \(2024\)](#) found that the main misconception among undergraduate primary education students is the confusion of WF with virtual water. Despite theoretical proposals for the use of constructivist models that connect WF with systems thinking ([Eskici, 2025](#)), there has been no systematic effort to develop a stable, objective tool suitable for assessing the understanding of a complex concept like WF and for large-scale implementation.

Therefore, to achieve genuine sustainability literacy and water literacy, research must bridge this gap. Assessment tools need to simultaneously embody the cognitive, value-based and behavioural dimensions of learning ([Dal-Farra et al., 2015](#)), enabling the connection of WF with real decision-making scenarios ([Chaney and Doukopoulos, 2018](#)). Furthermore, to complement the assessment of the cognitive basis with a metacognitive dimension, the present study introduces multi-tier diagnostic tools. Beyond mapping core content knowledge, these instruments incorporate distinct tiers of reasoning and confidence. This operational layout successfully combines the depth of qualitative misconception detection with the statistical power of large-scale sampling.

The evolution of diagnostic instruments: from single to multi-tier tests

In the field of educational assessment, traditional one-tier multiple-choice tests have, for decades, been the most widely used tools for measuring students' knowledge, due to their ease of grading and suitability for large-scale implementation ([Haladyna and Rodriguez, 2013](#); [Tamir, 1971](#)). However, international science education literature highlights two serious methodological disadvantages of these tools ([Adadan and Savasci, 2012](#); [Haladyna and Rodriguez, 2013](#)). Firstly, they allow participants to select the correct answer by guessing (lucky guessing), which distorts the accurate assessment of their knowledge. Secondly, and more importantly, when a student selects an incorrect answer, the one-tier test cannot determine whether this is due to a lack of knowledge or to a deeply rooted

misconception. This limitation has led researchers, with Treagust (Treagust, 1988, 1995) at the forefront, to seek more valid diagnostic instruments.

This evolution led to the development of multi-tier diagnostic instruments, which introduce deeper layers of evaluation to science education assessment. While two-tier tests added an explicit reasoning component to capture the underlying justifications for students' choices, a significant methodological upgrade was achieved through the integration of a confidence tier. As documented in contemporary literature (Arslan *et al.*, 2012; Caleon and Subramaniam, 2010; Ma *et al.*, 2025; Putica, 2023), requesting respondents to rate their certainty on a confidence scale explicitly embeds a metacognitive dimension into the diagnostic process. This structural advancement allows researchers to rigorously differentiate between stable alternative conceptions and random errors. When a student systematically selects a scientifically flawed answer but remains entirely certain of its correctness, it indicates a deeply rooted misconception. This framework acts as a robust barrier to learning rather than a temporary lack of knowledge.

Multi-tier tests have seen rapid development and validation in traditional science subjects such as physics, chemistry, and biology (Espinosa *et al.*, 2025; Ma *et al.*, 2025). However, their application in modern sustainability metrics remains limited. To date, such efforts have typically been restricted to macroeconomic environmental phenomena such as global warming and atmospheric degradation (Aksoy and Erten, 2022; Arslan *et al.*, 2012). Apart from the instrument developed by Liampa *et al.* for the ecological footprint, the literature lacks a valid multi-tier test for assessing complex environmental concepts. The absence of a tool for WF prevents academic instructors from identifying their students' specific cognitive difficulties. The present study addresses this significant gap by developing and validating the WFDI, thereby extending a rigorous diagnostic methodology for assessing complex sustainability metrics, such as the WF.

Conceptual change in science education: from cognitive structures to diagnostic tools

The foundational premise of the conceptual change literature, as synthesised by Duit and Treagust (2003, 2008), is that learning complex scientific and environmental concepts is not a simple process of accumulating new facts but rather a profound restructuring of pre-existing cognitive frameworks. Within science and sustainability education, students do not enter the classroom as blank slates; instead, they bring deeply embedded, intuitive ideas shaped by their daily interactions with the natural world. Duit and Treagust (2003) emphasised that for effective science teaching and learning to occur, instructional designs must explicitly acknowledge and target these pre-instructional conceptions. Furthermore, Treagust and Duit (2008) pointed out that navigating the theoretical, methodological and practical challenges of conceptual change requires robust diagnostic efforts. Identifying the exact nature of students' cognitive starting points is an essential prerequisite for designing targeted learning environments that foster genuine scientific literacy.

To understand the cognitive architecture of these alternative conceptions (or misconceptions), the literature is deeply influenced by two main competing models. The first model, Framework Theory, posits that students possess a relatively coherent and systematically organised knowledge structure derived from their daily experiences (Vosniadou, 1994, 2013). When confronted with scientific concepts that contradict their intuitive beliefs, learners attempt to assimilate the new information into their existing framework. This assimilation process often leads to the creation of synthetic models or robust misconceptions that display internal coherence and are highly resistant to traditional teaching methods (Vosniadou and Brewer, 1992). In contrast, the second model, known as 'Knowledge in Pieces,' rejects the notion of coherent pre-instructional knowledge structures

(diSessa, 1993, 2014). This perspective argues that students' knowledge is inherently fragmented, consisting of a loose network of distinct, intuitive elements termed 'phenomenological primitives' or 'p-prims' (diSessa, 1993). These p-prims activate automatically, depending heavily on the specific context of the task, meaning that alternative conceptions are not stable mental structures but rather momentary, *ad hoc* cognitive constructions generated on the spot due to the misactivation of these fragmented elements (diSessa *et al.*, 2004).

The methodological design of multi-tier diagnostic instruments offers a unique opportunity to investigate this theoretical debate between coherence and fragmentation while addressing the practical challenges outlined by Treagust and Duit (2008). By analysing the consistency of student response patterns across items and tiers, together with metacognitive confidence scores, researchers can empirically assess whether incorrect ideas are systematically stable or fluctuate with context (Gobert and Buckley, 2000; Özdemir and Clark, 2007). In the present study, this theoretical distinction between Framework Theory and Knowledge in Pieces is not used to validate one cognitive model over the other. Instead, in line with Duit and Treagust (2003) call to use conceptual change frameworks to improve classroom practice, it serves as an interpretive framework in the Discussion section. This framework enables a deeper investigation into whether pre-service teachers' difficulties with the WF concept stem from robust cognitive obstacles or fragmented knowledge, thereby providing actionable insights for designing effective sustainability modules and curricula in higher education.

Methodology

Instrument development and materials

The WFDI was developed following the methodology established by Treagust (1988, 1995) for the development of multi-tier tests in science education (e.g. Arslan *et al.*, 2012; Irmak *et al.*, 2023; Putica, 2023), and it was further elaborated and adapted in sustainability education by Liampa *et al.* (2019) for the investigation of the Ecological Footprint concept, which is a similar concept to the WF. Rather than adopting a rigid, uniform format for all items, the developed WFDI utilises a flexible, hybrid, multi-tier framework specifically designed to capture the unique conceptual complexities of the WF concept. "Depending on the nature of each propositional knowledge statement, individual questions combine up to three types of tiers. These include a content tier assessing *which* option is scientifically correct, a reasoning tier exploring *why* that choice is justified and a confidence tier measuring the respondent's certainty. This operational layout ensures that the scoring patterns can effectively distinguish between a genuine lack of knowledge (low confidence ratings) and deep-seated, confident misconceptions (incorrect cognitive choices paired with high confidence), fully aligning the instrument's structure with the distinct conceptual requirements of the targeted content areas. By evaluating these layers simultaneously, the instrument moves beyond purely descriptive testing; it maps the stability of students' underlying cognitive structures, aligning psychometric measurement with the mechanisms of conceptual change.

The development process of these tier tests comprised three main phases (Liampa *et al.*, 2019): conceptual definition, assessment of students' conceptions and development of the diagnostic instrument (Figure 1).

Initially (*Phase 1*), a concept map was constructed, and nine propositional knowledge statements (PKS) were formulated for the WF concept. These PKS were considered fundamental for determining whether someone understands the concept and were derived from the relevant literature defining WF (see Table 1). The content validity of these nine PKS

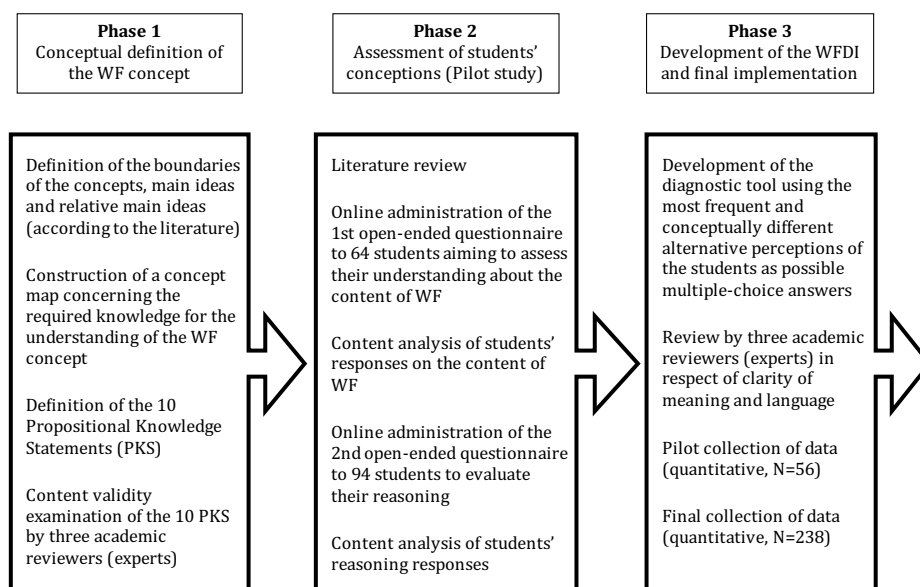


Figure 1. Phases of the development of the WFDI

Source: Authors' own work

was independently examined by three academics: two specialising in sustainability education and one in civil engineering with expertise in water resources management and water footprint. Feedback from the reviewers indicated that the instrument aligns with the conceptual principles of WF and that the PKS and the concept map are accurate and comprehensive. At the end of this process, the questions included in the WFDI were organised into the following five content areas as shown in [Table 1](#):

- (1) conceptual definition of WF (4 PKS, 7 questions);
- (2) usefulness and importance of WF (2 PKS, 2 questions);
- (3) determinants of WF calculation (1 PKS, 1 question);
- (4) determinants of individual WF and ways to reduce it (1 PKS, 5 questions); and
- (5) implications of an increase in individual WF (1 PKS, 1 question).

Through this procedure, content validity within the boundaries of the WF concept and appropriateness for students from the Departments of Education were ensured.

Phase 2 comprised two consecutive studies aimed at collecting qualitative data for the formulation of the content tiers ($n=64$) and the reasoning tiers ($n=94$). Due to the lack of previous research, these data were used to map and explore possible alternative conceptions about the WF among undergraduate primary education students. This sequential design was a deliberate choice to allow the bottom-up approach to guide the development of the diagnostic tool more effectively compared to a single combined study (e.g. [Liampa et al., 2019](#)).

In the first study, an open-ended version of the questionnaire was administered to 64 undergraduate students to collect qualitative data regarding all nine PKS. The collected answers were subjected to content analysis to produce conceptually and qualitatively distinct

Table 1. Content areas and propositional knowledge statements

Content area	Subcategories	Propositional knowledge statements	#Question (Q)
Conceptual definition of WF	1. Scientific definition of WF	Water footprint is “the cumulative virtual water content of all goods and services consumed by one individual or by the individuals of one country” (Hoekstra, 2003)	2
	2. Scientific definition of VW	Virtual water is “the volume of water that has been used to produce a commodity, and that is thus virtually embedded in it” (Allan, 1998)	3
	3. Units of measurement	The water footprint is measured in cubic meters per ton of production (m ³ /t) or in liters (L) per capita per day (L/c/d) (Hoekstra, 2003)	4
	4. WF Components	There are Three components of the water footprint: the <i>blue</i> component relates to the use of surface and groundwater; the <i>green</i> component relates to the use of precipitation water (e.g. rain, snow, etc.); the <i>grey</i> component relates to the use of freshwater required to dilute polluted water to meet existing water quality standards (Hoekstra et al., 2011)	5–8
Usefulness – importance of WF	5. Usefulness of WF	The water footprint provides valuable information about how individuals and countries use water and about their impact on the environment, which can in turn lead to more sustainable use (Hoekstra, 2003)	9
	6. Importance of WF	The significance of the water footprint lies in raising awareness about the water content of products and services, as well as the potential changes that can be achieved in production, diets and consumer behaviour (Lovarelli et al., 2016).	17 (Likert-type scale)
Determinants of WF calculation	7. Determinants of WF calculation	• Domestic water consumption • Outdoor water consumption • Consumer habits • Energy consumption • Recycling (Okutan and Akkoyunlu, 2021; Zhou et al., 2023)	10
Determinants of individual WF	8. Ways to reduce individual WF	• Saving water in household activities (e.g. turning off the tap while brushing teeth, using dual-flush toilets) • Saving water in outdoor activities (e.g. using a rain barrel to collect water for garden irrigation) • Reducing the purchase of clothing and footwear • Reducing meat consumption • Adopting a vegetarian or vegan diet • Saving energy (e.g. using A-class energy household appliances) • Reducing the use of cars and airplanes	11–15

(continued)

Table 1. Continued

Content area	Subcategories	Propositional knowledge statements	#Question (Q)
Implications	9. Implications of an increase in individual WF	• Recycling packaging and appliances (Aleksandrowicz <i>et al.</i>, 2016; Gerbens-Leenes and Hoekstra, 2011; Jalava <i>et al.</i>, 2014; Okadera <i>et al.</i>, 2014; Vanham <i>et al.</i>, 2013; Zucchinelli <i>et al.</i>, 2021) • Water scarcity • Water pollution • Food shortage • Destruction of flora, fauna and freshwater ecosystems • Acceleration of soil erosion (Chapagain and Orr, 2008; Hoekstra <i>et al.</i>, 2011; WWAP, 2009)	16

categories of students' ideas. These categories were classified by their frequency and importance, and the most common alternative ideas were directly used as distractors (multiple-choice options) for all content tiers of the tool.

Subsequently, a new version of the questionnaire was developed to formulate the reasoning tiers, in which all eight content tiers requiring justification were presented in a closed format. The first seven content tiers each had five options, adapted from Liampa *et al.* (2019), while the last content tier had eight options derived from the previous study. This version featured nine open-ended questions (one for each of the first seven content tiers and two for the last one) focusing on the PKS concerning "Ways to reduce individual WF" and "Implications of the increase in individual WF."

This instrument was administered to 94 students, who were asked to select a content option and provide a written, open-ended justification for their choice. As in the first study, content analysis of these responses was conducted to create distinct categories of students' ideas, which then served as the alternative answer options for the final reasoning tiers. This sequential step-by-step process ensured that the final reasoning distractors were directly anchored in authentic student reflections on specific content decisions.

In the *third phase* of the WFDI development, and to differentiate a lack of knowledge from a misconception, a 7-point certainty response index was added to 15 questions (29 items), ranging from "1-Absolutely unconfident (Just Guessing)" to "7-Absolutely Confident," with the balance point at "4-Neither Confident/Neither Unconfident." Through this answer option range, optimal discrimination and reliability of students' answers were ensured (Garbayo *et al.*, 2023).

Overall confidence was assessed only in Q10 and Q16, as each of these two questions contained many items, and it was considered that assessing confidence for each individual item would significantly increase response time and cause participant fatigue. No confidence rating was given for Q15 because its items were not homogeneous, meaning they contained conceptually distinct topics with varying levels of difficulty.

This version of the tool received feedback from three experts in sustainability education and civil engineering, and a Greek language expert, aiming to eliminate ambiguous expressions and refine clarity of meaning and language proficiency. Before the main administration, a pilot implementation of the final WFDI ($n = 56$) was conducted to examine face validity and item clarity. Minor wording adjustments were made before the full-scale implementation of the developed tool to 238 undergraduate primary school students.

The final instrument (see [Appendix](#)) consists of 37 closed-ended items: eight single-tier content items, 22 two-tier items (content and confidence), and six three-tier items (content, reasoning and confidence). In addition, it includes one attitudinal item measuring participants' self-assessed importance of the WF concept and one open-ended, two-tier item (Q1) for estimating the average individual WF, adapted from [Ferreira da Luz et al. \(2015\)](#). To prevent artificial forced choices, most content and reasoning questions offered seven options, including a "blank line option" allowing students to write an alternative answer if their specific idea was not listed ([Caleon and Subramaniam, 2010](#)). Confidence questions also included seven possible options. A typical three-tier item from the WFDI is illustrated below:

Q12.i. Content Tier: What do you think is the effect of meat consumption on a person's Water Footprint?

- Increases WF* ([Dota et al., 2016](#));
- Decreases WF;
- It also depends on other factors;
- Does not affect WF; and
- I do not know.

Q12.ii. Reasoning Tier: For what reason does this happen?

- Because of the food kilometres the meat travels until it reaches the end consumer.
- With the frequent consumption of meat, fewer animals remain and therefore less water is needed.
- It depends on the type of animal, because a chicken and a cow require different amounts of water.
- Because a large amount of water is used and polluted for the production, processing and distribution of meat* ([Dota et al., 2016](#)).
- Meat has a very large Ecological Footprint, mainly due to the feed the animals need to grow, but a very small Water Footprint.
- Other (write what you think).
- I do not know.

Q12.iii. Confidence Tier: How confident are you in your answer?

- Absolutely unconfident (just guessing);
- Very unconfident;
- Unconfident;
- Neither confident/neither unconfident;
- Confident;
- Very confident;
- Absolutely confident.

Note: Asterisks (*) indicate the scientifically accepted correct choices.

Participants

The selection of undergraduate students from Greek primary education departments provides a highly relevant context for developing and validating the WFDI. In Greece, sustainability

education is an increasingly vital component of teacher-training curricula. Undergraduate primary education students represent a critical target group because they will eventually be responsible for teaching environmental and water-saving concepts to young learners. Identifying their baseline understandings and alternative conceptions is therefore essential for improving teacher preparation programs. Furthermore, by involving participants from three regional institutions (Aristotle University of Thessaloniki, University of Western Macedonia and the University of Crete), the study captures a broader geographic and demographic representation, thereby strengthening the overall validity and reliability of the diagnostic instrument.

In the WFDI development process, four distinct groups of participants took part. For the development of the content tiers, a group of 64 undergraduate students (56 females, 8 males) participated. The participants were first-year undergraduates in the Department of Primary Education at Aristotle University of Thessaloniki and had already received a one-hour session on WF in an academic education for sustainability course.

A second group of 94 undergraduate primary school students (82 females, 12 males) provided data for the development of the reasoning tiers. These participants were also first-year undergraduates in the Department of Primary Education at Aristotle University of Thessaloniki. They had attended a similar one-hour session on the WF during the following academic year.

Before the large-scale implementation of the WFDI tool, a third group of 56 students (41 females, 15 males) was used for its final pilot testing. They were studying at the Department of Primary Education at the University of Crete and were enrolled in different semesters. These three separate groups of students were not included in the final main study.

In the final main study, 238 undergraduate students (87.0% female, 12.6% male, 0.4% other/self-described) from the Departments of Primary (87.0%) and Early Childhood Education (13.0%) participated. Specifically, 59.2% of participants were from Aristotle University of Thessaloniki, 24.4% from the University of Western Macedonia and 16.4% from the University of Crete. Regarding their year of study, 60.1% were in their first year, 18.5% in their second year, 10.1% in their third year and 8.4% in their fourth year. In addition, 2.9% had exceeded the standard four-year duration of their study programs. Finally, 64.3% of participants had attended a course on sustainability education, which also included WF, while 35.7% had not been exposed to such teaching.

Procedure

In three of the four groups of students ($n = 64, 94, 238$) who participated in this study, the questionnaires were administered individually in supervised classroom settings using the Google Forms platform. Only in the final pilot phase ($n = 56$), conducted before the main study, was the WFDI administered in a paper-and-pencil format to confirm face validity, ensure item clarity and gather direct feedback from participants.

Students needed approximately 30 min to complete the first open-ended version of the questionnaire and about 40 min for the second open-ended version. For the final version, in which all questions were in closed form, about 20 min were required. All students were informed that the test aimed to assess their understanding of the WF concept and identify potential difficulties. This feedback would be used to develop improved teaching and learning materials. They were also informed that their responses would be used solely for academic purposes and would not affect their course grades. In addition, they were assured of the anonymity and confidentiality of their responses and were asked to provide honest answers without collaboration.

The data collection period for the development of the WFDI was from spring 2022 to spring 2023, while the data for the main study (pilot and final) were gathered from May 2024 to April 2025.

Oral consent was obtained from all participants, and the study received approval from the Research Ethics Committee of the Department of Primary Education of Aristotle University of Thessaloniki.

Data analysis

The data from the two open-ended questionnaires were analysed using content analysis, with the basic unit of analysis being the unit of meaning, to create conceptually and qualitatively distinct categories of student responses to the topics under study (Krippendorff, 2013). This was performed for both the content and the reasoning students' answers using Atlas.ti software. To mitigate potential researcher bias, interrater reliability was interpreted using the benchmarks proposed by Landis and Koch (1977), ranging from slight to almost perfect agreement. In cases of disagreement, discussion among coders took place until they reached a common decision. In addition, to minimise participant response bias, all pilot questionnaires were completed anonymously.

For the final WFDI version, the data from the content and reasoning tiers were coded as correct (1) or incorrect (0). This approach was adopted because the primary focus was instrument validation rather than an in-depth analysis of misconceptions. Consequently, "blank line" responses were coded as incorrect (0) since none represented scientifically acceptable ideas. The confidence tier was coded on a 7-point Likert-type scale (1 = Absolutely Unconfident, 7 = Absolutely Confident), and the final attitudinal item was also coded on a 7-point scale (1 = Very Unimportant, 7 = Very Important). Following the literature (Espinosa *et al.*, 2025; Liampa *et al.*, 2019), confidence scores of 5–7 were considered confident and scores of 1–3 were considered unconfident, while a score of 4 was considered neither unconfident nor confident, carrying distinct interpretive value (Garbayo *et al.*, 2023).

Descriptive statistics (frequencies, means and standard deviations) were calculated using Microsoft Excel. Content and face validity were established as described in Methodology (Instrument development and materials) and were further supported by the pilot testing of the WFDI. Overall reliability for both the pilot and the final implementation of the WFDI was assessed using Cronbach's alpha on the total raw and standardised data with SPSS software. Item difficulty, item discrimination and confidence-related indices were computed using Microsoft Excel. Rather than treating these metrics strictly as psychometric filters, confidence-related indices were used as operational measures to pinpoint alternative ideas and assess the stability or fragmentation of students' underlying cognitive structures.

Results

Instrument validation and reliability analysis

For the data from the two open-ended questionnaires, interrater agreement was considered excellent according to Landis and Koch (1977) for both the content-focused and reasoning-focused questionnaires (Cohen's kappa = 0.87 and 0.91, respectively).

Analysis of the raw data from the closed-form pilot study ($n = 56$) yielded a Cronbach's $\alpha = 0.898$ for the overall WFDI, and $\alpha = 0.875$ for the standardised data. The standardised coefficient was calculated to ensure that dichotomous responses (0–1) and Likert-type confidence responses (1–7) contributed equally to the reliability estimate. Similarly high Cronbach α scores were recorded, and in the final implementation of the WFDI ($n = 238$), yielding an alpha coefficient of 0.866 for the raw data and 0.823 for the standardised. All values fall within acceptable reliability ranges, indicating that the WFDI demonstrates stable internal consistency and is suitable for assessing undergraduate primary education students' understanding of the WF.

To examine the internal structure of the WFDI more closely and evaluate the functioning of individual items, an item analysis was conducted. This analysis included item difficulty (Facility Index), item discrimination (Discrimination Index) and several confidence-related indices derived from students' responses (Table 2). The analysis provides further evidence of the diagnostic quality and validity of the WFDI.

FI = Proportion of students who selected the correct options. Range: 0.0% (min) – 100.0% (max).

DI = Item Discrimination Index (proportion of correct responses from the top 25% achievers in the total score – proportion correct responses in the bottom 25% of achievers). Range: –1.00 (min) – 1.00 (max).

CF = Mean Confidence of students in the item. Range: 0.00 (min) – 7.00 (max).

Table 2. Item difficulty, discrimination and confidence-related indices ($n = 238$)

Item	Facility index (FI) (% correct)		Discrimination index (DI)		Confidence				
	Content	Both tiers	Content	Both tiers	CF	CFC	CFW	CAQ	CB
1	0.4	–	0.02	–	3.63	–	–	–	–
2	50.8	–	0.28	–	4.04	4.25	3.82	0.33	–
3	59.7	–	0.59	–	4.10	4.30	3.79	0.33	–
4	50.8	–	0.30	–	4.35	4.28	4.43	–0.08	–
5	58.4	–	0.66	–	4.59	5.01	3.99	0.49	–
6	27.7	–	0.56	–	4.08	4.94	3.74	0.65	–
7	43.7	–	0.56	–	3.99	4.47	3.62	0.46	–
8	45.8	–	0.64	–	4.13	4.67	3.67	0.54	–
9	55.5	–	0.30	–	4.13	4.12	4.14	–0.01	–
10.a	96.2	–	0.10	–	4.14	–	–	–	–
10.b	85.3	–	0.26	–	4.14	–	–	–	–
10.c	94.5	–	0.11	–	4.14	–	–	–	–
10.d	79.4	–	0.19	–	4.14	–	–	–	–
10.e	97.9	–	0.06	–	4.14	–	–	–	–
10.f	73.9	–	0.22	–	4.14	–	–	–	–
10.g	59.7	–	0.21	–	4.14	–	–	–	–
10.h	76.5	–	0.27	–	4.14	–	–	–	–
11	44.5	28.2	0.61	0.50	4.04	4.45	3.88	0.38	0.15
12	54.6	31.9	0.53	0.44	3.94	4.50	3.78	0.50	0.10
13	56.7	17.6	0.40	0.29	3.94	4.33	3.89	0.31	0.22
14	38.7	34.9	0.43	0.35	3.76	3.95	3.77	0.12	0.06
15.a	84.9	–	0.35	–	–	–	–	–	–
15.b	68.9	–	0.29	–	–	–	–	–	–
15.c	72.3	–	0.45	–	–	–	–	–	–
15.d	54.2	–	0.42	–	–	–	–	–	–
15.e	15.1	–	–0.05	–	–	–	–	–	–
15.f	62.6	–	0.30	–	–	–	–	–	–
15.g	25.6	–	0.27	–	–	–	–	–	–
15.h	7.6	–	–0.05	–	–	–	–	–	–
16.a	87.8	–	0.16	–	3.99	–	–	–	–
16.b	44.5	–	0.54	–	3.99	–	–	–	–
16.c	93.3	–	0.14	–	3.99	–	–	–	–
16.d	12.2	3.4	0.07	0.00	3.73	2.63	3.78	–0.78	0.36
16.e	87.8	25.2	0.23	0.16	3.74	3.92	3.76	0.11	0.14
16.f	75.6	–	0.26	–	3.99	–	–	–	–
16.g	52.5	–	0.20	–	3.99	–	–	–	–
16.h	43.7	–	0.25	–	3.99	–	–	–	–
Mean	57.8	23.2	0.30	0.32	4.04	4.27	3.86	0.24	0.17
SD	25.7	12.9	0.19	0.19	0.19	0.57	0.21	0.36	0.11

CFC = Mean confidence of students who answered Correctly to the item. Range: 0.00 (min) – 7.00 (max).

CFW = Mean confidence of students who answered Incorrectly (Wrong) to the item. Range: 0.00 (min) – 7.00 (max).

CAQ = Confidence–Accuracy Quotient [$(CFC - CFW)/SD$] of all confidence ratings for the item].

CB = Confidence Bias [$(CF - 1)/7$ – proportion of students who answered both content and reasoning tiers, in the same question, correctly]. Range: –1.00 (min) – 1.00 (max).

More specifically, the *FI values* are the proportion of students who selected the correct options, so they reflect the difficulty level of each item. The wide variation in FI across questions indicates that the WFDI includes items that were very easy ($FI \geq 70.0\%$), moderately challenging ($70.0\% > FI \geq 30.0\%$) and highly demanding ($FI < 30.0\%$) for participants (Konakci, 2024). Such variation is desirable in a diagnostic instrument, as it ensures that the tool captures both well-understood and poorly understood facets of the WF concept. The mean FI value for content tiers was 58% (SD = 26%), and 23% (SD = 13%) for the combined content and reasoning tiers (both content and reasoning tiers had to be correct in a given question).

DI values were computed by comparing the proportion of correct responses between the highest- and lowest-performing quartiles (25%) of students. Most items demonstrated acceptable discrimination ($DI \geq 0.20$) (Konakci, 2024), indicating that they successfully distinguished between students with stronger and weaker understanding of the overall concept. The mean DI value for the content tiers was 0.30 (± 0.19) and 0.32 (± 0.19) for the combined tiers, i.e. content and reasoning.

The *confidence indices* (CF, CFC, CFW) provide further insight into students' self-perceived certainty when responding to the items. Overall, the mean confidence values (CF) were close to the balance point (mean = 4.04 ± 0.19 out of 7), suggesting that students were generally neither guessing indiscriminately nor expressing overly strong certainty. As expected, the mean confidence of students who answered correctly (CFC values, mean = 4.27 ± 0.57) tended to exceed that of those who answered incorrectly (CFW values, mean = 3.86 ± 0.21), indicating that students who answered correctly were typically more confident in their choices. In contrast, items where CFC and CFW are close suggest reduced metacognitive accuracy, with students having difficulty judging whether their responses were correct or not.

The confidence–accuracy quotient (CAQ index) contextualises confidence differences in relation to the accuracy of answers. High CAQ values indicate strong alignment between confidence and accuracy. Conversely, negative or near-zero CAQ values point to items where confidence does not reliably track accuracy, often the case when misconceptions or conceptual ambiguities are present.

The CB index further captures the extent to which students tended to be overconfident or underconfident. Positive CB values suggest overestimation of performance, e.g. in Q16.d, which is about WF and climate change, the CB is +0.36. This pattern commonly emerges when pervasive misconceptions are prevalent. Negative CB values indicate underconfidence, typically associated with items addressing unfamiliar or conceptually complex ideas. Together, these indices identify areas where students may require additional conceptual or metacognitive support.

Taken together, the item difficulty, discrimination and confidence-related indices present a comprehensive view of item functioning and offer strong evidence of the diagnostic utility of the WFDI. Overall, the instrument demonstrates acceptable psychometric properties for use with undergraduate primary education students.

Students' combined-tier performance. To obtain a more detailed and insightful view of participants' quality of understanding, two additional combined performance scores were

calculated for each item. The first combined score evaluated performance across two tiers (content and reasoning), while the second combined score accounted for performance across all three tiers (content, reasoning and confidence). For these metrics, a student's response to an item was treated as correct only if they provided the right answer to both tiers (two-tier combined score) or all three tiers (three-tier combined score). Following the methodology described above, responses to the confidence tiers were classified as correct if the participant selected a score of 5–7 (Confident to Absolutely Confident) on the 7-point Likert-type scale.

Figure 2 illustrates the descriptive trends for the first nine items of the instrument, which featured only content and confidence tiers. The mean percentage of correct answers was 43% ($\pm 19\%$) when only the content tier was considered, but it dropped significantly to 22% ($\pm 10\%$) when both tiers were required.

For the WFDI items that formed part of a broader, multi-item question (Figure 3), the mean percentage of correct answers for the content tier was 66% ($\pm 26\%$).

For the six three-tier items of the WFDI, a clear step-by-step decline in performance emerged as more layers of evaluation were added (Figure 4). The mean percentage of correct answers was 49% ($\pm 25\%$) when considering only the first tier. This dropped to 24% ($\pm 12\%$) when the first two tiers were evaluated together, and plummeted to just 9% ($\pm 5\%$) when all three tiers were factored in. The proportion of correct answers clearly decreases as more tiers are sequentially included.

Discussion

The aim of this study was to construct and validate a multi-tier diagnostic instrument for the WF concept. Overall, the findings indicate that the WFDI functions acceptably as a diagnostic tool and can assess students' conceptual understanding of the WF. A key point in our results is the clear misalignment between students' accuracy and confidence, especially when considering combined-tier performance. For instance, in the 2-tier items, the mean percentage of correct answers dropped from 44% in the first tier to 22% when both tiers were

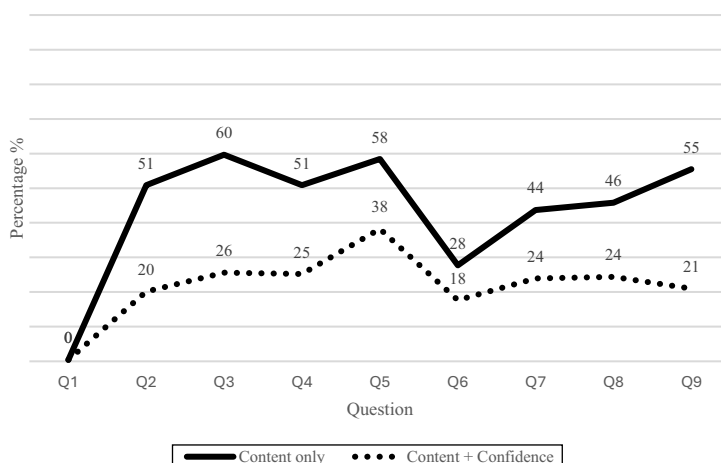


Figure 2. Percentage of correct responses per tier for the 2-tier items

Source: Authors' own work

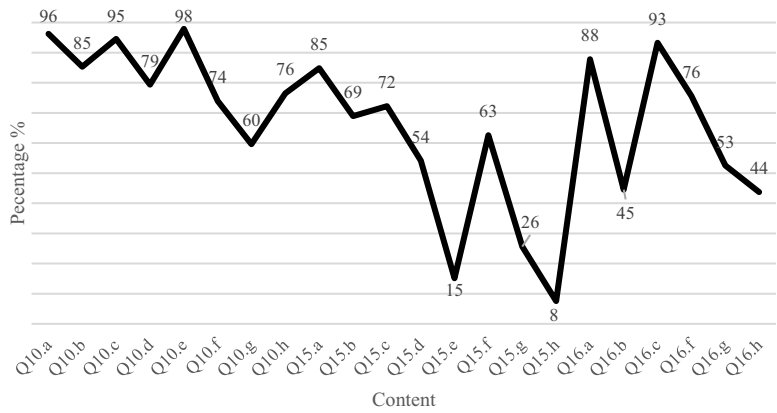


Figure 3. Percentage of correct responses for the multi-item 1-tier questions
Source: Authors' own work

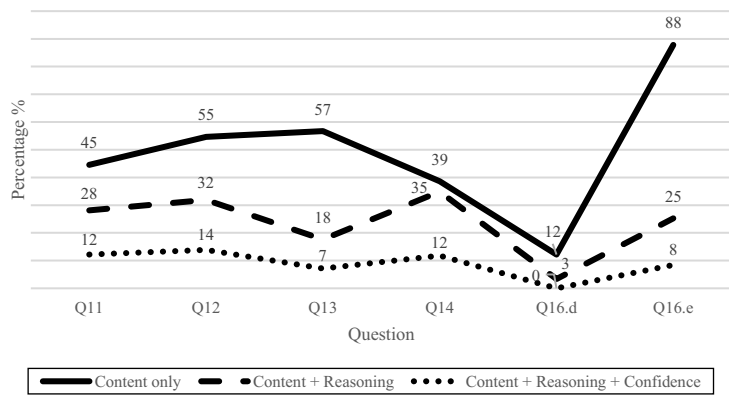


Figure 4. Percentage of correct responses per tier for the 3-tier items
Source: Authors' own work

considered, while at the 3-tier items the decline was even more dramatic, dropping from 49% to just 9% when all tiers were combined, a pattern consistent with multi-tier diagnostic studies (Ma *et al.*, 2025). This sharp decrease carries significant educational meaning and can be interpreted through the lens of the conceptual change literature, which is dominated by the debate between coherence (Vosniadou, 1994, 2013) and fragmentation (diSessa, 1993, 2014). The fact that only 9% of students could support a correct answer with proper reasoning and high confidence proves that traditional assessment methods, which mostly focused on content knowledge, overestimate student literacy.

This fragmentation of student knowledge is evident at two distinct levels in our study. At the intra-item level, steep performance drops within the same question indicate that students often hold an isolated piece of correct information that is completely disconnected from any valid scientific reasoning. At the same time, cross-item fragmentation was observed in the

large fluctuations in performance across items, such as the drop from 83% in Q10 to 0.4% in Q1. When students systematically selected specific incorrect answers with high confidence, they demonstrated robust alternative frameworks (Vosniadou, 2013). However, the widespread mixing of correct answers with flawed justifications across the instrument fits diSessa's (2014) view that pre-instructional knowledge is largely a collection of isolated, context-sensitive elements ("knowledge in pieces") rather than a unified theory.

The low percentage of students who selected the correct answers (FI values, < 30%) in four questions indicates that these items were difficult for most students and likely reflect underlying conceptual weaknesses or misconceptions, while the high FI values (>70%) in 11 questions suggest that the respective concepts were relatively well understood. Some items also produced low or negative Item Discrimination Index (DI values), indicating that the items do not effectively distinguish between students with stronger and weaker overall understanding of WF concepts. This may suggest that the question is unclear, too easy or too difficult, or that the scientifically correct option/reasoning is not sufficiently distinct from common misconceptions. A negative DI is especially problematic, as it means that lower-achieving students answered the item correctly more often than higher-achieving students, indicating that the item may need revision or removal. However, while a low DI is typically viewed as a structural flaw in traditional testing, within multi-tier diagnostic assessments designed to uncover alternative conceptions, this pattern is not unexpected (Arslan *et al.*, 2012; Liampa *et al.*, 2019). Therefore, these items do not weaken the instrument; rather, they serve their diagnostic purpose by highlighting specific cognitive obstacles that require more systematic instructional attention.

The students showed much lower means when both the content and reasoning tiers were considered. This high-content, low-reasoning pattern is typical for multi-tier diagnostic instruments and indicates surface-level rather than deep conceptual understanding. Liampa *et al.* (2019) reported similar content performance but considerably higher reasoning and confidence means, which may support the idea that the ecological footprint is more familiar and widely taught than the WF. In addition, reasoning about WF may be inherently more complex, as it involves tracking hidden, indirect resource loops and invisible processes such as virtual water flows (Hoekstra *et al.*, 2011).

The most challenging item was estimating the mean Greek WF magnitude (Q1), with only 0.4% of students selecting the correct answer. This is not surprising, as the WF of an average Greek (around 6,500 L/day according to Mekonnen and Hoekstra, 2011) is counterintuitive. Most students (74%) estimated values between 0 and 100 L/day, probably focusing on direct (domestic) water use rather than indirect consumption, even though the question explicitly explained the difference. These results are comparable to, but even lower than those of Barreiros *et al.* (2024), where only 18% of respondents reported an accurate consumption range. Findings from the tool's development phases reveal two main trends in students' perspectives that could explain this misunderstanding. The first trend was a strictly anthropocentric perspective grounded in direct experience, in which students consider personal consumption to be limited to what physically flows from their household taps. In contrast, a divergent trend emerged from students who acknowledged that product manufacturing requires water, but completely underestimated the industrial scale, viewing freshwater as an infinite, rapidly self-renewing resource. This qualitative contrast indicates that the 0–100 L/day students' estimation is probably not a random guess, but a structurally anchored misconception driven by the invisibility of virtual water (Hoekstra *et al.*, 2011), the pervasive societal beliefs and media communication that narrow water conservation strictly to domestic saving practices and miss to shed light to the indirect use which is the major part (Elmaslar Özbaş *et al.*, 2021).

In contrast, the determinants of WF calculation (Q10.a-h) were the easiest items for students, with a mean correct percentage of 83%. These items appear more intuitive, as they relate to everyday experiences. For instance, most students (FI = 98%) correctly identify the domestic water consumption (Q10.e) as a determinant of WF. This suggests that when WF concepts are assimilated concretely and linked to familiar everyday behaviours, students are more accurate and confident. This sharp contrast between abstract macro-metrics (virtual water volumes) and concrete micro-actions (domestic tap use) is highly consistent with [Vermehren et al. \(2025\)](#) findings on experiential grounding in science education. Students seem to achieve conceptual clarity mainly when a concept overlaps with their immediate lived experiences, leaving them vulnerable to fallacies when they deal with complex and global-scale topics.

Overall, the findings suggest that the WFDI is an acceptable and useful tool for assessing conceptual understanding of WF. In line with previous multi-tier diagnostic research (e.g. [Arslan et al., 2012](#); [Liampa et al., 2019](#)), the instrument displays adequate internal consistency and provides meaningful information about item functioning through FI, DI and confidence-related indices. As such, it serves as a valuable diagnostic tool for uncovering the underlying structure of student ideas, providing the empirical basis needed to synthesise how these alternative conceptions are formed.

Based on this conceptual synthesis, we propose two distinct cognitive mechanisms that describe how pre-service teachers process consumption-based sustainability metrics:

- (1) The Experiential Proximity Mechanism: This mechanism dictates that environmental literacy is high and metacognitive confidence is stable when the metric directly mirrors visible, daily personal routines (e.g. domestic water use in Q10).
- (2) The Industrial Hidden Barrier: This mechanism operates when a concept relies on hidden, macro-level supply chains (e.g. virtual water in food and product manufacturing in Q1). In these cases, students' cognitive frameworks fail to scale their intuitive ideas, creating a robust mental barrier that resists traditional, non-explicit sustainability instruction.

Implications

The findings indicate that undergraduate primary education students have only a partial and often fragmented understanding of the WF concept, especially when reasoning is required. However, these results reveal clear entry points for instructional planning. WF should be taught more explicitly in sustainability-related courses, with a focus on conceptual understanding and concrete examples that clarify the indirect link between everyday product consumption and freshwater use. Given their future professional roles, primary education students act as key conceptual gatekeepers. If their own misconceptions about hidden water resource consumption persist, they will inevitably instil similar flawed frameworks in their future students, limiting the development of broader awareness and understanding. Therefore, teacher training programs should move beyond generic environmental awareness and target specific sustainability concepts. Instructional interventions should explicitly address food and animal product consumption – which carry the heaviest virtual water weights – forcing future educators to confront the trade-offs between personal dietary patterns and global water resource use.

In addition, the results suggest a need for more systematic inclusion of WF in Greek school and university curricula. As WF is a multifaceted concept, instruction should address its components, determinants and implications, ideally supported by multi-tier diagnostic

approaches that can reveal alternative conceptions. In this regard, the WFDI provides a validated tool for identifying areas of weak conceptual understanding. To maximise its utility, the instrument features a structured framework that maps different students' profiles, while its format requires minimal educator training. However, for universities looking to implement this diagnostic approach, there is a practical trade-off to consider: while multi-tier instruments offer much greater analytical depth than standard multiple-choice tests, they require more time and effort to administer and evaluate. To balance this and guide educational decisions, institutions can focus on tracking two clear aspects: how effectively students' confident-but-mistaken beliefs decline after an instructional intervention, and what percentage of students develop a fully consistent, correct understanding across both their content choices and justifications. Furthermore, the size of each institution creates different conditions for application. Large university Departments with high student enrolment can distribute the test digitally through online learning platforms using automated scoring to easily manage the volume of data. In contrast, smaller Departments can leverage the tool more deeply by using the detailed student feedback directly in small-group reflective seminars to refine their day-to-day teaching practices.

Conclusions

This study developed and validated the WFDI, a multi-tier diagnostic tool specifically designed to assess pre-service primary teachers' understanding of the WF concept. The instrument was based on a robust conceptual framework, informed by extensive qualitative exploration of students' ideas and refined through expert review. Psychometric analyses indicate acceptable reliability and meaningful item functioning, supporting its use as a diagnostic tool.

Overall, the WFDI addresses an important gap in the literature by providing a structured method for assessing conceptual understanding of WF within teacher education. By moving beyond descriptive psychometrics, the WFDI proves that future teachers' primary vulnerability lies not in a raw absence of facts, but in the structural fragmentation of their systemic thinking. Addressing these specific cognitive mechanisms provides a clear, data-driven pathway for HEIs to reform sustainability curricula, bridge educational theory with classroom practice and ensure that future educators can confidently foster genuine environmental literacy.

Limitations and directions for future research

The study has several limitations that could be addressed in future research. First, the participants were exclusively Greek undergraduate primary education students from specific universities, which limits the immediate generalisability of our baseline performance findings to other student populations or different national education systems. Second, while the instrument's structural validity and reliability were thoroughly established, the reliance on Cronbach's alpha as the primary reliability metric presents known limitations in capturing the multi-dimensional nature of multi-tier assessments. Future studies should employ Item Response Theory (IRT) or Rasch modelling to overcome these limitations and further confirm the invariance of item parameters across a wider and more diverse range of students' abilities.

Another limitation is the absence of test-retest reliability; however, this design was intentionally omitted to avoid testing effects, where administering the exact same instrument twice within a short window can artificially alter students' genuine baseline responses.

Future research should test the instrument in other countries to examine its psychometric properties in various national and cultural settings. Moreover, future interventions should

adopt longitudinal pre/post research designs with participants who receive systematic instruction on the WF concept to evaluate the instrument's sensitivity in tracking genuine learning gains over time. Such efforts would also help determine whether correcting these alternative conceptions during teacher training leads to more effective and confident sustainability teaching practices in primary schools. To systematically guide these initiatives, future development will follow a structured, three-stage roadmap: short-term cross-cultural translation and validation, medium-term usability testing with university instructors and long-term longitudinal tracking to monitor how these alternative conceptions evolve over multiple academic years. Furthermore, while the WFDI is strictly domain-specific, its underlying multi-tier methodological framework can serve as a transferable blueprint for researchers looking to design similar diagnostic instruments for other complex sustainability metrics, such as carbon footprint.

References

- Adadan, E. and Savasci, F. (2012), "An analysis of 16–17-year-old students' understanding of solution chemistry concepts using a two-tier diagnostic instrument", *International Journal of Science Education*, Vol. 34 No. 4, pp. 513-544, doi: [10.1080/09500693.2011.636084](https://doi.org/10.1080/09500693.2011.636084).
- Akhtar, N., Syakir Ishak, M.I., Bhawani, S.A. and Umar, K. (2021), "Various natural and anthropogenic factors responsible for water quality degradation: a review", *Water*, Vol. 13 No. 19, p. 2660, doi: [10.3390/w13192660](https://doi.org/10.3390/w13192660).
- Aksoy, A.A. and Erten, S. (2022), "A four-tier diagnostic test to determine pre-service science teachers' misconception about global warming", *Journal of Baltic Science Education*, Vol. 21 No. 5, pp. 747-761.
- Aleksandrowicz, L., Green, R., Joy, E.J.M., Smith, P. and Haines, A. (2016), "The impacts of dietary change on greenhouse gas emissions, land use, water use, and health: a systematic review", *PLoS One*, Vol. 11 No. 11, p. e0165797, doi: [10.1371/journal.pone.0165797](https://doi.org/10.1371/journal.pone.0165797).
- Allan, J.A. (1998), "Virtual water: a strategic resource", *Ground Water*, Vol. 36 No. 4, p. 545+.
- Arslan, H.O., Cigdemoglu, C. and Moseley, C. (2012), "A Three-Tier diagnostic test to assess Pre-Service teachers' misconceptions about global warming, greenhouse effect, ozone layer depletion, and acid rain", *International Journal of Science Education*, Vol. 34 No. 11, pp. 1667-1686, doi: [10.1080/09500693.2012.680618](https://doi.org/10.1080/09500693.2012.680618).
- Barreiros, A.M., Durão, A., Galvão, A., Matos, C., Mateus, D., Araújo, I., Neves, L. and Mourato, S. (2024), "Higher education institutions' students' literacy in sustainable use of potable water", *Sustainability*, Vol. 16 No. 12, p. 5217, doi: [10.3390/su16125217](https://doi.org/10.3390/su16125217).
- Barth, M., Godemann, J., Rieckmann, M. and Stoltenberg, U. (2007), "Developing key competencies for sustainable development in higher education", Edited by Adomssent, M. *International Journal of Sustainability in Higher Education*, Vol. 8 No. 4, pp. 416-430, doi: [10.1108/14676370710823582](https://doi.org/10.1108/14676370710823582).
- Caleon, I. and Subramaniam, R. (2010), "Development and application of a three-tier diagnostic test to assess secondary students' understanding of waves", *International Journal of Science Education*, Vol. 32 No. 7, pp. 939-961, doi: [10.1080/09500690902890130](https://doi.org/10.1080/09500690902890130).
- Çamur, D., Konyalioglu, F.S., Ketrez, G., Güneş, İ.S. and Hasde, M. (2020), "Bir üniversitenin bazı fakültelerinde okuyan öğrencilerin su tüketimi konusundaki bilgi, tutum ve davranışları", *Turkish Bulletin of Hygiene and Experimental Biology/Türk Hijyen ve Deneysel Biyoloji*, Vol. 77.
- Cebrián, G. and Junyent, M. (2015), "Competencies in education for sustainable development: exploring the student teachers' views", *Sustainability*, Vol. 7 No. 3, pp. 2768-2786, doi: [10.3390/su7032768](https://doi.org/10.3390/su7032768).
- Chaney, P.L. and Doukopoulos, L. (2018), "An active learning exercise on sustainability and the water footprint of food: the dinner party menu challenge", *The Geography Teacher*, Vol. 15 No. 4, pp. 173-184, doi: [10.1080/19338341.2018.1491877](https://doi.org/10.1080/19338341.2018.1491877).

- Chapagain, A.K. and Orr, S. (2008), "UK water footprints: the impact of the UK's food and fibre consumption on global water resources", *Wwf-UK*, Godalming, UK.
- Cole, L.B., Priscilla, L., Zangori, L., Kania-Gosche, B. and Burken, J. (2024), "Raising the green roof: enhancing youth water literacy through built environment education", *Sustainability*, Vol. 16 No. 10, p. 4262, doi: [10.3390/su16104262](https://doi.org/10.3390/su16104262).
- Dal-Farra, R.A., da Costa, J.O., de Souza Proença, M., Assumpção, R. and Lopes, L.A. (2015), "Reflection: water footprint calculator, education and systemic thought with high school Brazilian students", *6th International Multi-Conference on Complexity, Informatics and Cybernetics, IMCIC 2015 and 6th International Conference on Society and Information Technologies, ICSIT 2015, Orlando, United States*, Vol. 2, pp. 140-143.
- diSessa, A.A. (1993), "Toward an epistemology of physics", *Cognition and Instruction*, Vol. 10 Nos 2-3, pp. 105-225.
- diSessa, A.A. (2014), "A history of conceptual change research: threads and fault lines", in Sawyer, R.K. (Ed.), *The Cambridge Handbook of the Learning Sciences*, 2nd ed., Cambridge University Press, New York, pp. 88-108, doi: [10.1017/CBO9781139519526.007](https://doi.org/10.1017/CBO9781139519526.007).
- diSessa, A.A., Gillespie, N.M. and Esterly, J.B. (2004), "Coherence versus fragmentation in the development of the concept of force", *Cognitive Science*, Vol. 28 No. 6, pp. 843-900, doi: [10.1207/s15516709cog2806_1](https://doi.org/10.1207/s15516709cog2806_1).
- Dota, A., Theodossiou, N. and Dots, V. (2016), "Pressures on water resources for sheep and goats breeding in the regional unit of Karditsa employing the water footprint approach", *European Water*, Vol. 55, pp. 21-30.
- Duit, R. and Treagust, D.F. (2003), "Conceptual change: a powerful framework for improving science teaching and learning", *International Journal of Science Education*, Vol. 25 No. 6, pp. 671-688, doi: [10.1080/09500690305016](https://doi.org/10.1080/09500690305016).
- Elmaslar Özbaşı, E., Sütirmak Pehlivan, T., Özcan, H.K., Öngen, A., Aydıñ, S. and Güneysu, S. (2021), "Distribution of water footprint components of university students and detecting the factors that affect those components", *International Journal of Agriculture Environment and Food Sciences*, Vol. 5 No. 4, pp. 456-463, doi: [10.31015/jaefs.2021.4.4](https://doi.org/10.31015/jaefs.2021.4.4).
- Eskici, E. (2025), "A theoretical framework for developing water footprint awareness at primary school level", *Zenodo*, doi: [10.5281/ZENODO.17473271](https://doi.org/10.5281/ZENODO.17473271).
- Espinosa, A.A., Koperová, D., Kuhnová, M. and Rusek, M. (2025), "Preservice chemistry teachers' conceptual understanding and confidence judgment: insights from a Three-Tier chemistry concept inventory", *Journal of Chemical Education*, Vol. 102 No. 1, pp. 53-65, doi: [10.1021/acs.jchemed.4c01146](https://doi.org/10.1021/acs.jchemed.4c01146).
- Ferreira da Luz Culpí, V.L. and Pereira Alves, J.A. (2015), "Introduction of the water footprint theme in science education: perceptions and perspectives of changes from the classroom", *Tecné, Episteme y Didaxis: TED*, No. 38, pp. 17-36.
- Garbayo, L.S., Harris, D.M., Fiore, S.M., Robinson, M. and Kibble, J.D. (2023), "A metacognitive confidence calibration (MCC) tool to help medical students scaffold diagnostic reasoning in decision-making during high-fidelity patient simulations", *Advances in Physiology Education*, Vol. 47 No. 1, pp. 71-81, doi: [10.1152/advan.00156.2021](https://doi.org/10.1152/advan.00156.2021).
- Gerbens-Leenes, W. and Hoekstra, A.Y. (2011), "The water footprint of biofuel-based transport", *Energy & Environmental Science*, Vol. 4 No. 8, p. 2658, doi: [10.1039/c1ee01187a](https://doi.org/10.1039/c1ee01187a).
- Gkitsas, S., Malandrakis, G., Theodosiou, N. and Lambrinos, N. (2024), "Undergraduate primary school education students' alternative ideas about the water footprint", In Kotsis, K. T., Stylos, G., Vakarou, G., Gavrilas, L., and Panagou, D. (Eds), *Proceedings of the 13th Panhellenic Conference on Science Education and New Technologies in Education: New trends and research in learning, teaching and technologies in science education*, University of Ioannina, Ioannina, Greece, pp. 816-821, doi: [10.12681/codiste.6823](https://doi.org/10.12681/codiste.6823).
- Gleick, P.H. and Cooley, H. (2021), "Freshwater scarcity", *Annual Review of Environment and Resources*, Vol. 46 No. 1, pp. 319-348, doi: [10.1146/annurev-environ-012220-101319](https://doi.org/10.1146/annurev-environ-012220-101319).

- Gobert, J.D. and Buckley, B.C. (2000), "Introduction to model-based teaching and learning in science education", *International Journal of Science Education*, Vol. 22 No. 9, pp. 891-894, doi: [10.1080/095006900416839](https://doi.org/10.1080/095006900416839).
- Godfrey, D.M. and Feng, P. (2017), "Communicating sustainability: student perceptions of a behavior change campaign", *International Journal of Sustainability in Higher Education*, Vol. 18 No. 1, pp. 2-22, doi: [10.1108/IJSHE-01-2015-0009](https://doi.org/10.1108/IJSHE-01-2015-0009).
- Haladyna, T.M. and Rodriguez, M.C. (2013), *Developing and Validating Test Items*, Routledge, New York.
- Hoekstra, A.Y. (2003), "'Virtual water trade', presented at the proceedings of the international expert meeting on virtual water trade, delft, The Netherlands", *Value of Water Research Report Series No.*, pp. 12-13.
- Hoekstra, A.Y., Chapagain, A.K., Aldaya, M.M. and Mekonnen, M.M. (2011), *The Water Footprint Assessment Manual: Setting the Global Standard*, Earthscan, London, UK.
- Holst, J., Fritz, H., Nölting, B., Lang, D.J., Löde, V.-L., Ober, S., Parodi, O., Potthast, T., Scheiding, C.T., Weynand, M. and Singer-Brodowski, M. (2025), "Exploring culture(s) of sustainability at higher education institutions: a systematic review of concepts and pathways", *International Journal of Sustainability in Higher Education*, Vol. 26 No. 9, pp. 489-506, doi: [10.1108/IJSHE-08-2024-0550](https://doi.org/10.1108/IJSHE-08-2024-0550).
- Horne, L., Manzanares, A., Atalan-Helicke, N., Vincent, S., Anderson, S.W. and Romulo, C.L. (2025), "An exploratory study of drawings as a tool to evaluate student understanding of the Food-Energy-Water (FEW) nexus", *Journal of Environmental Studies and Sciences*, Vol. 15 No. 2, pp. 235-249, doi: [10.1007/s13412-024-00929-x](https://doi.org/10.1007/s13412-024-00929-x).
- Institute of Educational Policy (IEP) (2021), "Subject uniform curriculum framework for environmental studies (in Greek)".
- Irmak, M., İNaltun, H., Ercan-Dursun, J., Yanış-Kelleci, H. and Yürük, N. (2023), "Development and application of a Three-Tier diagnostic test to assess pre-service science teachers' understanding on Work-Power and energy concepts", *International Journal of Science and Mathematics Education*, Vol. 21 No. 1, pp. 159-185, doi: [10.1007/s10763-021-10242-6](https://doi.org/10.1007/s10763-021-10242-6).
- Jalava, M., Kumm, M., Porkka, M., Siebert, S. and Varis, O. (2014), "Diet change-a solution to reduce water use?", *Environmental Research Letters*, Vol. 9 No. 7, p. 074016, doi: [10.1088/1748-9326/9/7/074016](https://doi.org/10.1088/1748-9326/9/7/074016).
- Kılıç, Z. (2020), "The importance of water and conscious use of water", *International Journal of Hydrology*, Vol. 4 No. 5, pp. 239-241.
- Konakçı, S. (2024), "Item analysis in multiple choice questions: a study on question difficulty and authors' evaluation", *Journal of Basic and Clinical Health Sciences*, Vol. 8 No. 2, pp. 490-497, doi: [10.30621/jbachs.1427464](https://doi.org/10.30621/jbachs.1427464).
- Krippendorff, K. (2013), *Content Analysis: An Introduction to Its Methodology*, Third edition. Sage, Los Angeles London New Delhi Singapore.
- Landis, J.R. and Koch, G.G. (1977), "An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers", *Biometrics*, Vol. 33 No. 2, pp. 363-374.
- Leal Filho, W., Frankenberger, F., Salvia, A.L., Azeiteiro, U., Alves, F., Castro, P., Will, M., Platje, J., Lovren, V.O., Brandli, L., Price, E., Doni, F., Mifsud, M. and Avila, L.V. (2021), "A framework for the implementation of the sustainable development goals in university programmes", *Journal of Cleaner Production*, Vol. 299, p. 126915, doi: [10.1016/j.jclepro.2021.126915](https://doi.org/10.1016/j.jclepro.2021.126915).
- Liampa, V., Malandrakis, G.N., Papadopoulou, P. and Pnevmatikos, D. (2019), "Development and evaluation of a Three-Tier diagnostic test to assess undergraduate primary teachers' understanding of ecological footprint", *Research in Science Education*, Vol. 49 No. 3, pp. 711-736, doi: [10.1007/s11165-017-9643-1](https://doi.org/10.1007/s11165-017-9643-1).
- Liu, J., Zhao, D., Mao, G., Cui, W., Chen, H. and Yang, H. (2020), "Environmental sustainability of water footprint in Mainland China", *Geography and Sustainability*, Vol. 1 No. 1, pp. 8-17, doi: [10.1016/j.geosus.2020.02.002](https://doi.org/10.1016/j.geosus.2020.02.002).

- Lovarelli, D., Bacenetti, J. and Fiala, M. (2016), "Water footprint of crop productions: a review", *Science of The Total Environment*, Vols 548-549, pp. 236-251, doi: [10.1016/j.scitotenv.2016.01.022](https://doi.org/10.1016/j.scitotenv.2016.01.022).
- Lozano, R., Merrill, M., Sammalisto, K., Ceulemans, K. and Lozano, F. (2017), "Connecting competences and pedagogical approaches for sustainable development in higher education: a literature review and framework proposal", *Sustainability*, Vol. 9 No. 10, p. 1889, doi: [10.3390/su9101889](https://doi.org/10.3390/su9101889).
- Ma, H., Yang, H., Li, C., Ma, S. and Li, G. (2025), "The effectiveness and sustainability of tier diagnostic technologies for misconception detection in science education: a systematic review", *Sustainability*, Vol. 17 No. 7, p. 3145, doi: [10.3390/su17073145](https://doi.org/10.3390/su17073145).
- Martins, F., Felgueiras, C., Smitkova, M. and Caetano, N. (2019), "Analysis of fossil fuel energy consumption and environmental impacts in European countries", *Energies*, Vol. 12 No. 6, p. 964, doi: [10.3390/en12060964](https://doi.org/10.3390/en12060964).
- Mekonnen, M.M. and Hoekstra, A.Y. (2011), "National water footprint accounts: the green, blue and grey water footprint of production and consumption. Volume 1: main report".
- Okadera, T., Chontanawat, J. and Gheewala, S.H. (2014), "Water footprint for energy production and supply in Thailand", *Energy*, Vol. 77, pp. 49-56, doi: [10.1016/j.energy.2014.03.113](https://doi.org/10.1016/j.energy.2014.03.113).
- Okutan, P. and Akkoyunlu, A. (2021), "Identification of water use behavior and calculation of water footprint: a case study", *Applied Water Science*, Vol. 11 No. 7, p. 127, doi: [10.1007/s13201-021-01459-5](https://doi.org/10.1007/s13201-021-01459-5).
- Owens, D.C., Pettit, D.N., Lally, D. and Forbes, C.T. (2020), "Cultivating water literacy in STEM education: undergraduates' socio-scientific reasoning about socio-hydrologic issues", *Water*, Vol. 12 No. 10, p. 2857, doi: [10.3390/w12102857](https://doi.org/10.3390/w12102857).
- Özdemir, G. and Clark, D.B. (2007), "An overview of conceptual change theories", *EURASIA Journal of Mathematics, Science and Technology Education*, Vol. 3 No. 4, doi: [10.12973/ejmste/75414](https://doi.org/10.12973/ejmste/75414).
- Panatsa, V.M. and Malandrakis, G. (2024), "Greek primary school students' moral judgments and motives about sustainable food consumption", *Cleaner and Responsible Consumption*, Vol. 12, p. 100173, doi: [10.1016/j.clrc.2024.100173](https://doi.org/10.1016/j.clrc.2024.100173).
- Pozo-Muñoz, M.P., Martín-Gómez, C., Velasco-Martínez, L.C. and Tójar-Hurtado, J.C. (2023), "Research and development of environmental awareness about water in primary education students through their drawings", *Education Sciences*, Vol. 13 No. 2, p. 119, doi: [10.3390/educsci13020119](https://doi.org/10.3390/educsci13020119).
- Putica, K.B. (2023), "Development and validation of a Four-Tier test for the assessment of secondary school students' conceptual understanding of amino acids, proteins, and enzymes", *Research in Science Education*, Vol. 53 No. 3, pp. 651-668, doi: [10.1007/s11165-022-10075-5](https://doi.org/10.1007/s11165-022-10075-5).
- Rieckmann, M. (2012), "Future-oriented higher education: which key competencies should be fostered through university teaching and learning?", *Futures*, Vol. 44 No. 2, pp. 127-135, doi: [10.1016/j.futures.2011.09.005](https://doi.org/10.1016/j.futures.2011.09.005).
- Soe, K.T. and Yeo-Chang, Y. (2019), "Livelihood dependency on Non-Timber forest products: implications for REDD+", *Forests*, Vol. 10 No. 5, p. 427, doi: [10.3390/f10050427](https://doi.org/10.3390/f10050427).
- Tamir, P. (1971), "An alternative approach to the construction of multiple choice test items", *Journal of Biological Education*, Vol. 5 No. 6, pp. 305-307, doi: [10.1080/00219266.1971.9653728](https://doi.org/10.1080/00219266.1971.9653728).
- Treagust, D.F. (1988), "Development and use of diagnostic tests to evaluate students' misconceptions in science", *International Journal of Science Education*, Vol. 10 No. 2, pp. 159-169, doi: [10.1080/0950069880100204](https://doi.org/10.1080/0950069880100204).
- Treagust, D.F. (1995), "Diagnostic assessment of students' science knowledge", in Glynn, S.M. and Duit, R. (Eds.), *Learning in the Schools: Research Reforming Practice*, Vol. 1 Lawrence Erlbaum, Mahwah, NJ, pp. 327-346.
- Treagust, D.F. and Duit, R. (2008), "Conceptual change: a discussion of theoretical, methodological and practical challenges for science education", *Cultural Studies of Science Education*, Vol. 3 No. 2, pp. 297-328, doi: [10.1007/s11422-008-9090-4](https://doi.org/10.1007/s11422-008-9090-4).
- United Nations (2015), "Transforming our world: the 2030 agenda for sustainable development", available at: <https://sdgs.un.org/2030agenda>

- United Nations (1987), "World commission on environment and development", *Our Common Future*.
- Vanham, D., Mekonnen, M.M. and Hoekstra, A. Y. (2013), "The water footprint of the EU for different diets", *Ecological Indicators*, Vol. 32, pp. 1-8, doi: [10.1016/j.ecolind.2013.02.020](https://doi.org/10.1016/j.ecolind.2013.02.020).
- Venckute, M., Silva, M. and Figueiredo, M. (2017), "Education as a tool to reduce the water footprint of young people", *Millenium - Journal of Education, Technologies, and Health*, No. 04, pp. 101-111.
- Vermehren, J.A.V., Trikoili, A. and Pittich, D. (2025), "Abstract thought in STEM education: an integrative literature review", *International Journal of STEM Education*, Vol. 12 No. 1, p. 54, doi: [10.1186/s40594-025-00573-z](https://doi.org/10.1186/s40594-025-00573-z).
- Vosniadou, S. (1994), "Capturing and modeling the process of conceptual change", *Learning and Instruction*, Vol. 4 No. 1, pp. 45-69, doi: [10.1016/0959-4752\(94\)90018-3](https://doi.org/10.1016/0959-4752(94)90018-3).
- Vosniadou, S. (2013), "Conceptual change in learning and instruction: the framework theory approach", *International Handbook of Research on Conceptual Change*, Routledge, New York, pp. 11-30.
- Vosniadou, S. and Brewer, W.F. (1992), "Mental models of the earth: a study of conceptual change in childhood", *Cognitive Psychology*, Vol. 24 No. 4, pp. 535-585, doi: [10.1016/0010-0285\(92\)90018-W](https://doi.org/10.1016/0010-0285(92)90018-W).
- Wiek, A., Withycombe, L. and Redman, C.L. (2011), "Key competencies in sustainability: a reference framework for academic program development", *Sustainability Science*, Vol. 6 No. 2, pp. 203-218, doi: [10.1007/s11625-011-0132-6](https://doi.org/10.1007/s11625-011-0132-6).
- World Water Assessment Programme (WWAP) (2009), *The United Nations World Water Developing Report 3: Water in a Changing World*, WWAP, UNESCO Publishing, Paris and Earthscan, London, Earthscan.
- Zhou, X., Zhai, Y., Zhang, T., Li, Z., Cheng, Z., Li, C., Xu, T. and Hong, J. (2023), "Uncovering the energycarbon-water footprint of waste rubber recycling: integrated environmental and economic perspectives", *Journal of Environmental Management*, Vol. 345, p. 118916, doi: [10.1016/j.jenvman.2023.118916](https://doi.org/10.1016/j.jenvman.2023.118916).
- Zucchinelli, M., Sporchia, F., Piva, M., Thomsen, M., Lamastra, L. and Caro, D. (2021), "Effects of different Danish food consumption patterns on Water ScarcityFootprint", *Journal of Environmental Management*, Vol. 300, p. 113713, doi: [10.1016/j.jenvman.2021.113713](https://doi.org/10.1016/j.jenvman.2021.113713).

Appendix. The instrument

The full version of the Water Footprint Diagnostic Instrument (WFDI) is available from the corresponding author upon request.

Corresponding author

Stergios Gkitsas can be contacted at: sggkitsas@eled.auth.gr