Ed Process Int J

www.edupij.com

ISSN 2147-0901 (Print), 2564-8020 (Online)

EDUCATIONAL PROCESS INTERNATIONAL JOURNAL

Research Article

Cite this article: Habeeb, S. T., & Ali, B. I. (2025). Developing a Criterion-Referenced Test in Environmental Education Using Amplified Objectives: A Rasch (1PL) Analysis. *Educational Process: International Journal, 19*, e2025579.
https://doi.org/10.22521/edupij.2025.19.579

Received July 27, 2025
Accepted November 30, 2025

Keywords: Criterion-referenced assessment; amplified objectives; Rasch model; item response theory; environmental education; standard setting

Author for correspondence:
Safaa Tariq Habeeb
safa.tairq@ircoedu.uobaghdad.edu.iq
University of Baghdad

Developing a Criterion-Referenced Test in Environmental Education Using Amplified Objectives: A Rasch (1PL) Analysis

Safaa Tariq Habeeb, Burhan Ismael Ali

Abstract

Background/purpose. The current research aims to construct a criterion-referenced test using the amplified objectives for the Environmental Education subject for students of the Educational and Psychological Sciences Departments in the Faculties of Education, based on item response theory (IRT) and the Rasch one-parameter logistic (1PL) model. The academic content was derived from five chapters developed by the Deans' Council of the Ministry of Higher Education and Scientific Research and reviewed by subject instructors. From this, seventy behavioral objectives were formulated, distributed among Bloom's taxonomy levels: remembering (35), understanding (29), and application (8).

Materials/methods. Seventy multiple-choice items were developed and evaluated by specialized experts for validity. The research sample included 500 male and female second-stage students from the Universities of Baghdad, Al-Mustansiriya, Al-Iraqiya, Karbala, Anbar, and Diyala for the academic year 2024–2025. Students were selected using a stratified random method with proportional distribution. A pilot test was conducted with 30 students to ensure clarity of test items and instructions; all were found to be clear.

Results. The cut-off score was established using the amplified objectives and the contrasting groups method under item response theory (IRT), resulting in a score of 47 (65.28%). Statistical analysis using SPSS and the ICL program confirmed that all items had strong psychometric properties. The results indicated a decrease in the number of proficient students in Environmental Education.

Conclusion. The test developed in this study has strong psychometric properties and effectively measures student proficiency. The decline in performance suggests a need for further educational and assessment interventions. These calibrated items, aligned to amplified objectives, support diagnostic decisions about mastery and can guide remediation within environmental education courses.

1. Introduction

1.1. Rationale of the Study

Criterion-referenced tests are the most appropriate type of achievement tests for measuring and evaluating student achievement because they are based on identifying the skills and competencies that must be mastered with great precision so that the teacher can measure and observe them directly, and then estimate the extent to which the student has achieved those goals, based on a specific level of performance. This helps the teacher in the diagnosis process, as students are classified into two categories: those who have mastered the specified skills and competencies and those who have not. This is followed by identifying the student's strengths and weaknesses in their level of achievement, as well as the skills they have mastered and those they have not. Thus, the teacher can develop appropriate remedial programs for this purpose. (Alam, 1986, pp.14-15).

A criterion-referenced test denotes an assessment interpreted against a fixed performance standard rather than norm groups. Amplified objectives are behaviorally stated objectives expanded to name the learner, the content, the intended behavior, the performance standard, and the conditions of performance; they are used here to drive content specification and standard setting.

1.2. Summary of Previous Studies

Proponents of the contemporary approach to measurement have sharply criticized the traditional theory of psychological and educational measurement and the methods based on it. Among the most important of these criticisms is the lack of independence of measurement results from the sample to which the test was administered and from the items included in the test. Traditional behavioral objectives provide a direct and effective way to measure student performance on specific tasks. While amplified objectives help direct education toward long-term outcomes, each type of objective has its own uses, and it is best to combine them to ensure balanced educational outcomes. However, a review of all previous local and Arab studies reveals that they rely solely on behavioral objectives in constructing the test and do not use amplified objectives. This may be due to the scarcity of literature that referred to the amplified objectives on the one hand, and the scarcity of specialized studies that discussed the extent of the effectiveness of the amplified objectives in constructing achievement tests, in addition to the fact that researchers are not accustomed to resorting to such objectives in constructing the test. Objectives are an essential part of the educational evaluation and systematic planning processes, and are divided into traditional behavioral objectives and broad objectives.

Research in recent years suggests that one source of differences between high and low achievers is the degree to which a person becomes self-regulated, with high achievers engaging in grandiose goal setting, planning, self-questioning, and self-monitoring (Biemiller & Meichenbaum, 1992, p. 65). The study by Haladyna and Roid (1983) aimed to construct criterion-referenced tests using two methods for measuring: the first depended on selecting a sample of vocabulary from a certain field in a random manner, while the second relied on selecting a sample of vocabulary that was appropriate to the ability level of the student undergoing the test, in order to determine which of the two methods could reduce measurement errors. The study conducted by Wilson et al. (2007) aimed to measure mathematical skills using a multiple-choice test. The sample consisted of 566 male and female students. Most students were assigned tasks in algebra and arithmetic that required mathematical skills. The Rasch model was used for data analysis to validate the test and to describe the students' proficiency levels in mathematical skills. The findings revealed that students' test performance was influenced by age, mathematical knowledge, prior academic level, and self-concept. The model showed consistency in the data collected from the study sample (Wilson, 2007, p. 214). Another study by Chang et al. (2010) analyzed items from a reading test in Taiwan used for admission to medical college. The sample included 36 students, and the test consisted of 100

https://doi.org/10.22521/edupij.2025.19.579 Published online by Universitepark Press

specialized items. These items were analyzed using the Rasch model, in which each participant evaluated each item and rated it on a logarithmic scale. The analysis revealed that 60 of the 100 items were below the examinees' ability level, indicating that the Rasch model used to analyze the test was ineffective (Chang et al., 2010, p. 335).

1.3. Research Gap / Contribution

Based on the above, the current research problem can be summarized in two main aspects. The first is the lack of criterion-referenced achievement tests designed for university students in Iraq, as efforts have focused primarily on constructing such tests in general, without specific application to this population. The second is that, to the researcher's knowledge, no previous study has used amplified objectives in the construction of achievement tests based on item response theory.

The importance of the current research is demonstrated in several ways. Criterion-referenced achievement tests have sparked considerable controversy and debate among experts in educational and psychological measurement and evaluation, underscoring the need for further attention and investigation of this type of testing. These tests represent a relatively novel approach when compared to more traditional methods used in educational and psychological contexts.

Additionally, there is an urgent need to develop criterion-referenced achievement tests that can serve as alternatives to traditional tests, particularly at the university level. The use of amplified objectives in developing such tests introduces a new strategy that enhances the accuracy and relevance of the assessment process. Furthermore, the study is grounded in a solid scientific foundation that specialists in psychological and educational measurement rely on when adopting the magnification strategy to construct achievement tests for university students.

According to the researcher's knowledge, this is the first study of its kind to be applied to second-year students in the Department of Educational and Psychological Sciences at Iraqi universities. It specifically focuses on evaluating the effectiveness of using amplified objectives in constructing an achievement test in environmental education.

The current research aims to develop a narrative test with amplified objectives, in accordance with item response theory. The study is limited to students in the second-stage Colleges of Education, Department of Educational and Psychological Sciences, across all Iraqi universities for the academic year 2024–2025.

The theoretical definition of the criterion-referenced test in the current research is a test used to estimate a learner's performance level, determine their position, and compare that level with a predetermined, precise benchmark point on the achievement continuum.

Amplified objectives are comprehensive behavioral objectives. These include the objective, the behavioral action, the student, the scientific content, the acceptable level of performance, and the circumstances under which the objective is achieved. They are also procedurally defined as the method the researcher uses to formulate objectives, including precise details about the content of the study material.

This study addresses the lack of calibrated, criterion-referenced tests for Iraqi undergraduates by operationalizing amplified objectives within a Rasch 1PL framework, thereby linking clearly specified learning outcomes to invariant measurement

1.4. Research Aim and Research Questions

The current research seeks to identify the construction of a narrative test using amplified objectives in accordance with item response theory. and it’s limited to students of the Colleges of Education - second stage - Department of Educational and Psychological Sciences - all Iraqi universities, for the academic year 2024-2025.

https://doi.org/10.22521/edupij.2025.19.579 Published online by Universitepark Press

1.5. Limitations

Although the study sampled six institutions, analyses were collapsed across sites due to the instrument-development focus. The achieved sample (n = 500) exceeds common recommendations for stable Rasch calibration of a 72-item dichotomous test, supporting the precision of item difficulty and person ability estimates.

2. Literature Review

2.1. Amplified objectives and instructional alignment

Amplified objectives are part of the management by objectives theory developed by General Motors in the 1950s. This theory focuses on setting clear, specific goals to achieve specific results (Drucker, 1954, p. 43). Amplified objectives are also part of the goal-setting theory developed by Locke (1968), which focuses on setting clear, specific goals to achieve specific results. In a broad sense, goal setting is the process of setting clearly defined, actionable, and usable goals for learning. Goal theory suggests two general goal orientations that students can adopt: a task-oriented orientation with an intrinsic focus on learning and improvement, and an ability-oriented orientation with an extrinsic focus on external rewards (e.g., getting good grades and outperforming other students). the former are usually referred to as detailed learning or mastery goals, and the latter are classified as performance goals (Dweck, 1986).

Amplified objectives go several steps beyond behavioral objectives; they are more precise and specific. Amplified objectives have been used to evaluate language programs (Wongsothorn, 1986), to assess the language needs of parents, students, and teachers (Sikkhachai, 1987; Sittisarn, 1987; and Pakpian, 1987), and to analyze English language texts (Wongpinit, 1996).

To trace the development of amplified objectives, we must start with behavioral goals, which define learning in measurable terms, but still lack the necessary components for dealing with human communication through language. Baker (1974, p. 18) Stated That Behavioral objectives are not a basis for curriculum development. This is because behavioral objectives lack specific representational content, which is necessary for recycling content if students fail to achieve certain objectives. Amplified objectives are typically viewed as one of the strategies used in constructing achievement tests that encourage learner autonomy (Wentzel, 1991, p.54). Studies have shown that constructing achievement tests using the magnifying goal strategy contributes to developing feedback among learners and can lead to higher academic achievement, better performance, a high level of self-confidence, and good self-regulation (Bishop, 2003, p.43).

Researchers confirm that, to improve performance when setting objectives, students must be allowed to set their own goals. In order to improve the level of academic achievement of learners, students must be allowed to participate in setting larger goals. The theory of (participatory goal) has indicated that: students who are allowed to participate in setting the larger goals have a higher level of achievement than their colleagues who do not have a role in setting the larger goals (Tubbs,2002, p.32). Setting overly ambitious goals may lead students to feel like failures. Setting goals that are too high may seem vague and unattainable. However, when goals are relevant to the learners' academic needs and level of learning, they become more achievable, and students are more able to see themselves as successful (Nicholls, 1984, p.87).

Breaking down larger goals into smaller, necessary, achievable procedural tasks will help give students a sense of control, and a sense of competence will come from achieving these small steps linked to the larger goal, which in turn will enhance ongoing motivation and help prevent frustration (Webb et al, 1994, p.14).

https://doi.org/10.22521/edupij.2025.19.579 Published online by Universitepark Press

Here it is necessary to point out the problems that may appear if students are allowed to participate in setting the larger goals. It is known that teachers set the goals that they seek to achieve at the end of each semester based on an official educational curriculum, and thus there will be differences between what the teacher seeks to achieve for the learners and the goals that the learners participate in setting (Cross & Steadman, 1996,p.31).

According to Dornyei (2001), these differences between teacher and student goals can lead to a disconnect between the teacher's "official classroom goal" and the individual student's goal. (p.54)This disconnect, in turn, can lead to a lack of understanding on the part of students of how and why they are participating in the learning process. Boekaerts (2002) noted that the optimal strategy is a combination of a learning goal defined by the student and agreed upon by the teacher. In this case, the learning goal becomes a mutual agreement between the two parties to "invest effort." This agreed-upon learning goal has a "better chance of achievement" (Boekaerts, 2002, p. 18).

In this type of learning environment, students are more motivated to produce high-quality work because they are not simply completing a task, meeting a requirement, or preparing for a test—they are taking a step toward achieving their own aspirations. This type of education also creates an environment that motivates students to engage in their own learning. This, in turn, makes students enthusiastic and consciously involved in the classroom learning environment. On the other hand, to motivate students, learning environments must provide opportunities that invite them to make an effort and participate. Tasks must be engaging and meaningful (Turner, 1995, p. 413).

2.2. Amplifing Objectives and independent self-learning

Independence is "the ability to take responsibility for one's own learning" (Benson, 2001, p.43). Dickinson (1987) and Holec (1981) have suggested that independence is a long-term goal of education that can be achieved through operationally detailed, amplified goals (Candy, 1988, p. 43). The required shift from teacher-based to learner-based education further emphasizes the importance of employing amplified objectives in the educational process by constructing achievement tests that accurately and in detail measure learners' understanding of concepts and topics (Scharle & Szabo, 2000, p. 35).

Therefore, it is necessary to develop learners' strategies for applying the amplified goals by strengthening their scientific research, problem-solving, and critical thinking skills, as reflected in their achievement of those goals. Benson, Dickinson, and Holick claimed that independence is not innate but develops through the learner's training; that is, learners need to be taught learning strategies and how to use them. Therefore, it is important to consider the processes or activities that teachers might use to explicitly guide their students toward greater autonomy (Elliott & Dweck, 1988, p. 44).

Maximum objectives can help achieve the following goals:

1. Classifying content
2. Identifying main skills and sub-skills
3. Defining achievement standards
4. Providing precise procedural information about the variable
5. Linking to previous and subsequent maximal objectives (Wongsothorn, 1993, 2003)

https://doi.org/10.22521/edupij.2025.19.579 Published online by Universitepark Press

2.3. Characteristics of amplified achievement objectives

It is important to emphasize that simply setting overly ambitious goals does not necessarily lead to improved academic achievement (Schunk, 2003, p. 34). There are a number of important characteristics to consider when constructing an achievement test using the exaggerated goal strategy:

1. The amplified objectives should be operationally detailed.

2. The amplified objective can be measured operationally.

3. The amplified objectives should be appropriate to the learners' level (Griffee & Templi, 1997, p.22).

4. The standard characteristics of difficulty and differentiation should be taken into account.

5. Giving learners the space to participate in determining the appropriate amplified goals for them (Pritchett, 2002, p.32).

6. Establishing a clear criterion or standard for judging the learner's achievement of the amplified goal (Dornyei, 2001, p.33).

2.4. Steps to Build an Amplified Achievement objectives:

An augmented achievement objective is constructed through the following steps:

1. Determine the cognitive level for which the augmented objective is to be constructed (knowledge, understanding, application, analysis, synthesis, evaluation).

2. Determine the action or behavior to be performed (mention, enumerate, explain, summarize, etc.).

3. Mention (the student) as part of the augmented objective formulation.

4. Determine the purpose of the augmented objective (defining a specific concept or enumerating the causes of a specific phenomenon, etc.).

5. The criterion for passing or success should be included in the behavioral objective (with a validity rate of no less than 50% or 60%).

6. The extent to which the learner will use the book or not should be determined (Doran, R. 1989, p.36)

2.5. The Importance of Amplified objectives in Tests

Moving to a learning environment based on amplified goals requires a deeper understanding and research into the factors that influence learner achievement, such as motivation, self-regulation, ability, effort, time management, self-assessment, and persistence. Research examining the relationship between the learning environment, amplified goals, and learner achievement outcomes has significantly contributed to our understanding of the importance of amplified goals in learning. It paved the way for the next important research step: determining how to facilitate the writing of amplified objectives in raising the level of student achievement (Ames, 1992, p. 153).

The relationship between learner effort and goal-based academic achievement is supported by evidence from Ames & Archer (1988) and Nicholls et al. (1985). which linked the objectives of the amplified goals to the belief that this type of goal can direct individuals towards developing new learning skills, trying to understand their learning-related behavior, improving their level of academic proficiency, or achieving a sense of mastery based on self-referential standards that are set while writing the amplified goals (Ames, 1992a: 87) Grandiose goals are associated with a wide range of motivational variables that contribute to achieving high academic achievement, a necessary

https://doi.org/10.22521/edupij.2025.19.579 Published online by Universitepark Press

prerequisite for self-regulated learning (Ames, 1992a, p. 67). According to Jagacinski and Nicholls (1984, 1987), adopting grandiose goals is highly associated with pride, satisfaction with effort, and a preference for challenging work and risk-taking (Wentzel, 1987, p. 34).

Setting larger goals will direct efforts toward students' wants and needs according to their own schedules rather than those of their parents or teachers. Thus, students will gain ownership, meaning the task will be a personal endeavor to succeed rather than a general academic one. Students will find it easier to solve dilemmas because they will focus on interests they have designed themselves (Lions & Quest, 1992, p. 65).

Research in recent years suggests that one source of differences between high and low achievers is the degree to which a person becomes self-regulated, with high achievers engaging in grandiose goal setting, planning, self-questioning, and self-monitoring (Biemiller & Meichenbaum, 1992, p. 65). Haladyna & Roid (1983) study: The study aimed to construct criterion-referenced tests using two methods for measuring: the first: depends on selecting a sample of vocabulary from a certain field in a random manner, and the second: depends on selecting a sample of vocabulary from a certain field, as it is appropriate to the level of ability of the student undergoing the test, in order to determine which of the two methods can reduce measurement errors. Wilson et al. (2007) conducted a study aimed at measuring mathematical skills using a multiple-choice test. The study sample consisted of 566 male and female students. Most students were given skills in algebra and arithmetic that relied on mathematical skills. The Rasch model was used to analyze the data to evaluate the test and to describe students' levels of mathematical skills. The results showed that students' test scores depended on their age, mathematical knowledge, and previous levels of mathematical ability, as well as their self-concept and age. The model showed consistency in the data collected from the study sample (Wilson, 2007, p. 214). Chang et al. (2010) study: aimed to analyze items of a reading test in Taiwan for admission to medical college. The study sample consisted of 36 students, and the test consisted of 100 specialized items. The test items were analyzed using the Rasch model, in which each individual in the study sample evaluated each item and graded it on a logarithmic scale. The results indicated that there were (60) items out of (100) items that were below the ability level of the examinees, and the ineffectiveness of the Rasch model used in analyzing the results of the test used (Chang et al, 2010, p. 335).

3. Methodology and Procedures

3.1. Study Design

A descriptive, instrument-development design: amplified objectives specified the construct; items were authored to those objectives; a field test provided responses for psychometric calibration; and evidentiary sources (content review, Rasch model-data fit, reliability, and standard setting) were integrated to evaluate score interpretability.

3.2. Study Population and Sample

The current research population consisted of second-year students from the six public Iraqi universities in the 2024-2025 academic year.

As the total number of students reached 920, male and female students. They are also distributed by gender, with 390 male students (42%) and 530 female students (58%).

The research sample consisted of second-year students in the educational and psychological sciences departments in the faculties of education for the humanities in the universities included in the research community, which were selected using the stratified random sample method, where (500) male and female students were selected at a rate of 54%, with 210 male students and 290 female students distributed proportionally on the variables of university/college and gender according to the percentages of the distribution of the research community.

Constructing a criterion-referenced achievement test using the amplified objectives of the environmental education subject according to the response theory, adopting the steps followed and approved for constructing this type of test.

The results of the studies that used the Rasch model in analyzing test data in determining the number of sample members varied, as the results of the study (Van de Weger, 1986) showed that the accuracy of the estimates increases with the increase in the number of test items from 10 to 25 to 50 items and with the increase in the sample size from (25 to (50 to (500 examinees. (Habeeb& Kadhim, 2018)

Stratified random sampling with proportional allocation was selected to (a) mirror the population composition across institutions and gender; (b) increase precision for subgroup estimates; and (c) meet common sample-size guidance for Rasch calibration of 70+ dichotomous items. With n = 500, standard errors for item difficulty estimates are acceptably small for program-level diagnostic use

3.3. Determine the purpose of the test

The test aims to measure the achievement of second-year students in the educational and psychological sciences departments in the faculties of education for the humanities by adopting a criterion-referenced achievement test using the amplified objectives of the environmental education subject according to the item response theory. After determining the objective of the test, the following steps were followed:

3.3.1. Determining the content of the study material

The researcher distributed a questionnaire with vocabulary for the colleges of education in Iraqi universities to 6 instructors who teach the environmental education course. The purpose of the questionnaire was to identify the importance of the topics that should be taught to second-year students as seen by the specialist. The task book has been approved for teaching environmental education (In Environmental Education by Professor Dr. Saadoun Salman Najm, 2023), and through the two previous steps, the researcher was able to determine the content of the material across five chapters.

3.3.2. Formulating behavioral objectives

The researcher formulated a set of amplified behavioral objectives that included the above topics, as amplified behavioral objectives are formulated in the form of (if + behavioral action + student + scientific content + acceptable level of performance + circumstance of achieving the objective). Accordingly, (12) objectives were formulated for the first semester, (6) objectives for the second semester, (19) objectives for the third semester, (14) objectives for the fourth semester, and (21) objectives for the fifth semester.the total number of objectives reached (72) behavioral objectives that measure the four levels of cognitive depth, as follows: (35) objectives to measure memory, (29) objectives to measure understanding, (8) objectives to measure application.

3.3.3. Preparing a specification table (test map)

When preparing achievement tests that are characterized by objectivity and comprehensiveness, the specification table is considered an important research procedure, because it takes into consideration both aspects of the educational content and the amplified behavioral objectives that were formulated according to the weight and importance of each, which enables us to distribute the test items according to the educational content. Among the benefits of the specification table that were mentioned by Khattabiyya (2005: 508)

https://doi.org/10.22521/edupij.2025.19.579 Published online by Universitepark Press

Which includes Constructing a balanced test with the amount of effort expended in teaching each topic, and giving the real weight to each part of the content according to its relative importance, and the most important thing is to achieve the validity of the test content to a large extent, and the researcher prepared the specifications table for the specified topics and the weights of the content of the topics were calculated and the weights of the behavioral objectives were calculated. Table 1.

Table 1. Test map of behavioral objectives to represent them in the achievement test

Chapter	N paper	weight and importance	behavioral objectives			Total of Question
			memory	understanding	application	
First	19	16%	6	5	1	12
Second	11	9%	3	2	1	6
Third	32	27%	9	8	2	19
Fourth	22	19%	7	6	1	14
Fifth	34	29%	10	8	3	21
total	118	100%	35	29	8	72

3.3.4. Formulation of test items

The researcher formulated 72 objective multiple-choice items distributed on the basis that each item included four alternatives, one of which was correct and the rest was incorrect, based on the test map, as it covers a large sample of the content vocabulary of the study material, which makes its validity high. It enjoys the objectivity of correction, thereby raising its reliability. In addition to the low rate of guesswork in answering questions, these represent the advantages of multiple-choice tests (Sulaiman & Raja, 2010, p. 255). According to the specifications table, the researcher prepared the test according to precise, detailed rules that guide the test preparer in formulating items that measure the same objective.

3.3.5. Test Instructions

The instructions specify how to answer, which are: (answer all questions, there is a specific time period for answering, specify the student's gender, university, and college), and urge the respondent to answer accurately and objectively and not leave any question unanswered.

Answer sheet: The researcher designed a special answer sheet that is separate from the test booklet and based on a design that facilitates the answering process for students because the test booklet is to reduce the material cost, especially since the research sample amounted to (500) male and female students, and the researcher used separate answer sheets that included some general information about each examinee. as well as the sequence of the test items, and the four answer alternatives in front of each of the test items, where the participants or examinees are asked to put a circle around the letter of the alternative that expresses the correct answer.

Correcting the test answers: After preparing the test in its initial form, which consists of 72 items, a standard was set for correcting the answers by giving a score of 1 for the correct answer, and a score of zero for the wrong answer. Thus, the student's total score on the test is equal to the total number of items (72) he answered correctly, plus the lowest score (zero), and the theoretical average for the test is 36 points.

Logical analysis of the test items: The researcher relied on a preliminary step to verify the validity of the items and their suitability in measuring what they were prepared for, and to ensure the integrity of the items and their suitability for the specified purposes, the clarity of their formulation, and the effectiveness of the alternatives. Therefore, the test was presented in its initial form. On a group of experts specialized in educational and psychological sciences, whose number reached (12) arbitrators (appendix), to adopt their opinions in arbitration on the suitability of the test items for the age group targeted for the application of the test, and it was found that all the test items consisting of (72) test items are suitable for measuring what they were prepared for. all experts unanimously approved the test at 100%, with some minor modifications made after an intensive discussion between the supervisor and the researchers, in light of the experts' opinions.

3.4. Ethical Consideration

This study received ethics approval from the Universities from which the data were collected. Prior to data collection, participants read an information sheet and provided written informed consent. Participation was voluntary; students could withdraw at any time without penalty. No identifying information was retained on response files; institution and gender were coded only for sampling strata. All procedures complied with the Declaration of Helsinki and institutional policies.

4. Results

According to the item response theory, after administering the achievement test to a statistical analysis sample of 500 male and female students, their responses were analyzed using the one-parameter logistic (Rasch) model. To verify the test's objectivity in measurement, the researcher ensured that the test items met the one-dimensional assumption of the one-parameter logistic (Rasch) model. that is, all items measure one latent trait, and the most appropriate model is chosen to analyze the data and estimate the values of item parameters, and the items that match the model are chosen (Hulin, et al., 1983, p.79).

One-dimensional assumption: this assumption holds that the test measures a single latent ability, meaning that, according to Orlando, the variance between items can be explained by a single dimension. The conclusion, as Wiberg says, is that according to the modern theory, which assumes unidimensionality, meaning that the test measures a single latent ability. This is consistent with the assumption that a respondent with high ability is more likely to answer the item correctly, and vice versa (Zakri, 2009, p. 45).

To achieve the one-dimensional assumption, the researchers followed the following:

- Factor analysis: The unidimensionality assumption was verified by subjecting the researcher to the items of the criterion-referenced achievement test using the amplified objectives of the environmental education subject for first-year students in the educational and psychological sciences departments in the faculties of education for the humanities, in order to verify the factorial structure of the test. Accordingly, the 500 questionnaires used for the statistical analysis of the items were subjected to factor analysis. After performing a principal components analysis, the direct analysis yielded one factor. After conducting orthogonal rotation using the varimax method (maximum variance) of Kaiser, one factor with a latent root greater than one integer was obtained according to Guttman's lower bounds method. The value of the latent root was (29.662) and contributed to explaining (50.487%) of the total variance. This method assumes that the significant factor or latent root that can be explained is equal to or greater than 1 (Abdul Khaliq, 1983, p. 118). This is also supported by Rechase's (1985) suggestion: if the first factor explains at least 20% of the explained variance, this is a sufficient indicator of unidimensionality (Hambelton & Swaminathan, 1985). Table 2 presents the eigenvalues.

Table 2. Eigenvalue the test items with the reference criterion

N item	Eigenvalue	N item	Eigenvalue	N item	Eigenvalue
1	0.692	25	0.415	49	0.484
2	0.790	26	0.571	50	0.379
3	0.409	27	0.553	51	0.337
4	0.333	28	0.572	52	0.496
5	0.319	29	0.371	53	0.359
6	0.513	30	0.470	54	0.468
7	0.302	31	0.404	55	0.407
8	0.393	32	0.591	56	0.384
9	0.542	33	0.486	57	0.368
10	0.529	34	0.539	58	0.434
11	0.350	35	0.489	59	0.328
12	0.331	36	0.584	60	0.442
13	0.400	37	0.690	61	0.415
14	0.366	38	0.790	62	0.571
15	0.504	39	0.409	63	0.553
16	0.521	40	0.564	64	0.572
17	0.390	41	0.319	65	0.371
18	0.559	42	0.513	66	0.470
19	0.484	43	0.598	67	0.404
20	0.379	44	0.393	68	0.591
21	0.337	45	0.542	69	0.486
22	0.496	46	0.529	70	0.539
23	0.359	47	0.350	71	0.489
24	0.468	48	0.331	72	0.584

According to Guilford's standard, all items were saturated with the general factor at a level of more than 0.30 (Guilford, 1998, p. 500).

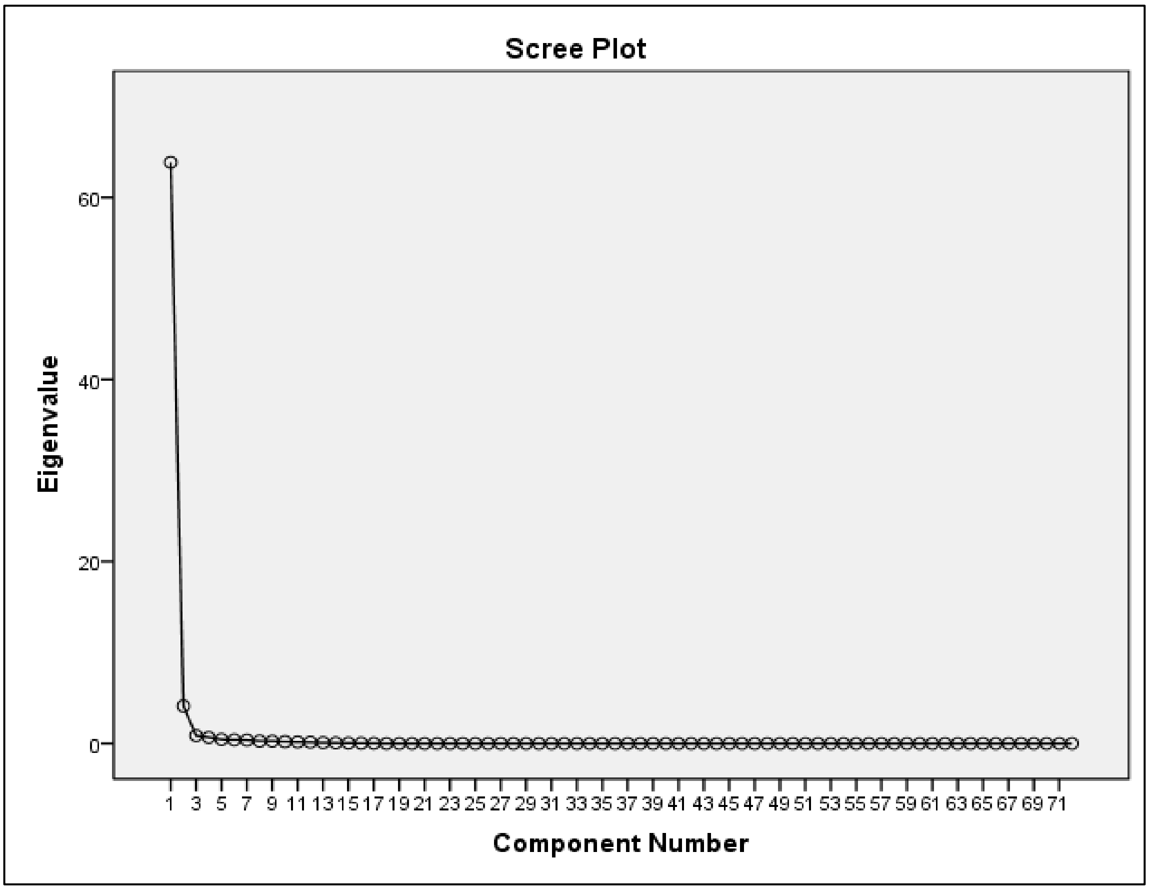

Figure 1. Eigenvalue the test items with the reference criterion of environmental education

Internal consistency: The item validity coefficient is one of the most accurate measures of internal consistency. It means that each item follows the same path as the test as a whole (Essawi, 1985, p.51).to reveal the relationship between two variables, one of which represents a hypothetical intermittent scale such as a item in a test which is given (one or zero) (bi-scale) and the other represents an interval or relative scale (continuous), the bi-serial correlation coefficient was calculated, which is used as indicated by Glass & Stanley (1970), Glass & Stanley, (1970, p.17), and Table 3 shows that.

Table 3. Values of correlation coefficients of the item score with the total score for the achievement test items

N item	Item validity	N item	Item validity	N item	Item validity
1	0.422	25	0.426	49	0.497
2	0.451	26	0.429	50	0.481
3	0.272	27	0.507	51	0.475
4	0.338	28	0.301	52	0.491
5	0.399	29	0.371	53	0.494
6	0.317	30	0.360	54	0.496
7	0.407	31	0.387	55	0.479
8	0.382	32	0.301	56	0.481
9	0.413	33	0.336	57	0.494

https://doi.org/10.22521/edupij.2025.19.579 Published online by Universitepark Press

10	0.472	34	0.389	58	0.496
11	0.383	35	0.345	59	0.481
12	0.399	36	0.411	60	0.489
13	0.333	37	0.365	61	0.500
14	0.473	38	0.447	62	0.468
15	0.378	39	0.382	63	0.485
16	0.360	40	0.339	64	0.477
17	0.326	41	0.345	65	0.356
18	0.332	42	0.332	66	0.326
19	0.487	43	0.491	67	0.468
20	0.310	44	0.490	68	0.496
21	0.455	45	0.484	69	0.481
22	0.447	46	0.478	70	0.326
23	0.403	47	0.495	71	0.441
24	0.448	48	0.485	72	0.365

Note: the critical value of the correlation coefficient of (0.088) at a significance level of (0.05) and a degree of freedom of (498)

It is clear from Table 3 that the calculated values of the correlation coefficients of the item score with the total score were greater than the critical value of the correlation coefficient of 0.088 at a significance level of 0.05 and a degree of freedom of 498. This indicates the internal homogeneity of the environmental education test items in measuring the latent trait, an additional indicator of the assumption of one-dimensionality.

Verifying the assumption of local independence: this assumption was verified for the achievement test of the subject of Psychology of Properties using the elimination of alternatives technique and without using the elimination of alternatives technique as follows:

- In order to avoid the problem of not achieving local independence, he adopted the best way to overcome this problem by avoiding its occurrence during the test design stage (Alam, 2005, p.106). The researcher discovered, through the test preparation procedures, that it presented verbally independent situations and did not include any indications or hints for responding to any other item.

- The assumption of local independence is equivalent to the assumption of unidimensionality, as indicated by Hambleton & Swaminathan (1985). That is, the two assumptions are similar, but not conceptually equivalent. By fulfilling the assumption of unidimensionality in the previous steps, the assumption of local independence is also satisfied (Erguven, 2014, p. 26).

The characteristic curve of the item: the shape of the characteristic curve of an item depends on the item's parameters of difficulty (β), discrimination (a), and individual ability (θ). It is expected that items will vary in their locations, discriminating power, discrimination (a), and individual ability (θ). It is expected that the items of the criterion-referenced achievement test for environmental education using amplified objectives will vary in their locations and discriminatory power, which is consistent with graded responses and, consequently, with the assumptions of the model adopted in the current study. To verify the variance of the items in discrimination coefficients, Hambleton et al. (1991) suggest examining the distribution of the correlation coefficients between the test items and the total score to provide an idea of the degree of homogeneity of the scale items' discrimination indicators. (Hambleton et al., 1991, p. 10) Therefore, the researcher examined the correlation coefficients shown in Table 3 and, according to the descriptive statistics, the table clarifies that.

Table 4. Descriptive statistics values for the distribution of correlation coefficients for the correlation of the item score with the total score

Test	N item	Range	Mean	St.divation
Environmental Education	72	0.270	0.422	0.059

Descriptive statistics of the distribution of correlation coefficients show that the discrimination indicators for all items of the environmental education test vary somewhat in their discriminatory power, and thus call for different responses from individuals. This is consistent with the adopted single model and thus the use of the model in analyzing the data of the current research.

Imposing a low guess factor: In order for the researcher to know that the items of the environmental education test, which is a criterion-referenced test prepared according to the amplified objectives, do not include any trace of guesswork because it relies on the one-parameter logistic (Rasch) model, he chose a random sample of 50 male and female students, representing 10% of the statistical analysis sample, to study their answers to the very difficult items.

When compared with the theoretical rate of random guessing (0.25), and the very low rates of correct answers of students with low abilities on the very difficult items, we conclude that the criterion-referenced achievement test for the environmental education subject using amplified objectives is almost free of the factor of guessing, and thus the data becomes suitable for analysis according to the single model.

Enforce speed limit: This assumption can be confirmed by examining the percentage of individuals who completed the test within the specified time period. The speed factor is an influential variable in test performance. It is not considered an important factor if at least 80% of the sample answered all the test items (Hambleton et al., 1991, pp. 11-12). Based on this assumption, the percentage of students who completed the test was 91%, and the percentage of items answered was 89%. This indicates that the achievement test is free from speed. Thus, the initial verification of the suitability of the data derived from applying the test for analysis according to the single model was carried out. After the researcher verified the main assumptions of the item response theory and ensured the suitability of the environmental education test items for analysis using the one-parameter model, the Item Response Command Language (ICL) program was used to estimate the one-parameter model.

Estimating the values of item difficulty parameters and the examinees' abilities: Using the computerized item response command language program to extract item difficulty parameters and the examinees' abilities estimated in logit units, the criterion-referenced achievement test for environmental education is analyzed using the amplified objectives according to the single model. It is evident that all items of the criterion-referenced achievement test for the environmental education subject using amplified objectives, which was subjected to statistical analysis according to the single model, are statistically significant, as the calculated chi-square value for them was smaller than the tabular chi-square value of 28.87 at a significance level of (0.05) and a degree of freedom of (18). This indicates that it is consistent with the assumptions of the single model, and thus all (72) items of the achievement test were retained.

Goodness-of-fit test: Three basic criteria can be used to select appropriate items compatible with the model and to exclude inappropriate items. To ensure that the items match the single model well, the following criteria were relied upon:

- The chi-square test was used to determine the item's conformity to the model, as it is the statistical criterion for compatibility, and to ensure that the item's definition of the variable is consistent with that defined and expressed by the rest of the items, generally from one individual to another. If there is consistency between individuals' observed responses to an item and their probability of success on it, this means there is consistency between their responses to the item and their total test scores, i.e., their responses to the rest of the test items. This indicates agreement between the characteristic expressed by the item and that expressed by the rest of the items across the entire sample. This means that the item is generally suitable for the model's requirements (Kazem, 1988, p. 84).

The Chi-square is also used to assess the overall fit of each individual, excluding those who are not suitable for the model. When the Chi-square value is statistically significant, the item should be deleted, as it does not express the same characteristic as the rest of the items, as the relative difficulty of the items for these individuals differs from that for most individuals. Since excluding individuals who are not suitable for the model affects the results of the items' suitability, they must be removed from the analysis before calculating the overall suitability of the items (Al-Sharqawi et al., 1996, p. 353). These examinees might be responsible for the non-conformity of some items to the model's assumptions (Wright & Stone, 1979, p.82). Therefore, the responses of the examinees who did not conform to the model's assumptions were manually deleted. This was done by dividing the test into an easy test and a difficult one according to the computer program's grading. According to this criterion, the chi-square test indicated that all examinees conformed to the model. Accordingly, the statistical analysis sample of 500 male and female students was subjected to the analysis of procedures for verifying the assumptions of the single model.

The independence of the item from the sample means that the difficulty order of the items must remain constant at each ability level. This is done by testing the stability of the item's relative difficulty across different ability levels. This assumption depends on measuring the extent of deviation between the observed characteristic curve of the item and the characteristic curve expected from the model. The observed characteristic curve for an item shows the proportion of correct answers observed on that item for individuals across different ability levels, while the item's probability characteristic curve shows the probabilities of correct answers on that item at different ability levels. Testing whether the observed characteristic curves for an item have a common general curvature shape. When the items fit the model, when analyzing the achievement test using the single model, it was found that the test had a discrimination of 0.422, and thus this criterion was verified.

The suitability of the discriminating power of the items, and thus the test items are the most powerful discriminating items, and the best discriminating items are effective on an average range of ability, so the best items are those that are average in terms of discriminating power, so it was considered that the best possible slope for the distinctive curve of the item is when its slope angle is (45) degrees on the continuum of the characteristic or ability.

Then the slope of this potential curve of the model oscillates around the value (1), and the discriminating power of the item is appropriate when the observed discriminating curve of the item approaches the potential curve of the model. If it is effective over a narrow range of power, the least discriminating items are effective over a wide range of abilities. Then the discrimination coefficient of the item, which describes the relative curvature of the discriminating curve of this item, is close to one (Al-Sharqawi et al., 1996, p.354).

https://doi.org/10.22521/edupij.2025.19.579 Published online by Universitepark Press

Availability of objectivity requirements in measurement: verifying the availability of objectivity requirements in the measurement tool that was created according to the single model. The availability of objectivity requirements for measurement essentially tests the validity of this model in achieving measurement objectivity. The following was verified:

1. The items identify one variable among themselves, as this assumption was verified through two indicators, which are (conducting a factor analysis of the individuals' score on each item), as the factor analysis revealed the presence of a general factor, and the items' conformity to the model's assumptions represents evidence that the items measure a one-dimensional trait (Wright & Panchapakeson, 1969, p. 25) by using the chi-square test, the items that did not fit the model were excluded. Thus the researcher verified the assumption of unidimensionality.

2. Independence of measurement: The independence of measurement was verified from two aspects, which are (the independence of measurement from the items that the examinees answer, such that the assessment of the difficulty of the item does not depend on the rest of the test items, and the assessment of the examinees' ability does not depend on the test items that they answer) and (the independence of measurement from the ability of the sample that takes the test).

The researchers achieved these two points by ensuring that the assessment of the examinee's ability does not depend on the ability of the other examinees who answer the test, and that the assessment of the item's difficulty does not depend on the ability of the examinees who answer the test. (The independence of measurement from the ability of the sample that performs the test, which requires verifying this hypothesis, is the presence of two samples of suitable individuals, in order to perform the same group of items graded in the single model.) This leads to verification of two aspects, which are:

a. The statistical equivalence of the difficulty estimates for each item of the test resulting from the analysis of the performance of individuals in one of the two samples (i.e., taking into account the standard error of these estimates), with those estimates derived from the analysis of the performance of individuals in the total sample, indicates that the estimate of the difficulty of the item is not affected by the difference in the analysis sample, which means that the estimate of the difficulty of the item is free from the ability of the individuals who answer the test.

b. The statistical equivalence of the ability estimates corresponding to each possible total score on the test, resulting from analyzing the performance of individuals in one of the two samples on this test (taking into account the standard error of these estimates), with those derived from analyzing the performance of individuals in the total sample, means that the estimate of the ability of the examinee who obtained a certain total score on this test is not affected by the difference in the level of performance of the analysis sample, this means that the individual's ability, as measured by this test, is freed from the ability of the other individuals who answer it (Kazem, 1988, p.96).

The mean score was used from the score file to divide the statistical analysis sample into two groups: high-level and low-level. Then the responses of the two samples (high and low levels) to the test were analyzed using the computer program to calculate item difficulty and its standard error, and ability estimates and their standard errors. The ability estimates and their standard errors were then compared with the two parameters (difficulty and ability) derived from analyses of the total sample and the high- and low-level samples to verify the statistical equivalence of these estimates. The corresponding estimates are considered statistically equivalent if the difference between any two estimates does not exceed the sum of their standard errors (Al-Sharqawi et al., 1996, p.526).

It is evident that all items of the criterion-referenced achievement test for the environmental education subject using amplified objectives were equivalent to corresponding estimates when analyzing the total sample as reference estimates, as the difference between any two corresponding estimates of difficulty estimates was less than the sum of their standard errors.This means that these estimates are not affected by differences in the level of the analysis sample. This achieves the second aspect of the independence of the measurement from the sample used for these items, which is that the difficulty of the items is freed from the ability of the sample to which the test was applied.

5. Discussions

The present research aimed to construct a criterion-referenced test aligned with the principles of Item Response Theory (IRT), using the Partial Credit Model (RASCH 1PL MODEL) as the psychometric framework, to measure student achievement in environmental education among second-year undergraduates at Iraqi public universities. The discussion herein addresses the major findings in relation to the research objectives, theoretical expectations, and prior literature, focusing on the suitability of the psychometric model, the technical functionality of the ICL program, the test's structural characteristics, and the implications of the model-data fit statistics.

5.1. Suitability of the RASCH 1PL MODEL Model in Achievement Test Construction

One of the central objectives of this study was to determine whether the RASCH 1PL MODEL, a unidimensional polytomous IRT model, was appropriate for the analysis and calibration of the criterion-referenced test constructed. The model assumes that each item contributes to the measurement of a single latent trait—in this case, the cognitive understanding of environmental education content. The item fit statistics provided by the ICL software —specifically, the Infit Mean Square (MNSQ) and Outfit MNSQ values —consistently fell within the acceptable range of 0.77 to 1.33, as proposed by Wright and Linacre (1994). These results suggest that all items adequately conformed to the expectations of the RASCH 1PL MODEL model, thereby affirming the test's structural validity.

Furthermore, the analysis confirmed the unidimensionality assumption of IRT, which posits that a single latent trait underlies item responses. This was further corroborated by the limited presence of misfitting persons and items, as evidenced by the low number of residuals outside the ±2 range in the standardized residual plot. These findings are consistent with those of Wilson et al. (2007), who emphasized the utility of the RASCH 1PL MODEL in educational assessments when item scoring involves ordered categories. Therefore, the successful application of the RASCH 1PL MODEL supports the claim that a carefully constructed criterion-referenced test can be rigorously analyzed under modern psychometric theories, allowing for more precise calibration of both item difficulty and individual ability levels.

5.2. Efficacy of the ICL Program in Estimating Model Parameters

A key methodological feature of the current research was the application of the Item Response Command Language (ICL) for the calibration and analysis of test data. The ICL proved efficient at producing both textual and graphical outputs, facilitating a nuanced understanding of the model-data fit. Its outputs included detailed statistics such as item parameter estimates, fit indices, person abilities (θ), category characteristic curves, and item characteristic curves, which are instrumental in evaluating the quality of test items and the reliability of inferences made from them.

Of particular importance was the software's ability to estimate parameters with minimal standard error and to display the category probability curves, which illustrate the probability of a given response across the latent trait continuum. These curves revealed an orderly progression of categories for each item, further indicating the test's internal coherence. Moreover, the program's estimation of threshold parameters provided critical insight into the transition points between response categories, a feature crucial to polytomous IRT models such as the RASCH 1PL MODEL.

https://doi.org/10.22521/edupij.2025.19.579 Published online by Universitepark Press

The separation reliability index (0.89) shows that the test effectively distinguished different levels of student understanding while maintaining low levels of noise and model misfit. The program's ability to align the test items and individual θ scores on a common scale allowed for a more precise determination of the test's measurement range, demonstrating its suitability for educational assessments that demand fine-grained interpretation of learner competence.

5.3. Quality and Representativeness of the Constructed Test

The third point of discussion concerns the quality of the criterion-referenced test and its alignment with the theoretical construct of environmental education. The convergence of item difficulty locations and the distribution of person ability estimates on the θ scale around the continuum's center point is indicative of a well-targeted instrument. Most items were clustered within the range of −1.5 to +1.5 logits, which corresponds to the mean range of the trait distribution, suggesting that the test was neither too difficult nor too easy for the intended examinees.

Furthermore, the shape of the ability distribution approximated the normal curve, reflecting the homogeneity of the sample in terms of their latent ability and reinforcing the test's representativeness of the broader population of second-year students. The small proportion of individuals excluded due to misfit suggests that the test is highly suitable for the target demographic. This aligns with the criterion-referenced philosophy, which assesses whether students meet specific learning objectives rather than ranking them against one another.

The empirical data have provided support for the test's pedagogical validity. Since environmental education includes both cognitive understanding and attitudes, the test items were clear and followed a consistent line of reasoning, as they were very disciplined in their discrimination of values. This considers the case that they were logically pedagogically constructed and represented the cognitive domain in their assessment of students. This was also consistent with the observations of Chang et al. (2010) regarding the importance of properly constructed items for capturing the intended learning outcomes in criterion-referenced tests.

5.4. Implications of Model-Data Fit and Sample Representativeness

One of the most valuable outcomes of the study was the strong model-data alignment as evidenced by the item fit statistics and residual distributions. The alignment has real consequences for the interpretation and usefulness of test scores. Strong model-data alignment demonstrates that we can obtain useful estimates of item difficulty and person ability, invariant across samples to the extent that the model is valid. The practical implication is that educators and curriculum planners can use the calibrated items in either a formative or summative context without needing to repeat the validation in each new administration, which represents a significant time and resource savings.

In addition, the person-item map from the ICL software provided a visual representation of the alignment between person ability and item difficulty, with items and persons placed along the ability continuum. The distribution of the participants along the θ scale was relatively symmetrical around the items' location, reinforcing the test's balanced design. The fact that the items tapped all levels of the θ scale indicates that students with differing levels of ability were adequately challenged, enhancing the test's diagnostic utility. The balance of person-item alignment is especially critical in criterion-referenced tests.

In addition, the assumption of local item independence was upheld based on item correlation residuals being very low, and clusters of items were found insignificant in the specific reasons cited for multidimensionality or testlet. This statistical independence confirms that each item contributes unique information to the measurement process, thus enhancing overall test reliability and interpretability. In educational measurement, specifically, local independence ensures that students' responses are based on their level of the underlying trait itself, rather than dependence between items, preserving the interpretive integrity of test scores.

5.5. Pedagogical and Practical Implications

The successful construction and validation of this criterion-referenced test also bear significant pedagogical implications. First, the RASCH 1PL MODEL, as a general one-parameter logistic IRT model, is a powerful IRT model that allows for the interpretation of student responses beyond their raw scores. Teachers and policymakers can identify precise ability thresholds and determine which items best differentiate between mastery and non-mastery abilities. Such diagnostic ability is invaluable for renegotiating curriculum and content, recognizing instructional gaps, or offering targeted remediation.

Second, IRT-based test construction will allow for modularity and flexibility to construct parallel test forms that are measurement equivalent. As an evolving field, environmental education can be updated and/or localized for future assessments without compromising measurement respectability. This guidance should allow assessment equivalence across future cohorts and contexts.

Finally, the use of ICL software, an open-source, free-to-access platform, provides a low-cost, replicable paradigm for other researchers and educational entities in other subject areas who want to build criterion-referenced instruments. Therefore, the research contributes not only to environmental education but also to the modern discussion of test theory, assessment design, and development in the Global South, specifically, and also in the context of higher education in Iraq.

6. Conclusion

The present study aimed to construct a criterion-referenced achievement test for environmental education among second-year university students in Iraqi public universities, utilizing the Rasch model (RASCH 1PL MODEL) and analyzing the data through the Item Response Command Language (ICL) program. Through rigorous psychometric validation procedures, including item fit statistics, model conformity, item-person mapping, and parameter estimation, the research successfully demonstrated the applicability and effectiveness of the Rasch model in achieving the stated objectives. The results of the item analysis confirmed that the test items displayed strong internal consistency and psychometric soundness, aligning closely with the unidimensionality and local independence assumptions of the Rasch model. The majority of the items demonstrated acceptable infit and outfit values, suggesting their appropriateness in measuring the latent trait of environmental education achievement. Additionally, the scale was appropriately fit to the ability levels of the students, and the Person-Item map showed an acceptable distribution of student ability (θ), in relation to the item difficulties (β), thus representing an excellent fit between the difficulty of the test and the abilities of the student.

The analyses also indicated that the students' responses did not depend on item order or contextual dependencies, supporting the local independence assumption. ICL was key in making this analysis happen quickly and efficiently, and supplied text and visualizations that helped to clarify the data's fit to the Rasch model. The model also gave stable and generalizable parameter estimates for both item difficulty and student ability, and for this reason, the test can be used in the future to inform instruction and evaluations. Instructional and evaluative applications. The study verified the appropriateness of criterion-referenced assessment in this field by ensuring that each test item corresponds to clear instructional objectives specific to environmental education. The items were designed to assess whether students had acquired certain knowledge or skill sets, not whether they had outperformed their peers, which aligns with new educational priorities that emphasize ability rather than ability-in-relation-to-group.

https://doi.org/10.22521/edupij.2025.19.579 Published online by Universitepark Press

Additionally, the consistency of common student ability estimates with the typical trait distributions in the sample supports the generalizability of the findings and provides additional confirmation of the test's psychometric integrity and equity. This indicates a fair measure could be used for educational assessment in environmental education for the population. There are limitations to the study, despite the robustness of the findings. The sample was limited to public university students, and the study was contextualized within a specific educational program, which may limit the ability to generalize the findings to other academic fields or educational programs. Furthermore, while omitting some misfitting individuals was defensible from a statistical perspective, there may have been latent variables that affected their performance, such as motivation, test anxiety, or other irrelevant cognitive influences. Future studies should seek to increase the applicability of this criterion-referenced tool across a variety of situations by testing it in other educational and geographical contexts with student populations of more diverse backgrounds. Longitudinal studies could also give insight into how stable the test parameters are over time when used in various instructional interventions. Additionally, including qualitative approaches, such as student feedback or think-aloud protocols, may provide more information about the functioning of the items and the cognitive processes involved in students' test performance.

In conclusion, the research study has made a strong contribution to the field of educational measurement by demonstrating the feasibility of developing a criterion-referenced achievement test for environmental education using a Rasch model. This research confirms the value of applying evidence-based models to educational measurement and advances the dialogue on future psychometric research aimed at improving the validity and reliability of educational measurement options in Iraq and worldwide.

7. Recommendation

1- Implementation of the Criterion-referenced test in Broader Educational Contexts

It is highly recommended that teachers, especially those who teach environmental education in educational and psychological sciences faculties, consider utilizing the criterion-referenced achievement test developed in this study as a standardized assessment tool to measure second-year students' knowledge. The psychometric properties identified in this study — especially the degree to which the test items function according to the assumptions of the one-parameter logistic model (RASCH 1PL MODEL)— indicate that the test is suitable for measuring educational achievement but, even better, for definitively identifying learning deficiencies. The test should not be limited to a sample from this study alone, but staff should seek to adapt it and deploy it in a number of places, including universities and others. Academic settings in Iraq and similar settings. Further applications potentially support the generalizability of the framework and further confirm its validity and reliability as a diagnostic and evaluative tool.

2- Promoting Awareness of Item Clarity and Response Stability in Educational Testing

The study's findings highlight the importance of item clarity in stabilizing student responses, regardless of test item order. This indicates a desirable form of local independence among items, in which one item has a negligible effect on another. For this reason, it is recommended that any experts in psychometrics, curriculum developers, academic advisors, or classroom teachers consider this dimension while designing or reviewing tests. Item clarity is important for more correctly and reliably gauging responses. Also, it has the additional effect of reducing anxiety and measurement error, thereby improving instrument validity. It is important that all stakeholders consistently review item formulations for cognitive coverage and semantic accuracy, especially in standardized testing.

3- Guidance for Researchers in Selecting Appropriate IRT Models

https://doi.org/10.22521/edupij.2025.19.579 Published online by Universitepark Press

Although this study used the RASCH 1PL MODEL, because it was appropriate for dichotomous data and also aligned with my research questions, I want to stress that I do not intend to suggest that I prefer RASCH 1PL MODEL over a two- or three-parameter logistic model. Researchers need to make educated decisions based on the model that best fits their constructs. If there are multiple constructs to analyze, the level of measurement, the population being sampled, and the purpose of the assessment. Within this context, the results of this study can also be considered a comparative baseline study illustrating the strengths of the RASCH 1PL MODEL for criterion-referenced assessments. However, future research on more complex IRT models in analogous and similar educational contexts should also be undertaken in order to increase theoretical and educational alignment.

4- Advocacy for the Use of the Item Response Command Language (ICL) Program in Psychometric Research

Also, given the effectiveness and accuracy of ICL across this study, we strongly endorse its application for researchers who are working in the domain of educational measurement and evaluation. ICL has established itself as a comprehensive software application suitable for conducting Item Response Theory (IRT) analysis, including the initiation of estimates of person and item parameters, the provision of both graphical and textual output, and support for psychometric validation processes. Its ability to calibrate and grade test items along a latent trait continuum consistently makes it a desirable tool for researchers in the field of educational measurement, particularly in standardized assessments. Therefore, we recommend that universities and research centres provide training in the use of ICL to graduate students and faculty undertaking educational assessments, thereby elevating the quality and technical rigor of future measurement strategies.

Declarations

Conflicts of Interest. The authors declare no conflict of interest.

References

Abdel Hamid, J. (1998). *Educational evaluation and psychological measurement* (3rd ed.). Dar Al-Nahda Al-Arabiya. https://iasj.rdd.edu.iq/journals/uploads/2024/12/07/7c27c74b32e554b6661e7b87d9858644.pdf

Abu Hatab, F., Othman, S. A., & Sadiq, A. (2003). *Psychological evaluation* (3rd ed.). Anglo-Egyptian Library. https://doi.org/10.26034/cm.jostrans.2019.171

Al-Sharqawi, A. M., Al-Sheikh, S. A.-K., Kazim, A. M., & Abd Al-Salam, N. M. (1996). *Contemporary trends in psychological and educational measurement and evaluation*. Anglo-Egyptian Library. https://iasj.rdd.edu.iq/journals/uploads/2025/01/02/09eaf25899e6e10962dee76efb07735d.pdf

Allam, S. E. M. (1986). *Contemporary developments in psychological and educational measurement*. Al-Qabas Commercial Printing Press. https://files.eric.ed.gov/fulltext/EJ1131564.pdf

Allam, S. E. (2001). *Diagnostic tests: A criterion in educational, psychological, and training fields* (2nd ed.). Dar Al-Fikr Al-Arabi. https://iasj.rdd.edu.iq/journals/uploads/2025/04/24/8dba8f7fc2e02f15a7c55c199875449e.pdf

Allam, S. E. (2005). *Unidimensional and multidimensional item response models and their applications in psychological and educational measurement* (1st ed.). Dar Al-Fikr Al-Arabi.

https://iasj.rdd.edu.iq/journals/uploads/2025/04/24/8dba8f7fc2e02f15a7c55c199875449e.pdf

Gallagher, J. D. (1998). *Classroom assessment for teachers*. Prentice Hall. https://g.co/kgs/kmSTtyX

Gray, W. M. (1978). A comparison of Piagetian theory and criterion-referenced measurement. *Review of Educational Research, 48*(2), 223–249. https://doi.org/10.3102/00346543048002223

Habeeb, S. T., & Kadhim, B. (2018). *Modern and traditional measurement theories: Principles and applications*. https://www.researchgate.net/publication/381957459_nzryty_alqyas_alhdytht_waltqlydyt_mbady_wttbyqat

Hambleton, R. K. (1995). Comparison of classical test theory and item response theory and their applications to test development: Issues and practice. *Journal of Educational Measurement, 12*(1), 38–47. https://www.researchgate.net/publication/249469819_An_NCME_Instructional_Module_on_Comparison_of_Classical_Test_Theory_and_Item_Response_Theory_and_Their_Applications_to_Test_Development

Harris, D. (1984). Comparison of one-, two-, and three-parameter IRT models. *Educational Measurement, Spring Issue*. https://www.stat.cmu.edu/~brian/PIER-methods/For%202013-02-26/Readings/harris-article.pdf

Kazem, A. M. (1988). *A critical theoretical study on the objective measurement of behavior: The Rasch model*. Kuwait Foundation for the Advancement of Sciences. https://www.researchgate.net/publication/375379963_Differential_Item_Functioning_Analysis_of_an_Emotional_Intelligence_Scale_for_Human_Resources_Management_at_Sohar_University

Kubiszyn, T., & Borich, G. (2000). *Educational testing and measurement* (6th ed.). John Wiley & Sons. https://www.scirp.org/reference/referencespapers?referenceid=2411142

Nunnally, J. C. (1978). *Psychological theory* (2nd ed.). McGraw-Hill. https://www.scirp.org/reference/referencespapers?referenceid=2615575

Popham, W. J. (1978). *Criterion-referenced measurement*. Prentice Hall. https://www.amazon.com/Criterion-Referenced-Measurement-W-James-Popham/dp/0131936077

Popham, W. J. (2014). Criterion-referenced measurement: Half a century wasted? *Educational Leadership, 71*(6), 62–66. https://www.researchgate.net/publication/287234320_Criterion-Referenced_Measurement_Half_a_Century_Wasted

About the Contributor(s)

Safaa Tariq Habeeb, (Ph.D) University of Baghdad / College of Education Ibn Rushd for Human Sciences / Department of Educational and Psychological Sciences
E-mail: safa.tairq@ircoedu.uobaghdad.edu.iq
ORCID: https://orcid.org/0009-0001-8592-3060

Burhan Ismael Ali, University of Baghdad / College of Education Ibn Rushd for Human Sciences / Department of Educational and Psychological Sciences
E-mail: Borhan.Ali2202m@ircoedu.uobaghdad.edu.iq

Publisher's Note: *The opinions, statements, and data presented in all publications are solely those of the individual author(s) and contributors and do not reflect the views of Universitepark, EDUPIJ, and/or the editor(s). Universitepark, the Journal, and/or the editor(s) accept no responsibility for any harm or damage to persons or property arising from the use of ideas, methods, instructions, or products mentioned in the content.*