International Journal of Innovative Technologies in Social Science

e-ISSN: 2544-9435

Scholarly Publisher
RS Global Sp. z O.O.
ISNI: 0000 0004 8495 2390

Dolna 17, Warsaw,
Poland 00-773
+48 226 0 227 03
editorial_office@rsglobal.pl

ARTICLE TITLE	DEVELOPING A UNIVERSITY LEARNING QUALITY ASSESSMENT SCALE
ARTICLE INFO	Habiche Bachir, Bakhta Chettouh. (2025) Developing a University Learning Quality Assessment Scale. *International Journal of Innovative Technologies in Social Science*. 2(46). doi: 10.31435/ijitss.2(46).2025.3297
DOI	https://doi.org/10.31435/ijitss.2(46).2025.3297
RECEIVED	17 February 2025
ACCEPTED	25 March 2025
PUBLISHED	21 April 2025
LICENSE	The article is licensed under a **Creative Commons Attribution 4.0 International License.**

DEVELOPING A UNIVERSITY LEARNING QUALITY ASSESSMENT SCALE

Habiche Bachir
University Center Elbayadh, Algeria

Bakhta Chettouh
University Center of Aflou, Algeria

ABSTRACT

The current study aimed to attempt to construct a scale for evaluating and assessing the quality of education from the perspective of university professors. In the theoretical framework of the construction, we relied on Item Response Theory (IRT), which is considered the modern theory used in scale development. The study found the following:

- The data fit the Rasch model.
- The scale demonstrated acceptable validity.
- The scale exhibited high reliability.

Based on the findings of the current study, it becomes evident that there is a necessity to employ modern measurement theory and its various models to achieve the highest possible level of objectivity in psychological and educational measurement.

KEYWORDS

Learning Quality, Higher Education, Quality Assessment, Educational Evaluation, University Teaching, Assessment Scale Development

CITATION

Habiche Bachir, Bakhta Chettouh. (2025) Developing a University Learning Quality Assessment Scale. *International Journal of Innovative Technologies in Social Science*. 2(46). doi: 10.31435/ijitss.2(46).2025.3297

The Problem:

The field of developing and updating research tools, scales, and tests is one of the vital and fundamental areas that researchers in various psychological and educational sciences cannot do without. It represents one of the most important domains that have contributed to the development of these sciences through systematic and statistical foundations, enabling them to provide methodological basics that can be relied upon in making decisions related to individuals or groups.

Those who follow the movement of psychological and educational measurement, particularly in relation to scale development, will notice a shift in the mechanisms of constructing scales and verifying their validity and reliability. This shift involves a rapid transition from classical test theory (CTT) to the new theory (Item Response Theory, IRT). Most previous research and studies validated measurement tools—i.e., the validity and reliability of tests and scales—according to classical test theory, which is based on specific principles and rules (criticisms have been directed at it, negatively impacting it). In this theory, a test is constructed, and its suitability and adequacy for good measurement are verified by referring to its theoretical assumptions and foundations. For example, to confirm reliability coefficients, one refers to a theoretical model describing the extent to which random errors affect the total test score, known as the true score model (Crocker, 2009, p. 148).

The idea of this model is based on the **obtained score** that an individual achieves in a test, which is often marred by error. If we subtract the **error score** from this score, we obtain the **true score** according to the following equation:

$X = T - EX = T - E$

Cronbach argues that the weakness of classical test theory stems from its oversimplification of the concepts of validity and reliability on one hand, and researchers' mechanical use of these two concepts on the other. This has led to ambiguity in their true meaning and practical application. Even if this statement is somewhat ambiguous, we know that reliability in classical test theory is linked to the results obtained from applying a scale, while the concept of validity refers to a qualitative rather than a quantitative meaning.

Alam (2008, p. 700) goes further, stating that classical test theory in psychological and educational measurement is one of the weak and limited theories (**True Score Theory Weak**) in addressing the concepts of validity and reliability because it does not distinguish between the multiple measurement errors related to one of the test forms used by the researcher under certain conditions. Consequently, methods for estimating reliability and validity vary accordingly.

Mimi Sayed Ahmed cites three main reasons for classical test theory's inability to achieve objectivity in measurement (Sayed Ahmed, 2014, p. 45):

- The theory assumes that test scores representing the trait are a steady linear function.
- Discrimination, difficulty, and reliability coefficients (which are of interest in this research) in classical test theory depend on the characteristics of the sample on which the test is applied.

All these criticisms and others have led measurement scientists to undertake innovative research efforts since the 1970s, accelerating the emergence of modern measurement theory, known as **Latent Trait Theory (LTT)** or **Item Response Theory (IRT)**.

Item Response Theory (IRT) is based on strong assumptions that must be verified in the data to produce reliable results. One of the most prominent and widely considered assumptions by researchers is **unidimensionality**, which means that the test items measure only one trait or ability that explains an individual's performance on the item—in other words, all items measure a single dimension.

Item Response Theory includes several psychometric models that seek to determine the relationship between an individual's performance on a test and the latent trait or ability underlying that performance. According to Alam, these models are probabilistic mathematical functions that vary depending on the number of parameters.

One of the most prominent and widely used models is the **Rasch model (RM)**, which can provide the requirements for objective measurement if its conditions are met. The Rasch model is based on the results of the interaction between individuals' abilities and item difficulty, with the outcomes of this interaction taking the form of observable responses. Through these responses, item calibrations and individual estimates can be derived, fulfilling the requirements of measurement objectivity (Kazem, 1986, p. 43).

Modern measurement theory, in its various models, has become indispensable in research, and it is now imperative for us to adopt it, especially when verifying the quality of measurement tools and striving for greater objectivity. This study falls within this framework, focusing on the variable of **education quality**, which has received increasing research attention and attracted the interest of researchers, particularly in educational sciences and educational psychology, due to its importance in various aspects of life.

In this study, we attempted to employ the **unidimensional Rasch model** to develop a scale for diagnosing the level of education quality from the perspective of a sample of university professors. It is no secret that Algerian universities, schools, educational centers, and even families suffer from a noticeable decline in education quality. However, the Algerian state is making every effort to introduce radical reforms in the education system to achieve the desired quality. A clear example of this is the introduction of English language instruction at the primary level.

Research Questions:

1. To what extent do the data derived from the education quality scale fit the Rasch model?
2. Does the difficulty calibration of the education quality scale items differ when using the Rasch model?
3. What is the validity level of the education quality scale according to the Rasch model?
4. What is the reliability level of the education quality scale according to the Rasch model?

Hypotheses:

1. The data derived from the study scale fit the Rasch model.
2. The difficulty calibration of the study scale items differs when using the Rasch model.

3. The study scale has an acceptable level of validity according to the Rasch model.
4. The study scale has an acceptable level of reliability according to the Rasch model.

Study Importance:

This study is significant for several reasons. It addresses a relatively recent variable in the literature of educational psychology and educational sciences: education quality. Additionally, the diversity of models and theoretical approaches explaining this variable has led to a variety of scales. Each model presents a specific scale and theoretical framework for interpreting education quality, making it difficult to choose the optimal scale.

There are self-assessment scales, which treat the variable as a trait and focus on consistency in the components of the researched variable across different situations, represented in specific behaviors. In contrast, performance scales view it from the perspective of inconsistency in factors surrounding the educational environment. However, as previously mentioned, all these scales were developed in light of psychometric measurement theory.

The study's importance also lies in its focus on a modern trend in psychological and educational measurement (Item Response Theory), which has gained widespread recognition globally. Researchers in psychological and educational measurement have recommended its adoption, whether in developing psychological and educational tests and scales or in adapting and standardizing other tests.

Study Objectives:

Most educational tests applied in the Algerian context were initially constructed based on psychometric theory and were developed in societies different from ours. This raises doubts about the credibility of their results and necessitates their adaptation to our local environment and its specificities. In general, the research objectives can be summarized as follows:

- To draw researchers' attention to the cautious use of scales and not to rely solely on their results, especially when using traditional theory methods.
- To verify the validity of the education quality scale according to the Rasch model.
- To attempt to provide a scale for diagnosing and measuring education quality.

Operational Definitions of Study Terms:

1. **Education Quality:** In the educational context, quality refers to the good type of education that prepares graduates capable of adapting and effectively dealing with modern developments and their various outputs, while fulfilling the requirements expected by all stakeholders inside and outside the university (Louchène Hussein & Magawsi Saliha, 2008, p. 271).
2. **Reliability According to the Rasch Model:** Measurement reliability in the Rasch model is evident through the independence of measurement from the sample of scale items and the group of individuals to whom the test is applied.
3. **Validity According to the Rasch Model:** In the current study, the scale's validity is verified by examining the extent to which the items measure what they were intended to measure through fit statistics, including infit and outfit statistics.

Methodology:

The study adopted a descriptive-analytical approach because it aligns with the nature of the researched topic: employing the Rasch model to develop a diagnostic scale for education quality.

Stages of Test Construction According to the Model:

Al-Shafi'i (Al-Shafi'i, 1996, p. 383) summarized data analysis according to the simple Rasch model in six steps, illustrated in the following figure:

1. Answer Correction and Data Entry
2. Preliminary Analysis
3. Exclusion of Non-Fitting Individuals
4. Secondary Analysis
5. Removal of Non-Fitting Items
6. Final Analysis

Results Analysis:

1. Testing the First Hypothesis:

After preparing the data for preliminary analysis according to the Rasch model by creating a response matrix and examining it to exclude any item answered uniformly by all individuals (a non-discriminating item) or any individual who chose the same response option for all items, no items or individuals were excluded. The data were then ready to test their fit with the Rasch model using the Winsteps program.

For individuals, fit statistics (infit and outfit) were calculated to determine how closely the data aligned with the model—i.e., whether the data derived from the scale fit the Rasch model. In this step, all individuals had fit indices between (0.60 and -1.40), allowing us to rely on 160 individuals for data analysis (no individuals were excluded).

After verifying the fit of individuals' abilities to the model, we then checked the fit of the items to the model using the same statistical method (infit and outfit statistics). Bond (2001) indicates that this statistic should range between (0.60 and -1.40) for acceptable fit limits, confirming that the items fit the Rasch model in the Winsteps program. The results are shown in the following table:

Table 1. Fit Statistics (Infit and Outfit)

Items	Infit Mean Square	Outfit Mean Square	Items	Infit Mean Square	Outfit Mean Square
19	1.23	1.21	39	1.17	1.13
4	1.07	1.09	29	1.03	1.09
22	1.11	1.11	41	1.10	1.15
23	1.01	0.98	31	0.78	0.79
61	1.11	1.11	54	1.02	1.01
24	0.92	0.93	09	1.04	1.00
36	0.89	0.90	18	1.03	1.00
15	0.95	0.90	48	0.88	0.94
08	1.07	1.05	11	1.04	1.03
40	0.83	0.83	17	0.98	0.99
21	0.87	0.91	03	1.02	1.04
34	0.86	0.86	44	0.96	0.96
45	1.02	1.02	50	1.04	1.08
60	1.08	1.10	06	1.06	1.03
42	1.09	1.09	47	1.05	1.03
07	0.97	0.98	05	0.94	0.92
16	0.80	0.83	02	1.01	0.96
38	0.84	0.81	55	0.97	0.92
26	0.79	0.84	51	1.00	0.94
13	0.90	0.91	01	0.97	1.07
46	0.88	0.94	10	1.02	0.96
43	0.95	0.94	63	1.12	1.07
33	1.07	1.08	14	1.04	1.04
35	1.18	1.18	62	1.05	1.02
30	0.94	0.95	20	1.16	1.16
32	0.93	0.92	52	0.99	0.96
28	0.90	0.91	37	1.19	1.14
56	1.06	1.06	59	1.40	1.40
25	0.99	1.02	58	1.19	1.14
57	0.99	1.01			
49	1.00	0.97			
64	1.05	1.08			
12	1.01	1.07			
27	0.94	0.92			

From the table above, we observe that all items of the education quality scale had infit and outfit mean square values within the acceptable fit range (0.60 to -1.40). The infit mean square ranged from (0.78 to -1.40), while the outfit mean square had a minimum value of (0.79) for item (31), which is above the lower limit of acceptable fit (0.60). The highest value was (1.40) for item (59), equal to the upper limit of acceptable fit (1.40).

Thus, based on the infit and outfit statistics, all scale items fell within the acceptable fit limits, allowing us to conclude that the data derived from the items fit the Rasch model. Therefore, the hypothesis is confirmed: the data fit the Rasch model.

2. Testing the Second Hypothesis:

The second hypothesis states that the difficulty calibration of the study scale items differs when using the Rasch model.

To test this hypothesis, individuals' responses were analyzed using the **Winsteps** program by calculating difficulty coefficients measured in logits and standard errors.

Table 2. Difficulty Coefficients of the Scale in Logits and Standard Errors

Items	Difficulty	Standard Error	Items	Difficulty	Standard Error
1	19	0.41	49	-0.02	0.06
2	4	0.40	64	-0.02	0.06
3	22	0.39	12	-0.03	0.06
4	23	0.26	27	-0.03	0.06
5	61	0.25	39	-0.04	0.06
6	24	0.25	29	-0.04	0.06
7	36	0.23	41	-0.04	0.06
8	15	0.23	31	-0.06	0.06
9	8	0.21	54	-0.06	0.06
10	40	0.18	09	-0.06	0.06
11	21	0.17	18	-0.07	0.06
12	34	0.17	48	-0.08	0.06
13	45	0.16	11	-0.08	0.06
14	60	0.14	17	-0.08	0.06
15	42	0.12	53	-0.09	0.06
16	7	0.12	3	-0.09	0.06
17	16	0.10	44	-0.10	0.06
18	38	0.10	50	-0.10	0.06
19	26	0.09	6	-0.12	0.06
20	13	0.09	47	-0.12	0.06
21	46	0.09	5	-0.12	0.06
22	43	0.07	2	-0.14	0.06
23	33	0.07	55	-0.17	0.06
24	35	0.06	51	-0.17	0.06
25	30	0.05	1	-0.18	0.06
26	32	0.04	10	-0.19	0.06
27	28	0.03	63	-0.19	0.07
28	56	0.02	14	-0.21	0.07
29	25	0.00	62	-0.21	0.07
30	57	-0.01	20	-0.22	0.07

From the table above, we observe that the item calibration differed when using the Rasch model. Items numbered (44, 50, 6, 47, 5, 2, 55, 51, 1, 10, 12, 14, 32, 20) were below the mean, with difficulty coefficients ranging from (-0.10 to -0.22).

Hamilton and Swaminathan (1905) note that, theoretically, item difficulty values range between **(+∞, -∞)**, but in practice, they vary depending on the software used to extract them. In the **Winsteps** program, the acceptable range for item difficulty is between (+2). In the current study, the difficulty coefficients ranged from (-0.22 to 0.41), all within acceptable limits.

Amina Al-Kazem (1988) states that for items with average difficulty, the logit score is (0), as is the case with item (25). Items with higher difficulty deviate from zero positively (above average), while easier items deviate negatively (those with difficulty coefficients between -0.10 and -0.22 in this study).

This can be easily visualized through Wright's item-person map, which illustrates the new item calibration:

From the figure (Wright's map), we can easily identify items with difficulty coefficients of (0) logits or close to it—i.e., items of average difficulty, such as (11, 31, 40, 41, 50, 60, 23, 44, 10, 46, 18, 2, 25, 30, 34, 44).

The more difficult items—which, in this study using a multi-response scale, indicate strong agreement—were concentrated at the top of the map, such as items (14, 15, 7).

The easier items—indicating disagreement in this study—were concentrated at the bottom of the map, such as items (55, 8, 26, 9).

3. Testing the Third Hypothesis:

This hypothesis states that the education quality scale has an acceptable level of validity according to the Rasch model**.** To test this hypothesis, we relied on the Rasch model's principal component analysis of residuals using the Winsteps (3.72.3) program to verify whether the scale measures an independent factor more than other shared factors constituting the scale. The results are shown in the table below:

Table 3. Results of Unidimensionality Verification According to the Rasch Model

Component	Eigenvalue	Observed %	Model %
Total Variance of Responses	30.7	100	100
Variance Explained by 1st Factor	18.7	60.9	60.9
Unexplained Variance	12.0	39.1	39.1
Variance Explained by 2nd Factor	1.9	6	15.6

The table above (Table) shows the results of unidimensionality verification using Item Response Theory (IRT), specifically the Rasch model, through principal component analysis of residuals. The variance explained by the first factor was **(60.9%)**, a strong criterion for judging unidimensionality, as noted in the Winsteps program guide (John M. Linacre, 2011).

Additionally, the eigenvalue for the variance explained by the second factor was less than (3), measured at (1.9) in this study—another strong criterion for confirming unidimensionality and thus validating the hypothesis.

4. Testing the Fourth Hypothesis:

This hypothesis states that the research scale has an acceptable level of reliability according to the Rasch model**.** To address this hypothesis, reliability and separation coefficients were calculated for individuals and items, as shown in the table below:

Table 4. Reliability and Separation Coefficients for Items and Individuals According to the Rasch Model

Coefficients	Individuals	Items
Mean	0.32	0.00
Standard Deviation	0.32	0.18
Highest Score	1.23	0.41
Lowest Score	-0.35	-0.49
Separation Index	2.72	2.54
Reliability	0.88	0.87

The table above, extracted to verify the reliability of the current study's tool according to the Rasch model, shows that the reliability coefficient for individuals was (0.88)—a high value indicating that the sample's ability levels were sensitive in distinguishing between high and low levels of the measured trait. The separation index for individuals was (2.72), exceeding the required criterion (2).

For items, the separation index was (2.54), also exceeding the required criterion (2), indicating the hierarchical ordering of test items according to the new calibration.

The item reliability coefficient was (0.87), a high value confirming the adequacy of the scale's items.

Conclusions.

Through this study, we see the necessity of working diligently to overcome the limitations of classical test theory, particularly its normative group aspect, to develop high-quality tests and scales. The properties of classically constructed tests were influenced by group characteristics, leading to less accurate results.

The ultimate goal was to achieve objectivity in measuring behavior or variables, whether related to learning disruptions (negative factors associated with the learning environment), abilities, or psychological aspects. Collaborative efforts led to Latent Trait Theory, later known as Item Response Theory (IRT), which evolved through stages and includes several models:

- The one-parameter model (difficulty and ability),
- The two-parameter model (adding discrimination),
- The three-parameter model (adding guessing).

The Rasch model, one of the simplest and most important models of this theory, was used in this study as an encouraging starting point for developing a scale to diagnose education quality. It is widely used in the Algerian context and has a significant impact—often negative—on individuals and their environment (university, professor, student, family, school).

REFERENCES

1. Ibrahim Mubarak Al-Dosari. (2000). The Reference Framework for Educational Evaluation, 2nd ed., Arab Education Library for Gulf Countries, Riyadh.
2. Ismail Mohamed Al-Faqi. (2005). Psychological and Educational Assessment and Measurement, Gharib Publishing, Cairo.
3. Amina Mohamed Kazem. (1988). A Theoretical and Critical Study on the Objective Measurement of Behavior: The Rasch Model, Kuwait Foundation for the Advancement of Sciences, Kuwait.
4. Anwar Mohamed Al-Sharqawi, et al. (1996). Contemporary Trends in Psychological and Educational Measurement and Evaluation, Anglo-Egyptian Library, Cairo.
5. Saad Hassan Al-Ghamdi. (2003). The Extent to Which Psychometric Properties of Measurement Tools Vary with Response Options and Study Stage—A Case Study of the Likert Scale, Master's Thesis, Umm Al-Qura University.
6. Al-Shafi'i, Mohamed Mansour. (1996). The Effect of Test Score Equating and Normative Controls on Item Calibration Using the Rasch Model, Doctoral Dissertation, Faculty of Education, Mansoura University.
7. Salahuddin Mahmoud Alam. (2002). Educational and Psychological Measurement and Evaluation—Foundations, Applications, and Contemporary Directions, Dar Al-Fikr Al-Arabi, Cairo.
8. Salahuddin Mahmoud Alam. (2006). Educational and Psychological Tests and Scales, 1st ed., Dar Al-Fikr, Amman.
9. Safaa Tariq Al-Habib, Balqis Hamoud Kazem. (2018). Modern and Classical Measurement Theories: Principles and Applications, 1st ed., Dar Al-Manhajiya, Jordan.
10. Abdulrahman bin Sulaiman Al-Turairi. (1997). Psychological Measurement and Educational Evaluation: Theory, Foundations, and Applications, 1st ed., Al-Rushd Library, Riyadh.
11. Abdelaziz Bousalem. (2008). Employing the One-Parameter Rasch Model in Building an Achievement Test in Psychological Measurement and Achieving Objective Interpretation of Its Results—A Psychometric Comparative Study Between Latent Trait Theory and Classical Test Theory, Doctoral Dissertation in Educational Sciences.
12. Abboud Jawad Al-Safi. (2000). Construction and Application of a Learning Strategies Scale for Preparatory School Students, University of Al-Qadisiyah, Wasit Journal of Human Sciences.
13. Ali Maher Khattab. (2005). Measurement and Evaluation in Psychological, Educational, and Social Sciences, 2nd ed., Anglo-Egyptian Library, Cairo.
14. Mohamed Rabie Shatat. (2009). Personality Measurement, 2nd ed., Dar Al-Masirah, Kuwait.
15. Musa Nabhan. (2004). Fundamentals of Measurement in Behavioral Sciences, 1st ed., Dar Al-Shorouk, Amman.
16. Mimi Sayed Ahmed. (2014). Modern Trends in Psychological and Educational Measurement, 1st ed., Dar Al-Kitab Al-Hadith, Cairo.
17. Cecil R. Reynolds, Ronald B. Livingston. (2013). Mastering Modern Psychometric Theories and Methods, trans. Salahuddin Mahmoud Alam, 1st ed., Dar Al-Masirah, Jordan.
18. Bond, T.G., & Fox, C.M. (2001). Applying the Rasch Model: Fundamental Measurement in the Human Sciences, Mahwah, NJ: Lawrence Erlbaum.
19. Crocker, L., Algina, J. (1986). Introduction to Classical and Modern Test Theory, New York: CBS College Publishing.
20. El-Korashy. (1995). Applying the Rasch Model to the Selection of Items for a Mental Ability Test, Educational and Psychological Measurement.
21. Hambleton, R., & Swaminathan, H. (1989). Item Response Theory: Principles and Applications, Boston: Kluwer Nijhoff Publishing.
22. Linacre, J.M. (2012). A User's Guide to Winsteps Ministep Rasch-Model Computer Programs, Winsteps.com.
23. Weinstein, C.E. (1988). Assessment and Training of Student Learning Strategies, New York: Plenum