Designing a reference based test in statistics for students of the Faculty of Physical Education, Sadat City University, using the logarithmic model of the modern theory of measurement

*Dr. Ahmed Rabie Mahmoud Saad

introduction:

- The process of educational and psychological measurement relies on tools and measures that provide quantitative data that allow educators to understand the educational phenomenon in a specific way. Achievement tests are an important part of the educational process, through which it is possible to judge the extent to which educational goals have been achieved. It also helps the student to identify the level of his academic achievement(12:17) .

- By following the developments in the construction of educational and psychological standards, we notice a trend towards the use of the standard-referenced measurement, and in return there is another trend towards the use of the standard-referenced measurement, and the use of the test item response theory. The standard reference scale depends on the interpretation of grades in the light of specific criteria, that is, the student’s scores in the test are attributed to standard scales, and then the relative performance level of the student is determined based on the location of his score compared to the average achievement of the standard group in the test, and this trend focuses on individual differences between students so that It shows the extent of the difference between one student's level and another, that is, it measures the ability of a particular student compared to the ability of other students. The standard-reference measures were subjected to some aspects of criticism, as it was found that these tests depend on comparing the performance of the student with the members of his class group, and therefore it is possible that the location of the student differs according to the characteristics of his standard group, in addition to that the average of the standard group does not necessarily represent the performance required for success(3:54) .

- These tests are also concerned with the individual differences between students without regard to the extent of the student's proficiency in the skills and information to be measured in order to qualify him for new training and educational programs. With accurate information that helps in making appropriate educational decisions about the level of adequacy of the student and the curriculum, and these criticisms played a role in the emergence of modern perceptions in the methodology of educational tests and standards, which led to the emergence of another trend called analogical reference(2:74) .

- The reality that we must not overlook is that education is no longer limited to

Assistant Professor, Department of Fundamentals of Physical Education, Sadat City University *

simply distinguishing between students in the measured characteristic, but rather it must focus on their acquisition of certain skills and achievement of specific goals, and even mastery of skills and information, and thus comparing the performance of the individual with a specific level of performance in a specific field of Behavior This is what spoken tests aim at(11:65) .

- Also, referenced tests are the most suitable types of achievement tests for measuring and evaluating student achievement, because they accurately determine the skills and competencies required to be mastered, as they can be measured and observed directly, which helps in diagnosis, classifying students into proficient and non-proficient, and determining appropriate remedial programs. More detailed information about the performance of the learners, and the emergence of the idea of learning with perfection led to the emergence of the referenced tests, where the movement of the spoken measurement took an important place among the scholars of educational measurement, and with the development of the referenced tests, the variation of defining the concept of the referenced test(63 :5)
- And the test that covers a narrow range and is used to make empowerment decisions, as he defined it as a way to determine the level of the individual in relation to a specific behavioral field.(74 :9)
- The referenced test is a way to verify the learner's acquisition of basic competencies or skills that express specific educational outcomes(4:55) .
- And for the referenced tests, it is concerned with comparing the student's performance with a certain level of performance, regardless of the group's performance(12:54) .
- However, the classification process is based on a comparison of the learner's performance with the predetermined point on the achievement continuum, which corresponds to the so-called cut-off degree that separates masters and non-masters of the measured academic content(2:13) .
- It is a point on the observation scale that distinguishes between proficient and non-proficient persons, and the cut-off degree is used to separate between persons who succeeded in the test and those who did not. The Item Response Theory is one of the recent developments in the field of educational and psychological statistics, as it presented psychometric methods of Great effectiveness in building educational and psychological measures and the method of interpreting the degrees of those scales compared to the traditional theory of measurement(8:77) .
- The theory of responding to the optional item has advantages, including: The estimate of the individual's ability is independent of the sample of the items to which it is applied, that is, the assessment of the capabilities of individuals is free from the characteristics of the items, and the estimates of the features of the items are statistically independent from the sample of individuals that were used to estimate these characteristics, and that it is possible to obtain a statistical (Standard error in estimation) to

estimate the degree of accuracy in measuring the ability for each individual and the statistician differs from one individual to another(4:88) .

- The independence of measurement is the essential difference between the traditional theory and the individual response theory. This theory has made assumptions that must be achieved in the data in order to lead to reliable results. Among the most important assumptions are:

- Unidlmensionality: It means that there is one ability that explains the performance of the individual in the test, so it is called one-dimensional models, while models that assume the existence of more than one ability that underlies this performance are called multi-dimensionalالاستقلال الموضعي. Local Independence**:** That is, the individual's responses to the different items in the test are statistically independent at a certain level of ability, that is, the individual's response to a specific item should not affect positively or negatively his response to another test item.

- Item Characteristic Curve ICC: It is a mathematical relationship linking the individual's probability of responding correctly to the test item, and the ability measured by the group of items that contain that item. It clearly indicates that the probability of answering the item with a correct answer increases with the increase in the ability of the individual, because the curve is cumulative upward, and these curves are described in the models of tests designed to measure one characteristic (one-dimensional) in terms of one teacher, two teachers, or three parameters. (18:71)

- Speed in performance Speedlness: It assumes that the speed factor has no role in answering the item, ie that the individual's characteristics in answering the test items are due to his low ability and not to the effect of the speed factor in his answer.

There is a set of assumptions specific to each of the single response models, as the triple model is the general case for these models, because in it the difficulty parameter (b), the discrimination parameter (a), and the estimation parameter (c) are estimated. As for the unilateral model, only the difficulty parameter (b) is estimated, and it is assumed that the discrimination is equal for all test items, and it is equal to one, and the guess (c) is equal for all test items and is equal to zero. As for the binary model, the difficulty parameter (b) and the discrimination parameter are estimated (a), and it is assumed that the guess parameter (c) is equal for all test items, and is equal to zero, and these assumptions are difficult to achieve, that is, there is a group of items with equal discrimination, equal to one, and the probability of guessing is equal to zero for items of multiple choice type. (16:20)

The individual response theory includes several models, the most important of which are:

Rash Model: It is one of the most common of these models, as it allows the vocabulary to differ in its difficulty, and it is mathematically expressed by the following relationship:

The two-parameter model: allows the test items to differ in their difficulty and distinction, and is expressed in the following relationship:

And the three-parameter model: (difficulty, discrimination, guessing), and this model is used in the current research, and it is represented by the following relationship:

Where denotes the probability of an individual chosen randomly from the level of ability () answering the item (a) correct answer, b: difficulty parameter, a: discrimination parameter, c guessing parameter, ability parameter, D = 1.7 and represents the scaling factor.

The test information function plays an important role in the response theory of the individual, as it is possible to determine the standard error in the estimate, and the test information function is characterized as representing the sum of the information functions of the individual items at a certain level of ability, and the test information function is independent of the sample of the testers, and thus the response theory of the individual offers advantages The error in estimation is inversely related to the square root of the test information function according to the following relationship:

Test information function:

Standard error of estimation:

This means that the standard error in estimation is as low as possible at the ability levels that correspond to the maximum estimate of the information. One of the useful uses of the individual information function is the ability to know the extent to which each item contributes to the test information function independently of the other items of the test. If we know the capabilities of a group of testers, we can Selection of test items that increase the information provided by the test in the extent to which their abilities are distributed in the test (7:24)

research importance:

The importance of this research lies in its endeavor to build a referenced achievement test in statistics that is characterized by objectivity and accuracy in measuring the trait, by creating vocabulary free from the characteristics of individuals so that it becomes valuable in its use in revealing the level of achievement of students of the Faculty of Physical Education in the field of statistics.

The importance of this research also lies in its use of the three-teacher logarithmic model in developing standardized tests, which would open doors for subsequent studies that highlight the practical applications of parametric models in psychological and educational measurement.

research aims :

Constructing a referenced test in the educational statistics course that has acceptable psychometric characteristics according to the three-teacher model as one of the response theory models for the test item.

Research hypotheses :

1. To what extent are the assumptions of the three-teacher model achieved in the current research data?

2. To what extent do the respondents' responses match the three-teacher model?

3. What are the features of difficulty, discrimination and guessing for the test item?

search terms:

Reference Test Criterion:

It is the test that is used to evaluate an individual's performance in relation to a criterion (an absolute level of performance) without the need to compare his performance with that of other individuals.

The Three Parameter Model: It is one of the two-graded test item response theory models, where this model can estimate three parameters of the test item: difficulty, discrimination, and guessing.

Research plan and procedures:

- research community: The research community consisted of male and female students of the Faculty of Physical Education, Sadat City University, who were registered during the academic years 2021-2022.

The research sample:

The research sample consisted of all (1000) male and female students of the Faculty of Physical Education, Sadat City University, the quartet enrolled in the statistics course during the academic years 2021-2022, as the number of male and female students reached (450) and (550) female students.

Data collection methods and tools

The construction of the research tool went through the following stages:

•Determine the purpose of the test.

•Determining the test content by referring to the course description and the appropriate literature. The content was limited to the following topics: basic concepts in statistics, types of tests, building tests, and analyzing test results.

•Formulating the behavioral objectives that are covered by the topics and expected to be mastered by the students.

•The researcher formulated (45) multiple-choice items with four alternatives, one of which represents the correct answer to measure each of the behavioral goals, and some of them required measuring more than one item. In terms of content and knowledge level.

Validity of the scale:

In order to ensure the validity of the content, the test items were presented to five arbitrators from the faculty members at the university who study the educational statistics course, and each of them was asked to express his opinion on: the degree of agreement between the behavioral goals and the main goal of the test and their representation of them, and the extent to which the vocabulary measures the behavioral goal related to it, And the extent to which the test as a whole is consistent with its main objective, and based on their observations, five items were deleted, and the arbitrators unanimously agreed that there are items that fulfill their purpose, and that they are a representative sample of the behavioral field that the test measures. The number of test items was (40) items, (Appendix 1).

Tool stability:

The internal consistency stability coefficient of the test was extracted according to the Cronbach Alpha equation using the SPSS 15 program. The stability coefficient was 0.90. This value indicates that the test has a high degree in measuring the trait.

Search procedures:

•The test was applied in its final form (40) items to students before the end of the second semester of the year 2021-2022, and after completing the application process and collecting the response sheets, the researcher corrected the answer sheets, and the correction was done by giving one (1) mark for the correct answer and a grade of zero (0). 0) for the wrong answer, and the total score of the student that he obtains in the test represents the total number of the test items for which he answered correctly.

•The testers' responses to the test items were dumped into special files in the computer's memory in preparation for conducting the necessary statistical treatments using special programs (BILOG - MG3, SPSS 20) and to answer the research questions.

Presentation and discussion of the results:

The first question: What is the extent to which the assumptions of the response theory for the test item are achieved according to the three-teacher model؟

First: Unidimensionsllty assumption:

This assumption has been verified using factor analysis through the SPSS program using the Principle Component Analysis method. Table (1) shows the values of the latent roots, the explained variance ratios for the first four factors, and the result of dividing the first root by the second factor.

Schedule (1)
The values of the latent root, the explained variance, and the cumulative explained variance of the first four factors

e The first latent root	cumulative explained	Variance	Percentage of Explanable	Latent Root Factor
The second underlying root				
4.741	42.081	42.081	16.832	1
	50.955	8.874	3.550	2
	61.846	5.891	2.967	3
	61.764	4.918	1.967	4

It is noted from Table (1) that the value of the explained variance for the first factor exceeded 20% as an indicator of one-dimensionality, and the result of dividing the first potential root by the second potential root exceeded the value (2) as a second indicator of one-dimensional realization (Hattie, 1985). A one-dimensional hypothesis by graphically representing the underlying roots, using the SPSS program, which is known as a Scree Plot.

Second: the assumption of local independence

Since the assumption of one-dimensionality is achieved according to what was previously indicated, this means that the assumption of positional independence is fulfilled.

The second question: To what extent do the responses of the test items match the three-teacher model؟

To verify the conformity of individuals and vocabulary to the three-teacher model, the BILOG-MG 3 program was used. The results of the analysis showed that all the responses of the individuals were identical to the expectations of the model, with the exception of four students whose numbers were (307, 405, 501, 605), where the chi-square values were statistically significant at the level of Significance (a ≤ 0.05), and in order to verify the degree of conformity of the test items to the three-parameter model, the same program was used, and based on the chi-square index at the level of significance (a ≤ 0.05), which considers the test item to be inconsistent with the model if the probability of this item is less Or equal to 0.05. Chi-square values, significance level, and individual matching status of the three-parameter model were found. Table (2) shows the values of these indicators.

Schedule (2)

Chi-square values and level of significance for the three-parameter model

statistical significance	ka 2 for good fit	The item number	fit statistical significance	ka 2 for good	The number of the item
0.204	3.4	21	0.330	9.4	1
0.002	86.5	22	0.150	12.7	2
0.000	145.8	23	0.165	11.7	3
0.000	148.8	24	0.287	36.6	4
0.236	64.9	25	0.172	30.6	5
0.092	47.4	26	0.120	9.1	6
0.244	11.5	27	0.168	26.6	7
0.176	14.1	28	0.261	15.9	8
0.173	52	29	0.183	9.3	9
0.218	29.6	30	0.296	12	10
0.006	22.7	31	0.760	7.5	11
0.221	8.6	32	0.317	8.6	12

Follow Schedule (2)

Chi-square values and level of significance for the three-parameter model

statistical significance	ka 2 for good fit	The item number	fit statistical significance	ka 2 for good	The number of the item
0.151	10.1	33	0.189	21	13
0.155	26	34	0.278	17.5	14
0.418	12.3	35	0.095	32.6	15
0.000	117.8	36	0.257	32.7	16
0.001	26.2	37	0.171	4.7	17
0.182	11.4	38	0.120	10.2	18
0.090	13.2	39	0.240	4.4	19
0.001	26.2	40	0.261	19.8	20

Statistical significance level (a ≤ 0.05)

Table (2) shows the chi-square test values and their statistical significance at the level of (a ≤ 0.05), and that (33) items out of (40) items are not statistically significant at the level (a ≤ 0.05), which indicates their conformity to the three-parameter model. The vocabulary is: (1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 25, 26, 27, 28, 29, 30, 32, 33, 34, 35, 38, 39). Appendix (1), as for the test items (22, 23, 24, 31, 36, 37, 40), the chi-square value indicated a statistical significance at the level (a ≤ 0.05), and this indicates that these items do not conform to the three-factor model the teacher.

The third question: What are the values of the parameters estimates (discrimination, difficulty, and guessing) of the test and the standard errors in their estimation according to the three-parameter model؟

In order to answer the research question, the parameters of the test vocabulary were found: difficulty, discrimination, and guessing, as shown in Table.(3)

table(3)

Evaluate test vocabulary parameters (difficulty, discrimination, and guess) and standard errors

Standard Error of Gue	)Guess(	Standard Error of Difficulty	)Difficulty(	Standard Error	)(Discriminatio(	Item Number
0.103	0.318	0.327	0.952 -	0.217	1.266	1
0.068	0.292	0.198 -	0.262 -	0.296	1.641	2
0.033	0.213	0.496	2.408	0.429	1.265	3
0.068	0.273	0.204	0.041 -	0.272	1.431	4

Follow table(3)
Evaluate test vocabulary parameters (difficulty, discrimination, and guess) and standard errors

Standard Error of Gue	)Guess(	Standard Error of Difficulty	)Difficulty(	Standard Error	)(Discriminatio(	Item Number
0.092	0.288	0.253	0.978 -	0.235	1.527	5
0.050	0.175	0.129	0.209 -	0.276	1.892	6
0.061	0.452	0.102	1.025	0.275	0.856	7
0.083	0.241	0.292	0.636	0.086	0.551	8
0.038	0.239	0.118	1.761	0.215	0.992	9
0.047	0.169	0.118	0.359 -	0.101	1.180	10
0.039	0.224	0.084	0.175	0.078	1.073	11
0.024	0.293	0.146	1.556	0.718	0.894	12
0.067	0.247	0.204	0.613	0.156	0.778	13
0.039	0.296	0.113	0.882	0.446	1.091	14
0.102	0.283	0.272	1.635 -	0.205	1.431	15
0.049	0.354	0.147	1.006	0.184	0.843	16
0.062	0.479	0.251	0.160	0.188	0.620	17
0.048	0.244	0.371	0.041	0.072	0.471	18
0.063	0.256	0.173	0.280 -	0.289	1.728	19
0.044	0.370	0.137	0.792	0.433	1.889	20
0.049	0.208	0.123	0.008	0.311	1.943	21
0.052	0.314	0.155	0.434	0.431	1.203	25
0.067	0.254	1.596	0.812 -	0.038	0.107	26
0.061	0.594	0.182	0.155 -	0.379	1.933	27
0.045	0.123	1.570	2.457	0.031	0.091	28
0.062	0.149	0.562	1.611	0.149	0.470	29
0.043	0.243	0.110	0.640	0.352	1.918	30
0.027	0.152	0.105	1.341	0.353	1.010	32
0.055	0.281	0.196	0.853	0.305	1.251	33
0.074	0.175	0.257	0.223	0.224	0.928	34
0.050	0.115	0.239	1.125	0.222	0.918	35
0.061	0.163	0.067	1.362 -	0.037	0.104	38
0.012	0.240	0.034	0.748	0.158	1.368	39
0.06	0.27	0.28	0.41	0.23	1.111	المتوسط الحسابي
0.02	0.10	0.35	0.99	0.12	0.54	الانحراف المعياري

It is clear from Table (3) that the value of the discrimination parameter ranged between (0.091 - 1.943) with an average of (1.111) and a standard deviation of (0.54), and that the value of the difficulty parameter ranged between (- 1.635 - 2.457) with an average of (0.41) and a standard deviation of (0.41). 0.99), and that the value of the estimation parameter ranged between (0.115 - 0.594), with a mean of (0.27) and a standard deviation of (0.10).

In light of these results, the researcher recommends conducting more studies on the psychometric properties of the test, which was built using models of other latent features of the test item response theory, such as the Rasch model and others, in order to identify the extent to which the test items match other models of the test item response theory models, and to identify estimates Parameters (discrimination, difficulty, and guessing) of test items and standard errors in the light of these models.

Arabic references:

1- Al-Jubouri Rashid, (2012 AD): Building an achievement test for students of educational administration and supervision according to the theory of latent traits, Teachers Preparation Institutes, Al-Ustad Magazine, (203), 1392-1421 AH.

2- Hijazi, Taghreed and Al-Khatib, Abdullah (2014): The compatibility between the classical theory and the two-parameter model in matching the paragraphs of the spoken reference test in the provisions of recitation and intonation, An-Najah University Journal (Human Sciences), 28 (10), 2241-2266.

3- Hamadneh, Iyad. (2009 AD): Using the Response Theory for the Item in Constructing a Referred Test in Mathematics According to the Three-Teacher Logistic Model, Journal of Educational and Psychological Sciences, University of Bahrain, 10 (2), 215-238.

4- Al-Samarrai, Muhammad and Al-Khafaji, Ahmed (2012): Constructing an achievement test based on the reference in the subject of private psychology, educational and psychological sciences departments, Al-Ustad Journal, (3), 64-1002.

5- Al-Sharifin, Nidal. (2006): Psychometric properties of a standard reference test in educational statistics according to the modern theory of measurement, Journal of Educational and Psychological Sciences, University of Bahrain, 7 (4), 79-110

6- Ababneh, Imad. (2004): The effect of sample size, method of selection, number of paragraphs, and method of selection on the accuracy of estimating vocabulary parameters and ability to test mental ability using vocabulary response theory, unpublished PhD thesis, Amman Arab University for Graduate Studies, Jordan.

7- Al-Azzawi, Rahim Younis, (2008): Al-Manhal in Educational Sciences: Statistics in the Educational Process, 1st Edition, Baghdad: Dar Dijla for Publishing and Distribution.

8- Allam, Salah El-Din, (2001 AD): Diagnostic Tests, Reference Test in the Educational, Psychological and Training Fields, Cairo: Dar Al-Fikr Al-Arabi.

9- Allam, Salah El-Din, (2005 AD): One-dimensional and multi-dimensional test response models and their applications in educational and psychological measurement, Cairo: Dar Al-Fikr Al-Arabi.

10- Ali, Nidaa Bahaa El-Din, (2012): The Effectiveness of Using the Rasch Model in Constructing a Narrated Test Reference for the Statistics Course in Education, Unpublished PhD Thesis, University of Damascus.

11- Al-Ithawi, Ahmed, (2009): Using the Rasch model according to the theory of latent traits in constructing an achievement test in the psychology of individual differences, an unpublished master's thesis, University of Baghdad, College of Education, Ibn Rushd.

12- Al-Nabhan, Musa, (2004): The Basics of Measurement in Behavioral Sciences, 1st Edition, Mutah University, Amman: Dar Al-Shorouk for Publishing and Distribution, Jordan.

13- Foreign references:Allen, M& . Yen, W. M., (1989): Introduction to Measurement theory. California, Broks Cole publishing company.

14- Baker, F.B. (2001): The basics of item response theory (2nd Ed). College Park, MD: ERIC Clearing Honse on-Assessment and Evaluation.

15- Biddle, R. (1993): How to Set Cutoff Scores for Knowledge Tests Used In Promotion, Training, Certification, and Licensing. Public Personnel Management, 22(1), 69-73.

16- Brannick, M. (2003): “Basics of IRT one - linefile”: //a:/item response theory.htm.

17- Crocker, L& Algina, J. (1986): Introduction to classical and modern test theory. New York: Holt, Rinehart and Winston.

18- Hambleton, R & Swaminathan, H. (1985): Item response theory principles and applications. Boston: Nijhoff Publishing.

19- Hambleton, R, K. Swaminathan, H& . Rogers, H. (1999): Fundamentals of item response theory. Sage Publication. Newbury Park, CA.

20- Hattie, 1985. J. Hattie, Methodology review: assessing unidimensionality of tests and items. Applied Psychological Measurement 9 (1985), pp. 139-164.

21- Klein, M & Velden, U. (2009): Intuition who will pass dental OSCE? Comparison of the Angoff and borderline regression standard setting methods. European Journal of Dental Educational, 13(1). 162-171

22- Lord M.F. (1980): Application of item response theory to practical testing problems, Hillsdale, NJ. Er/ Baum.