BALTIC JOURNAL OF LAW & POLITICS
A Journal of Vytautas Magnus University
VOLUME 15, NUMBER 2 (2022)
ISSN 2029-0454

Cite: *Baltic Journal of Law & Politics* 15:2 (2022): 1332-1357
DOI: 10.2478/bjlp-2022-001086

Effect of outliers on standard characteristics according to item response theory

Karrar Ghalib Jader Altaweel
Prof. dr Mohammed Anwar Mahmoud
saidkarar506@gmail.com
Baghdad University/Ibn Rushd College of Education for Humanities

Received: August 11, 2022; reviews: 2; accepted: October 12, 2022.

Summary

The current research aims at the effect of extreme values on the standard characteristics according to the paragraph response theory . To achieve this goal , the researcher followed precise scientific steps . The researcher adopted a test Abstract inference prepared by the educational institution (pmteducation) in (2019), and followed the procedures for preparing it according to the paragraph response theory (the two-parameter model), as the researcher verified the translation validity procedures for the test.

And after completing the translation validity procedures , and to verify the descriptive validity of the test, which consists of: (55) paragraphs , The researcher presented the test to a group of arbitrators specialized in psychological sciences, measurement and evaluation, their number is (14) arbitrators , and no paragraph of the test was modified or excluded because it obtained a percentage of agreement (100%), and thus the descriptive validity of the test was confirmed.

To verify the clarity of the test instructions and paragraphs The test was applied to a sample of (10 0.). (A male and a female student chosen at random from the middle school students - for the fourth and fifth grades) scientific and literary) And it became clear from this experience that the instructions and paragraphs of the test are clear and understandable.

For the purpose of statistical analysis of the paragraphs of the abstract inference test and extracting its standard characteristics , the test was applied to a sample consisting of (1000 A male and female student from the fourth and fifth grades (scientific and literary) of Najaf Governorate, were chosen. b , by random sample method , and thus the researcher obtained a file containing (1000) participants in the case of (extreme values), and (903) participants in the case of (absence of extreme values), i.e. after deleting the (97) extreme values , (34) of

which are a high extreme value and (63) a minimum extreme value for the purpose of measuring the effect of those values on the standard characteristics (honesty, difficulty, discrimination, stability) according to the traditional measurement theory and paragraph response theory.

The test was prepared according to the paragraph response theory, specifically (the two-parameter model), which contains (55) paragraphs with three 3 alternatives and some paragraphs with four 4 alternatives in its initial form, and after verifying the factor analysis, one paragraph 1 of the sequence (3) of the test paragraphs And two paragraphs of the sequence (25, 43) were tested for fit to square Ka2, and the number of test items became (52) in its final form and adopted in the final statistical analysis and for the two cases (the presence of extreme values) and (the absence of extreme values) according to the paragraph response theory.

The statistical analysis according to For the paragraph response theory , it was done through the researcher's dependence on the two-teacher model as one of the paragraph response theory models in analyzing the paragraphs of the abstract inference test, as calculated by the Bilog -mg3 program The coefficients of (honesty, difficulty, discrimination, stability) of the paragraphs were calculated, and the extent of the influence of extreme values on them was measured, once (with the presence of extreme values) for (1000) participants, and another (without the presence of extreme values) for (903) participants.

To achieve the model assumptions, the researcher followed the following:

To verify the one-dimensional assumption, the researcher subjected the test to factor analysis using the basic components method, as one meaningful factor was obtained for the test. 1), and the adoption of a saturation percentage (0.30) to accept the saturation of each paragraph of the tests with the general factor , and (3) paragraphs of the sequence (3, 11, 15) were deleted in the case of (the absence of extreme values) because the saturation of the paragraphs was less than (0.30), and the second indicator is the relationship of the paragraph's degree with the total degree to verify one-dimensionality.

In addition to matching paragraphs to a model Two-parameter is evidence that the items measure a one-dimensional trait , based on the value of the chi-square at the level of significance (0.05) And as calculated by the program, and accordingly it was done exclusion A paragraph introduced by the program, and two (2) because it is a function according to the value of the chi-square, and the analysis was re-analyzed again and the paragraphs whose difficulty exceeded (- 2.5) to (+ 2.5), as well as the paragraphs whose discrimination coefficient exceeded (5, 0) to me (5, 2 And its number was (3) paragraphs The validity of the test was verified through (descriptive honesty, functional honesty), and the reliability of the test, and in light of the results of the current research, the researcher reached some conclusions, recommendations and suggestions.

Problem of the research

Studies have indicated that the presence of extreme values in a set of data may lead to an inflation rate of errors, as it increases the error of the second type, and thus reduces the power of statistical tests, and leads to significant distortions in the estimates of statistics and parameters, whether when The use of his teachers ' tests or tests of his teachers, and is likely to lead to a bias in the estimates of interest (Salama and Ibrahim, 2017; Nasrollahzadeh & Koramaz, 2020).

The extreme values are located at both ends of the distribution, so the low extreme values lead to a decrease in the values of the statistical parameters such as the mean and the variance, while the high extreme values lead to an inflation in the values of these overestimation parameters. Liu, 2005) in (Al-Dawy, 2011; Ottuh, 2020; Rygielska, 2020).

The researcher must know how to influence these values in appropriate ways and not ignore them in the data, and there are a number of studies that have been conducted to detect extreme values in the data, but it is still not clear the nature of the impact of extreme values on the standard characteristics (honesty, difficulty, discrimination, stability) according to the theory Responding to IRT , so the problem of the current research is determined by answering the following question:

"What is the effect of extreme values on standard characteristics according to the paragraph response theory?" And within the researcher's knowledge, there is a dearth of studies and research that dealt with this aspect.

Research Importance

The Significance of the Research

Measurement is not limited to estimating material things, but rather includes estimating psychological properties and characteristics that can be measured and represented quantitatively in numbers. Indirect methods are often used to describe what the thing contains of this characteristic, then it is expressed in numbers. i.e. subjecting the phenomenon to a quantitative estimation by using standardized digital units or agreed- upon units garage , 1997 : 97-98 A phenomenon can be accurately understood by using objective, honest and consistent tools and methods that result in data and information that increase our knowledge and understanding of this phenomenon, and these data and information are used To make practical decisions related to the phenomenon in a better way (Abu Nahia , 1994: 17).

There is no doubt that mental abilities are one of the important characteristics in the human personality, and psychologists have focused great

efforts to study them and develop mental measures to measure them. 2005.[1](7)

and that for cause Preparation the exams Mental and mentality status Distinguished in area Science self in general and measure psychological particularly command that they call to me Preparation building Tests Psychological and mentality trusted in its efficiency you use in fields different , and if we moved to me area Measure capabilities mentality We will find that this the field Represent to forbid the corner in measurement Psychological (Abu wood 2011 : 321).

as you play theories Psychological and educational measurement a role Whatever in Processes the college for methodology search And that by asking Techniques General in Procedures measurement or Measure Pain changes that care out , and that From Yes a test sensitivity and accuracy modalities measure that It was completed developed (Crocker and Gina , 2009: 30).

and for a purpose reach to me higher Accuracy in measurement psychological and educator then specialists in measurement and researchers are invited to me development Techniques and tools measurement picture ongoing and benefit From theories contemporary and techniques Modern (Abdel-Dayem, 1973 : 21).

and that Feature the most Importance in theory response for paragraph (IRT is _ independence Estimation Statistics Paragraph About capacity people Users to find this is Statistics , as well independence Estimation Capabilities people About Milestones vertebrae second hand in proces Appreciation, has Launched this is the theory From Possibility Forecasting perform the individual on me a test a certain group From factors or adjectives function sports growing steady , Can From through it Determine Prospect the answer correct About Paragraph for levels different From capacity (Embretson & Hershberger, 1999: 44).

has did experts in measurement enriched this entrance , and make it More objectivity by theory response for paragraph (IRT), gesticulate Back Of which From Forms logarithmic Multiple , Might From the most important and most of them frequently used in area the exams achievement she models logarithmic probability mono The dimension , And the fit models (mono), the couple , triangular) landmarks with vertebrae that be the answer on her bi staging meaning the answer About Paragraph As for given degree (zero), or given Degree (one) (mark, 1995: 15).

It features model couple the teacher with characteristics sports make it more frequently used in Applications psychometric From form Synthesis temperate, presumably in form logarithmic couple the teacher as such he is adverb in Most Forms response for the paragraph non affected the answers in general Guessing (Al -Najjar, 2010: 320).

The efforts of researchers in building standards and tests focused on extracting the psychometric characteristics of the test and the paragraphs related

[1] children of a. h . , & Jassim A. c. (2019). Psychometric features of inductive reasoning among university students according to the item-response-theory theory . Professor's Journal for Humanities and Social Sciences , 58 (4) , 273–298. https://doi.org/10.36473/ujhss.v58i4.1006

to the validity and reliability of the tests, the difficulty of the paragraphs and their distinction, and their importance in that (Bani Atta, 2018: 158).

And the interest in the issue of extreme values has increased because they are usually treated with disregard, and this leads to estimates with less efficiency and bias in the results and their impact on them, because the presence of extreme values that differ fundamentally from the rest of the data leads to significant differences. And fundamental changes and distortion in the results representing the sample data, and therefore it is necessary to research and verify these unusual observations and search for ways to control their impact. (Al-Atyan , 2018: 1485).

Research Objective: Aims of the Research

The current research aims to (measure the effect of extreme values on standard characteristics according to the paragraph response theory).

Search Limits: Limits Of The Research

1- students stage prep class (the fourth and fifth) and from both Both sexes (males and females) The two specializations (scientific and literary) for schools Affiliate for directorates the public to breed Governorate An Najaf, for the year Academic year (2021-2022) _ .

2- The abstract inference test prepared by the educational institution (pmteducation) in (2019), which is intended for the age group (16-17 years), and whose number of paragraphs (55) has 3 alternatives and some of the paragraphs have 4 alternatives, binary Degree Gives a correct answer (1) and a false answer (0).

3- The two-parameter model Paragraph response theory .

4- Transactions (honesty, difficulty, discrimination, stability) .

5- The Tukey method) to detect outliers.

6 - Programs computational used Which program (SPSS) and a program Statistical analysis (BILOG MG 3).

Definition of Terms :

First : the extreme values : Outliers

Known by:

- **Barnett & Lewis (1977) :**

A value or a set of values appears heterogeneous and illogical if it is compared with the data set and other values and is separate from it (Barnett & Lewis, 1977: 9).

- **Dan & Ijeoma (2013)**

That value that has a standard remainder is relatively large compared to the rest of the values in a set of data (P. 308 : Dan & Ijeoma, 2013).

Standard Properties: Psychometric Properties

- **Zekri (2009):**

Characteristics associated with the same scale or test , which can be expressed in numerical terms , both those characteristics related to paragraphs Or those characteristics related to the overall score of the scale or test, according to the traditional measurement theory : (Difficulty parameters , discrimination for the paragraph ( ) , and the validity and stability of the scale or the test) , According to the theory of response to the paragraph) : Difficulty parameters , discrimination , and guessing for the paragraph ( ) , and the validity and reliability of the scale or test ( ) zakri , 2009 : 15th) .

Item Response Theory (IRT) :

- **Hambleton 1995) Hambleton : _**

This theory assumes that the performance of individuals can be predicted or their performance in a particular psychological or educational test can be explained in light of a characteristic or characteristics of this performance called Traits , or in other words, the presence of one or more of the basic characteristics or traits. that determine an individual's observed responses to a test item These features are not directly observed However, it appears through the performance of individuals, which can be observed and measured directly through a set of test items (Hambleton, 1995: 6).) .

Theoretical Framework:

Extreme Values

looks at extreme value as defined by Johnson and Wichern (Johnson & Wichern, 2007) as a value that is inconsistent with the rest of the data, while Barnet and Lewise (Barnet & Lewise, 1994) defined it as a value or more that is inconsistent with the total data set so that it appears far from the rest of the data, whether it is a large or small value , and they describe two types From the extreme value , the first type: polluting value (Contaminant Value , a value that comes from a different distribution, and the second type An extreme value , which is either a large or a small value, but of the same distribution While Evans (Evans, 1999) considered the extreme values to be unusual values that could lead to negative changes in the results of the statistical analysis , from the above it is clear that the extreme value is a value that deviates from the characteristic pattern of the data set , and Barnett pointed out and Lewis (Barnet & Lewis, 1994) that the data outliers are due to either computational errors, reading errors, or recording errors, while Green (Green, 1976) indicated that outliers in the data set may appear because the data belong to the distributions asymmetrical, meaning that it may have a high torsion either towards the right or towards the left , while Hawkins attributed (1980 , Hawkins (extremism of the data on the grounds that the data

comes from two distributions, one of which is the primary distribution) distribution Basic which generates good data, while the other is distribution contaminating which generates extreme values (Bani Atta, 2018: 158)

Reasons for the emergence of extreme values

There are a number of reasons that lead to the appearance of extreme values in a set of data, and these reasons are:

(Dan, Ijeoma, 2013 Osborne & Overbay, 2004 ; Lin & Zumbo, 2007)

1- Data errors: They occur due to human errors such as errors when collecting, recording, entering or preparing data for analysis.

2- Unexpected measurement errors from individuals: they occur due to the respondents' use of guesswork, or the lack of desire among individuals to respond due to fatigue, negligence and lack of attention, or poor response by individuals due to lack of understanding of instructions or with the aim of thwarting the research.

3- Wrong distribution assumptions: False assumptions about the distribution of data can lead to the emergence of extreme values.

4- Extremist cases in the original community: The permissible sample size from the community plays an important role in the emergence of extreme values, as the larger the size of the data set, the greater the possibility of withdrawing more samples from the community, thus increasing the possibility of the emergence of extreme values.

5- Failure in the calibration process: Extreme values may appear due to errors in the research methodology, such as extreme things happening during the application of the study to individuals .

6 - Errors in proces Selection samples: Possibility Response some individuals in the form of Different About Rest Individuals the sample (Ibrahim, 2016: 13-14).

and detecting extreme values in the data:

1. Tukey 's method for detecting outliers:

Al-Bayati and Dagha (1996) reported that (Tukey) was the first to discover methods for analyzing statistical data and methods for detecting and estimating anomalies.

Tukey mentioned that one of the most important and best graphical methods is the box and plots method with the five labels, which depends on the drawing in detection and the statement of extreme observations, their identification and diagnosis in the case of one variable, and the five labels include: (median value M, minimum quartile value Q1, the highest quartile value is Q 3, a lower anomalous value (E 1) and a higher anomalous value (E 2).

The data can be represented through a box , where it is found from the bottom of the lowest quartile, and from the top of the highest quartile, and the

median is between these two quartiles, and it is used to identify and detect outliers (Al-Bayati and Dagha, 1996: 45).

The researcher has adopted the Tukey method , which is the box-graph method to detect extreme values because most researchers and previous studies have adopted it, and because it does not depend on the arithmetic mean and standard deviation, and it is a powerful method when dealing with large data, and it can be used to make comparisons between number of aggregates.

2- Standard Deviation:

It is considered one of the simplest and one of the simple traditional methods for checking outliers in a set of data, where the values can be considered outliers if they are outside the intervals:

$$2SD\ Method: \bar{x} \pm 2\ SD ------$$
$$3SD\ Method: \bar{x} \pm 3\ SD -----$$

Where: ($\bar{X}$: arithmetic mean of the data set), (SD) standard deviation of the data set

3- Z-Score:

It is another method that can be used in order to check the extreme values in the data, by tracking using the mean and deviation, and the basic idea of this rule is if (x) follows a standard normal distribution, then (Z) follows a standard normal distribution, where if the absolute value of the standard degree For observation exceeds the value (3) Observation is considered an extreme value, and this method is considered simple when the data follow a normal distribution, and the standard score is calculated as follows:

$$Z_{i=} \frac{x_{i-} \bar{x}}{S} -------$$

Where: (Z_i istandard score), (X_i iobservation), ($\bar{x}$ arithmetic mean), (S). Standard Deviation (Seo, 2006: 36).

Methods of dealing with extreme values:

After the extreme values have been identified , detected and diagnosed in the data, and the reasons for their anomalies and types are explained, they are treated in one of the following ways:

1- How to delete

Qassem and Ismail (2008) see that the emergence of extreme values in the data set greatly affects its analysis, and most of the time the outliers in the data must be deleted and rejected, and stated that it is necessary to know "Are there extreme values in the data? And disclose them." Because detecting the presence of outliers sometimes has a positive effect, especially in some medical circles and analyzes, the presence of outliers may indicate cancer cells or a specific disease.

mentioned (Rahman and AL Amri, 2011: 41 .). It deletes outliers in the data after it is diagnosed and detected in the data, then statistical analysis is done on the rest of the other data, in order to improve and develop the accuracy of the estimation coefficients and increase them better.

2- Maintaining extreme values

Al-Mukhtar (1980) suggested that there are a number of possibilities that are used to deal with outliers in the data, including keeping and maintaining outliers in the data set, because their presence in the data is of great interest and importance and has an impact on the results of statistical analyzes, which may be a reason for explaining phenomena studied, meaning that its presence may not indicate any bad condition (Al-Mukhtar, 1980: 22).

3- Replace the outliers in the data

Where the outliers in the data are dealt with using appropriate statistical methods, and outliers are replaced by different substitution methods, the most important of which are the following :

A - Trimmed Mean:

Al-Fattal and Intranik (2009) indicated that the truncated arithmetic mean method is one of the best methods used to deal with outliers and replace them in the data set, and this method is characterized by efficiency and accuracy (Al-Fattal and Intranik, 2009: 64).

b- Winsorized Mean:

Narrated by light (2010 The method of treatment through the compensatory arithmetic mean is an arithmetic mean of a set of data in which the outliers were estimated by a value close to it instead of being treated by deletion (Al - Nour, 2010 : 10).

C - Increasing the sample size:

Al Nuaimi noted 2012) that it is possible to rely on increasing the sample size to treat outliers in order to avoid excluding outliers on the one hand, and to obtain good and realistic results that serve studies on the other hand (Al-Naimi, 2012: 32).

inference in theories configuration mental

Prepare Inference Skill essential to acquire knowledge and made up fundamentally for intelligence (Goswami, 1991) (Hosenfeld & Vender Maas and Venden boom 1997: 530) , can say that inference Represent to forbid the corner

in intelligence the humanitarian has use it spearman as one Indications important for intelligence general From During measurement or the acting , acting in the sense logical he is judgment on me something a certain as What to have this is Adjective itself in something else a certain similar for him in characteristic or characteristics other (complete strings Numbers or letters, problems Classification) (Oils , 1995: 292) .

Abstract Inference

dating back Tests inference the abstract to me Search make it Scientist self (Charles spearman) in twenties horn twenty, Lost use spearman Technique Statistic it's called Analysis amali (Factor Analysis) from Yes scan Relationship between grades that Gets on her individuals in Tests different, infer that Whose they do perform Good in some Tests intelligence they achieve consequences good also in Tests Other (as tests vocabulary, or Maths , or Capacity spatial), and the the opposite then Whose they achieve consequences weak in a test a certain From Tests intelligence inclined also to me Investigation consequences weak in the exams The Other (Paul, 2009: 3).

justifications Use Tests Abstract Inference

Supporters of abstract (nonverbal) reasoning tests believe that spatial tests measure the inferential ability of an individual in his abstract form away from the influence of any previous knowledge, which are known as tests that are far from cultural influences, meaning that they are free from any cultural bias because they do not give any additional advantage to the individual Who belongs to a particular culture without others who do not belong to the same culture, that is, these tests exclude any factors related to language or other skills that may be associated with a culture without others, and these tests are designed to measure the individual's ability to deal with problems according to a systematic analytical method Contributes to the development of individuals' ability to infer, and the use of verbal tests and scales with individuals who suffer from problems in using language leads to a misjudgment of the abilities of these individuals (Carter, 2005: 38).

Item Response Theory (IRT)

to get to know theory response for the paragraph in measurement Basim Features latent for her interest by connecting between Response the individual for a paragraph a test self Properties certain And his ability (Al-Shafi'i, 2008 : 43) .

the paragraph response theory seeks to achieve is:

response theory aims to reach objectivity in behavioral measurement, similar to standards and tests Used in the natural sciences, this requires two things: **first:** that the difficulty of the paragraphs does not depend on the characteristics

of the sample of respondents to whom the test was applied.

second: that the test gives an estimate of the performance and does not depend on the difficulty of the items included in the test (Allam, 1986: 118).

The main assumptions on which the paragraph response theory is based

Assumptions of Item – Response Theory

first assumption: (one-dimensional: Unidimensionability)

one most important Assumptions the basic for the theory measurement he is that Collection From vertebrae that form a tool What all of which share in Measure something One Just , and this is Assumption Presents the foundation for most Forms measurement sports , then Application theory response Paragraph (IRT) for problems measurement, Complete Procedure Assumption common that there Worker One dominant or Ability that Can that represent the performance on me Paragraph , and this is What mean with it assumption mono Dimension (Unidimensionality) , meaning that Trait or capacity one Just measured by vertebrae , and that for the theory response for the paragraph model to describe or Clarification Relationship between Response the individual Examined and the variable hidden latent (often What he is called by the ability or attribute) be measured with a tool measurement, has is being variable latent Which phenomenon described behavioral with (Theta θ) , and (θ) compound single The dimension continuous that Describe Heterogeneity between responses the paragraph (Habib and Aziz, 2018: 75).

second assumption: (Local Independence)

Lord (1980) points out that local (local) independence is a definite consequence of one-dimensionality. If the test items measure one feature and it turns out that the respondents' answers to the test are one-dimensional, then it is evidence of the fulfillment of the assumption of local (local) independence. (Lord, 1980: 19).

third assumption: (Assumption of the paragraph characteristics curve: Item Characteristic Curve)

And that the characteristic curve of the paragraph (ICC) is a non-linear regression of the degree of the paragraph on the ability or characteristic measured through the scale or test, and the difference between the models of latent features is due to the different forms of their mathematical functions, and then the different shapes of their curves, and this difference is a mathematical function. Correlation between the probability of an individual 's success in answering the item and the ability measured by the test or the set of items that the measure includes (Allen

& yen, 1979: 128).

Fourth assumption: **(The speed of performance in the test: speediness)**

paragraph response theory focuses on the accuracy of the answer and neglects the speed factor, and its models impose that the speed factor does not play an important role in answering the paragraph, so when the paragraph is not answered due to the speed factor and the tightness of the test time, the individual's performance on the test is not a function of the latent feature (Tinsley & Dawis, 1975: 326) .

The main types of paragraph response theory models

1 - A two-parameter logarithmic model (Birenbaum's model)2 (difficulty and discrimination)

Two- Parameter Logistic Model (Birnbaum Model) (2-PL)

(Crocker and Aljina, 2009) mentions that it is more appropriate to start with the anagram - the dual- teacher model because it is the most representative of the anagram - the natural model (Crocker and Aljina , 2009: 466), and promises this form more models likeness by theory measurement traditional From Where Properties psychometric as such pointed out Allam (1990) , As for it's a more in agreement with Indications Classic (traditional) from models the other (Hernandez, 2009: 13), in addition to to me priority this form in selection paragraphs , and in this form Different vertebrae the test in its difficulty and distinguish it Just , in when Supposedly form non affected responses factor Conjecture (Pelton, 2002: 118).

2- The A one-parameter logarithmic model (A Rasch model) (difficulty)

Logistic Model (Rash Model) (1-PL) One-parameter

This is a one-parameter logarithmic model known as a Rash model .) is the most widely used among the paragraph response theory models (Harris, 1989), and it is considered a special case of a Birnbaum model . The parameter is two-way logarithmic. In this model, it is assumed that the discrimination parameter and the measurement line parameter are constant , meaning that all the items

[2] Through reviewing the literature and many studies related to the field of contemporary educational and psychological measurement, the researcher found that there is a difference in the names of the binary and triple models with the names of scientists. There is a lot for both of them in developing these two models, but the researcher has relied on defining the name of each of the binary and triple models according to their names mentioned in the current research depending on the source (Response models for one-dimensional and multi-dimensional test item and its applications in psychological and educational measurement) by the author Professor Dr. Salah El-Din Mahmoud Allam, first edition, year 2005 - Arab Thought House (source is included in the list of sources).

distinguish to the same extent between the individuals, but they differ only in their difficulty (Al-Najjar, 2010: 218).

3- The A three-log logarithmic model of the teacher (Lord model) (difficulty, discrimination and guesswork)

(3-PL) Three – Parameter Logistic Model

It's called Pal a triple- parametric logarithmic model (3-PL) where he added a third parameter in the probability of individuals reaching the correct answer , and he is a teacher of guesswork as well as a teacher (Difficulty and Distinction) Thus, this must be taken into account when matching the data from the test to this model (Hambleton & Swaminathan, 1985: 37-38).

4 - The A logarithmic model Quad j the teacher

(4-PL) Four-Parameter Logistic Model

McDonald's Foot (Mcdonald, 1976) and Barton and Lord (Barton & Lord, 1981) describe the mathematical formula for this model, suggesting a solution to a problem facing estimating the true values of the responses of some tested individuals. Or the respondents on the test items, as sometimes the examinees with high ability do not answer the test items correctly, and this may be due to their lack of interest, or that they have information other than what is supposed to be measured by the test builder or preparer , and therefore they choose The answer that does not agree with the answer of the correction key , and this model differs from the three-teacher model in adding a fourth parameter that represents the inability of the tested individuals or Respondents with high ability to answer the test items, and here the value of the teacher is less than one (Hambleton & Swaminathan, 1985: 48) .

Standard characteristics (ideometric) of test items according to item response theory

1- Difficulty Parameter (β)

Paragraph difficulty is a point on the continuum of the trait at which we expect the probability of an individual to respond correctly to it equal to (0.50) (And without guessing), that is, estimating the difficulty of the paragraphs is a function of the number of individuals who answered correctly about the paragraph and the ability of these individuals (Kazim , 1988, 137). , The difference in the test items usually gives different levels of difficulty, this can be demonstrated graphically , as shown in Figure (14) below , which represents the level of possession of the trait) or the sum of the scores obtained (along the horizontal axis) sigmoid (And the probability of getting the correct answer for a paragraph on the vertical axis) y) in three test items (Coaley, 2010: 39) .

2- Item Discrimination: (a) Item Discrimination

Clause discrimination in item response theory is defined as the rate of change in the probability of the correct response of the individuals tested or responding to the item relative to the ability level (Al-Sharqawi and others , 1996: 342) , meaning that the paragraph whose degree of discrimination is high means that the percentage of those who answered it correctly from the members of the upper group is greater than the percentage of those who answered it correctly from the members of the lower group (Murad and Suleiman, 2002: 218), The value of (a) is determined theoretically m n (□- to □+) , and in practice it is less than or equal to (0.2) Hambleton and Swaminathan (1985) mentioned that the practical value of a) It can range from (0.4) to (2) (Al - Zahrani, 2008: 18) .

Standard characteristics (ideometric) of the test as a whole according to the paragraph response theory

First : the truth the test : Validity Test

The concept of honesty for the referenced test is not very different from it for the reference test except in terms of the nature of the purpose for which it is designed, and the lack of great difference does not prevent us from addressing the types of validity of the reference test, and here we can talk about three types of honesty as follows:

(Descriptive honesty, functional honesty, behavioral range selection validity).

1- Descriptive Validity

This type of validity has several names, sometimes it is called "content validity " as in standard-reference tests, or " Instructional Validity ", but Yalow , Popham and Linn do not prefer this last name because the concept of descriptive validity It is more general, and this type of honesty is the basis for other types of honesty. Allam, 2001: 281) .

2- Functional honesty:

This type of validity corresponds to what is known as empirical validity in reference standard tests, and despite the diversity of functions and uses of referenced tests, which are of interest to evaluation experts, some of them do not include prediction of specific criteria, so the concept of functional validity is more applicable to tests Spoken – Marji[3] (Popham, 1978: 143 - 144)

[3] Kazem A. B. h. (2020). Verifying hierarchical reference-based assumptions of criteria for higher order thinking skills using item response theory. Professor Journal of Humanities and Social Sciences, 59 (1), 19-44. https://doi.org/10.36473/ujhss.v59i1.1050

3- Domain Selection Validity

The behavioral range that we choose from among the other behavioral ranges must allow us to generalize on the comprehensive range of the dimension measured by the test. Allam, 1995: 21).

Second: Test stability: Test Reliability

Reliability in psychometrics means the accuracy of the test in measurement or observation and not contradicting itself its consistency and its exclusion in what it provides us with information about the respondent's behavior) Abu Hatab , and others , 2008 : 135) .

The researcher has adopted the reliability methods specified in the study (Al-Sudani, 2016), as follows:

1- first method

First: a laboratory that requires the application of the same test to the same sample.

- Livingstone coefficient:-

$$K(X,T) = \frac{Q^{2(T)} + (M-C)^2}{Q^2(x) + (M-C)^2}$$

- Harris coefficient: -

$$\mathrm{MC} = (\mathrm{ssb}\ )/(\mathrm{ssb} + \mathrm{ssW})$$

- Kappa-Spikoviac coefficient: -

$$SK = \frac{SP\ - Pc)}{1 - Pc}$$

- Sabkoviac's coefficient of agreement:

$$P_c^{(i)} = \frac{\sum_{i=1}^{n} P_c^{(i)}}{N}$$

- Haina's kappa coefficient:

$$K = \frac{P_Z - P_{ZZ}}{P_Z - P_{ZZ^2}}$$

Second: A parameter that requires the application of two parallel images:

- Carver factor

$$= \frac{A + C}{N}$$

- Hambleton – Novec Po. agreement coefficients

$$po = \frac{A + B}{N}$$

- Kappa coefficients for Swaminathat, Hambleton and Gania.

$$Pc = \frac{(A + c)}{N} \times \frac{(B + D)}{N}$$

2- The second method:

- Contrast Ratio Index pVariance Ratio Index according to the following equation:

$$R = \sigma^2_T / \sigma^2_O$$

- Information Function Index according to the following equation:

$$R = 1 - (SEE)^2 \ldots\ldots or \ldots\ldots R = 1 - (1/I(\theta))$$

- Separation Coefficient Index according to the following equation:

$$R = G^2/(1+G^2)$$

3- The third method:

- can be calculated using statistical programs (Al-Sudani, 2016: 68-69).
- previous studies
- Bani Atta study (2018):

The effect of outliers on the differential performance of the mathematics test items in the international study TIMES according to the gender variable

The study aimed to verify the effect of outliers on the differential performance of the mathematics test items in the international study " TEMS " for the eighth grade according to the gender variable . From those who took the test for the year 2011, the final image of the test consisted of (26) items after deleting items that do not match the three-parameter model. Tukey's method (box graph) and the performance indicator method for the item and the test were used to detect outliers, and the differential performance of the test items. The results of the study revealed: The presence of (26) extreme values, including (17) values for females and (9) values for males, and the results also showed the presence of seven paragraphs that showed differential performance with the quality of regular and irregular in the presence of extreme values out of (26) items with a percentage

of (27 %) according to the gender variable After deleting the outliers from the data set and re-analysis, the results showed that there are four paragraphs that showed a differential performance in the quality of regular and irregular by a percentage (15 %), where this percentage decreased by (12 %) when outliers were excluded from the analysis .

Keywords: the differential performance of the item, the three-parameter logistic model, the mathematics test in the international study, the item response theory (Bani Atta, 2018: 157).

A study by Christos et al. (Christos , et. al. , 2021) entitled:

"A comparative study of methods to handle outliers in multivariate data analysis"

A comparative study of ways to deal with outliers when analyzing multivariate data.

This study aimed to develop ways to deal with outliers when analyzing multivariate data communities, where several methods were used, including: methods based on the depth and rooting of outliers within the data community, and others based on the interface dimension between different outliers, and other methods based on Density of values within a certain range (ranges) within the data community, another based on the MAHALANOBIS dimension factor , and another based on the distribution pattern of outliers. The following R Packages were used : depthTools, Lopez-Pintado and Torrente, chemometrics, Filzmoser and Varmuza., DDoutlier, Madsen, mvoutlier, Filzmoser and Gschwandtner. and OutlierDetection, Tiwary and Kashikar) and (OutliersO3, Unwin) In addition to the use of some functions , the results of the study showed that the application of all these methods (individually or collectively) can achieve their purpose, which is to reach the most appropriate ways to deal with outliers when Analysis of multivariate data communities, and the researcher recommended completing these studies to reach the best possible formula to find the best possible solutions to address the problem of statistical outliers (Christos, et. al., 2021)

Research Methodology and Procedures

First, the research methodology: The Methodology of the Research

The method used in this study is the descriptive comparative method It is a method of scientific research This method is based on the study of reality or phenomenon as it is , and cares as an accurate description It is expressed qualitatively , in terms of describing the phenomenon and clarifying its characteristics , or a quantitative expression in terms of giving a numerical description that shows the amount or magnitude of the phenomenon , and the degree of its relationship with other phenomena The descriptive approach does not aim to describe phenomena or reality as it is Rather, reaching conclusions and

generalizations that contribute to the development and understanding of reality The descriptive method is a form of analysis that explains and depicts the phenomenon or problem in a scientific and organized way. , and categorize , and analyzes it and subjects it to _____ study meticulously. Obeidat and others , 2000: 247) .

Second: search procedures procedures of the Research

1- Society of the Research

It is the statistical population or the study population, which is any known grouping of things , people , or accidents , and it is the comprehensive group from which samples are being selected (Odeh and Hebron, 1988: 171).

current research community consists of middle school students (for the fourth and fifth grades of middle school). For morning studies , both males and females, scientific and literary, who are regular in public and private schools affiliated to Najaf Governorate, and for the academic year (20 21 – 20 22), and the total number of students reached (52410), distributed according to gender, at (25,300) male students, representing a percentage of (48 %) of the total community , while the number of females (27110) and those who represent (52 %) of the whole community.

Distributors according to the General Directorate of Education in Najaf Governorate to which they belong, 47496 Male and female students in public schools by a percentage of (91 %) While the number of students in private schools reached (4,914 male and female students and their percentage (9 %).

They are distributed according to the study sub-variable, amounting to (42684) of the students studying in the scientific branch of the study, representing (81%) of the community . And the mother of students in the literary study branch (9726) male and female students representing (19 %) of the community . and their gender and academic branch , as well as obtaining the website list containing the names and addresses of secondary education schools for the academic year (2021-2022) through the researcher's personal visit to the General Directorate of Education in Najaf Governorate and the Department of Educational Planning (Statistics Division of the General Directorate of Education in Najaf Governorate) .

2 - Ain ta search: Samples of the Research

and that Completing the current research requires many procedures Therefore , the researcher will explain the method of his selection of samples and their sizes , each according to the procedure followed at the time , and it should be noted that the researcher chose three samples from the research community , and as follows:

- a sample (Experience clarity of instructions and understanding paragraphs) Its purpose is to know the clarity of the test instructions and to

understand the paragraphs in the test about for the sample , and this sample was formed Out of (100) male and female students .

- a sample (The final statistical analysis It is a random sample from the actual research community and its size was (1000) male and female students from the research community, and its purpose is to organize the data of this sample in two files. And a student, after deleting the extreme values of (97) extreme values, with (34 high extreme values) and (63 minimum extreme values), and that the purpose of the current research is not to measure the students' abstract inference ability, but rather to obtain students' responses that match the expectations of the model. Logarithmic in order to know the effect of extreme values on the standard characteristics according to the traditional measurement theories and the paragraph response theory, and these samples and procedures are presented in some detail within the steps taken to achieve the objectives of the current research.

3- Instrument of the Research:

happened The researcher on the abstract reasoning test _) and it is a test issued About the educational institution (Pmteducation) year (2019) in Britain , which consists of 55 paragraphs with 3 alternatives and some of them 4 with 4 alternatives .

Presentation, interpretation and discussion of the results

Browse search results:

The aim of the quest: to measure the effect of extreme values on the standard characteristics (sincerity, difficulty, discrimination, stability) according to the item response theory.

With regard to the goal of the research, the researcher sought to measure the effect of extreme values on the standard characteristics (sincerity, difficulty, discrimination, stability) that he reached by analyzing the data of the abstract inference test according to the paragraph response theory, specifically (the two-parameter model).

Statistical analysis of the abstract inference test according to the paragraph response theory

1- Verify that the data fits the paragraph response models

The researcher used the BILOG-MG program to verify the suitability of the test items to the two-parameter model, through the scores of a sample that consisted of (1000) participants in the case of (extreme values) and (903) participants in the case of (no extreme values), as follows:

The researcher used the "Ka2" test to find out the significance of the

differences between the number of appropriate and inappropriate items in the cases of the presence and absence of extreme values, and table (1) shows this:

Table No. (1)

The value of " Ka2" and its statistical significance for the differences between the number of appropriate and inappropriate paragraphs
The two-parameter model has both the presence and absence of extreme values

the case	repetitions	appropriate paragraphs	Inappropriate paragraphs	The value of " K2 "	Indication level
Having extreme values	Observation	52	0	2,387	0.05
	expected	52	0		
No outliers	Observation	52	0		
	expected	52	0		

It is clear from the results of the above table that there are no statistically significant differences between the number of appropriate and inappropriate paragraphs in the two cases of the presence and absence of extreme values, as the calculated "Ka2" values amounted to (2.387), which is a non-statistically significant value.

2- Evaluate the coefficients of difficulty

The researcher used the program (BILOG-MG). in extracting the values of the coefficients of difficulty, through the responses of a sample that consisted of (1000) participants in the case of (extreme values), and (903) participants in the case of (no extreme values), and the following table illustrates this:

The researcher used the statistical test for the percentage difference of the number of omitted paragraphs in the two cases of the presence and absence of outliers, according to the values of the paragraphs' difficulty coefficients, and table (2) illustrates this:

Table No. (2)

The results of the Z test for the percentage difference of the number of omitted paragraphs according to the paragraph difficulty coefficients for the abstract inference test according to the paragraph response theory

Statistical procedure	variable	number of paragraphs	The number of paragraphs deleted	percentage	z value		Indication level	judgment
					calculated	tabular		
Paragraph difficulty coefficients	The existence of extreme values	52	0	%0	1.405	1.980	0.05	nonfunction
	No outliers	52	3	%6				

It is evident from the above table that there are no statistically significant differences between the percentages of the number of omitted paragraphs in the two cases of the presence and absence of outliers, according to the values of the paragraphs difficulty coefficients, where the calculated value reached (1.405), which is not statistically significant.

3- The values of the discrimination coefficients:

The researcher used the program (BILOG-MG). In extracting the values of discrimination coefficients, through the responses of a sample that consisted of (1000) participants in the case of (extreme values), and (903) participants in the case of (no extreme values), and the following table illustrates this:

T-test was used For two independent samples to find out the significance of the differences between the mean values of the discrimination coefficients in the two cases of the presence and absence of the mentioned values, the results were as shown in Table (3).

Table (3)

The calculated T-value to know the significance of the differences between the average values of the discrimination coefficients for the items of the abstract inference test according to the item response theory

Statistical procedure	variable	number of paragraphs	Arithmetic mean	standard deviation	T value		Indication level	judgment
					calculated	tabular		
Paragraph discrimination parameter values	The existence of extreme values	52	0.589	0.145	2.942	2.373	0.01	A function in favor of the second formula
	No outliers	52	0.674	0.143				

It is clear from the results of the above table that there are statistically significant differences between the average values of the discrimination coefficients in the two cases of the presence and absence of extreme values, and in favor of the case of the absence of extreme values, where the calculated T value reached (2.942), which is a statistically significant value at the level of significance (0.05).

Test reliability: reliability tests

Contemporary measurement theory asserts that tests that include a good quality of items can be more stable regardless of the number of items they have (mark, 2005: 56).

4- The values of the stability coefficients:

The researcher used the program (BILOG-MG). In extracting the values of indicators of total stability, through the responses of a sample that consisted of (1000) participants in the case of (extreme values), and (903) participants in the

case of (no extreme values), and the following table illustrates this:

The z- test was used to denote the difference between the two reliability coefficients (followed by correlation coefficients), and the results were as shown in Table (4).

Table No. (4)

The offset value indicates the difference between the two correlation coefficients, which indicates the calculated stability

In both cases (the presence of extreme values, and the absence of extreme values) according to the paragraph response theory

Statistical procedure	variable	stability coefficient	Standard Fisher value	face value		level indication	judgment
				calculated	tabular		
Calculation of the values of stability coefficients	The existence of extreme values	0.803	1,099	2.267	1,980	0.05	function for formula the second
	No outliers	0.917	1.557				

It is clear from the results of Table (4) that there are statistically significant differences between the values of the reliability coefficients for the abstract inference test and for the two cases (the presence and absence of extreme values) and at the level of significance (0.05), because the calculated z-value of (2.267) indicates the difference between the two coefficients is greater than the value The table value of (1,980), and that the trend of this difference is in favor of the second formula (the absence of extreme values).

Conclusions

1- There are no statistically significant differences between the number of appropriate and inappropriate paragraphs in the two cases of the presence and absence of extreme values, where the calculated "Ka2" values amounted to (2.387), which is a non-statistically significant value.

2- There are no statistically significant differences between the percentages of the number of omitted paragraphs in the two cases of the presence and absence of outliers, according to the paragraphs' difficulty coefficients, where the calculated Z-value reached (1.405), which is not statistically significant.

3- There are statistically significant differences between the average values of the discrimination coefficients in the two cases of the presence and absence of extreme values and in favor of the case of the absence of extreme values, where the calculated T-value amounted to (2.942), which is a statistically significant value at the level of significance (0.05).

4- There are statistically significant differences between the values of the reliability coefficients for the abstract inference test and for the two cases (the presence and absence of extreme values) and at the level of significance (0.05),

because the calculated Z-value of (2.267) indicates the difference between the two coefficients is greater than the tabular Z-value of (1,980). And that the trend of this difference is in favor of the second formula (the absence of extreme values).

Recommendations

1- The necessity of taking into account the presence of extreme values in the responses when constructing and preparing mental abilities tests , based on the paragraph response theory, in order to reduce measurement errors, which increases the accuracy of the measurement.
2- Introducing those in charge of preparing and building mental abilities tests, of the importance and danger of extreme values in the data, by holding awareness sessions for them, which would enhance their awareness and expand their knowledge and convictions.
3- The need to review the data file or matrix after entering it to ensure that there are no errors in monitoring the corresponding values of the responses.
4- Choosing the appropriate time to apply the measurement tools so that the respondents avoid fatigue or exhaustion.
5- It is necessary to ensure that respondents understand the instructions for answering the paragraphs of the measurement tools before starting the response.

Suggestions

1. The effect of extreme values on standard characteristics according to the three-parameter model.
2. Using the unilateral and ternary parameter A model in addition to the dual parameter A model and comparisons of the results according to these models , and its impact on the standard characteristics , and on the parameters of the paragraphs , in the light of the paragraph response theory.
3. Comparing the methods of exploring extreme values and determining the best ones.
4. The effect of processing extreme values on the accuracy of estimating the parameters of paragraphs according to the three-parameter model.
5. The effect of test length and sample size on the outliers according to the item response theory.

Sources

- Abu Hatab, Fouad (2011) : mental abilities The Anglo-Egyptian Library , Cairo . , and Osman , Mr Ahmed , and honest , Hopes (2008) Psychological Calendar , 4th Edition , Cairo , Anglo-Egyptian Library .
- Abu Nahia , Salahaddin (1994) : educational measurement , Cairo The Anglo-Egyptian Library .
- Ibrahim, Muhammad Farid Hussein (2016): The effect of the method of dealing

with extreme values on the effectiveness of a model equation test , PhD thesis, Yarmouk University, College of Education, Jordan.

- Bani Atta, Zayed Saleh Ibrahim (2018): The effect of extreme values on the differential performance of the mathematics test items in the international study TIMs according to the gender variable , Educational Sciences Studies, Volume 45, Number 4, Supplement 2, University of Jordan, Deanship of Scientific Research.
- Bani Yassin , Omar Saleh Mofdi (2004) : The psychometric properties of the reference test in chemistry for first year secondary scientific students estimated according to the classical and modern theories of measurement Unpublished doctoral thesis Amman University .
- Al-Bayati, Mahmoud, Wadgha, Delir (1996): Determining outliers using exploratory methods and comparing them with parametric methods, unpublished master's thesis , University of Baghdad. Iraq .
- Zechariah Ali bin Mohammed Abdullah (2009) Psychometric properties of the test : Otis - for neon Mental ability is estimated according to the classic measurement and the Rasch model among middle school students in Sabya Educational Governorate PhD thesis, , Faculty of Education , Umm Al Qura University , Kingdom Saudi Arabia .
- Zahrani , Bandar bin Hamdan (2008) The effect of the difference in sample size and the breadth of the ability to accurately estimate the true degree estimated using the traditional theory and one-dimensional models in the modern measurement theory , Ph.D , measurement and calendar , Umm Al Qura University , Kingdom Saudi Arabia .
- Al- Zayyat, Fathi Mustafa (1995) : Cognitive bases for mental formation and information processing, Cognitive Psychology Series , 1st Edition, Dar Al-Wafaa, Mansoura.
- Al- Sudani, A comprehensive interview with Khalaf (2016): Standard characteristics of tests of logical problems among middle school students according to the traditional measurement theory and paragraph response theory, unpublished doctoral thesis, College of Education Ibn Rushd, University of Baghdad, Iraq .
- Sharqawi , Anwar Muhammad , and Sheikh , Suleiman Al-Khudari , kazem , Omnia Muhammed , and Abdul Salam , Nadia Muhammad (1996) Contemporary trends in psychological and educational measurement and evaluation , Cairo The Anglo-Egyptian Library .
- Al-Dawy, Mahsob Abdel-Qader (2011): Investigating the effect of extreme degrees and the number of response categories on estimating Cronbach's alpha coefficient , Journal of the College of Education, Volume 27, Number 1, Part One, University Assiut - College Education .

Allam, Salah El-Din Mahmoud (1986) : Contemporary Developments in Psychological and Educational Measurement , Kuwait University, Kuwait .

(1995) Future directions for evaluating student achievement in light of the

requirements of the twenty-first century , Journal of Education , Al Azhar university , the number (49) .

(2001): Diagnostic tests reference criterion , second edition, Dar al-Fikr al-Arabi, Nasr City, Cairo.

- Eyal, Yassin Hamid (2005): Standardization of the Henmon-Nelson Test of Mental Abilities for University Students, PhD thesis, College of Education Ibn Rushd, University of Baghdad, Iraq .

- Carter , philip (2005) IQ and psychometric tests , i 1 Jarir Bookstore , Riyadh , Saudi Arabia .

- Garage , Abdul Qadir (1997) Measurement and evaluation in psychology . new vision)) , 1st floor, Amman Al-Bazuri Scientific House .

Crocker, Linda , and alginate , James (2009) : Introduction to traditional and contemporary measurement theory , Translation : let's Zeinat Youssef , 1st floor , Dar Al-Fikr, publishers and distributors .

- Al-Mukhtar, Suleiman (1980): Extreme values and their impact on statistical data analysis, unpublished master's thesis in Statistics, University of Baghdad, Iraq.

- Najjar, Nabil Juma Saleh 2010) : measurement and calendar An applied perspective with spss software applications , first edition Dar Al-Hamid for publishing and distribution , Jordan .

- Al-Naimi, Aswan (2012): Detection and treatment of outliers in the fortified way and comparing them with other methods, Tikrit Journal of Administrative Sciences and Economics , 8 (25).

- Al-Noor, Nadia (2010): A comparison of some resilient methods for estimating the location parameter of some probability distributions, Kufa Journal of Mathematics and Computers , 1 (1).

foreign sources

, MJ, & Yen, WM (1979): Introducation to Measurement Theory . California: Cole Publishing Company.

- Barnett, V., & Lewis, T. (1977) : Outlier in statistical data. and Sons . The Edition . New York : John Wiley .

- Dan, E . D. & Ijeoma, O. A. (2013a) : Statistical Analysis / Methods of Detecting Outliers in A Multivariate Data in Regression Analysis Model . Journal of International Academic for Multidisciplinary,1(3),302-337 .

- Hambelton, Thomas (1995) : Assessment of abitities : ERIC learning house on counseling and student service greenshoro NC. ERIC NO. 389960, p(1-5).

- Popham, W.J (1978) : Criterion Referenced Measurement. Englewood cliffs ,NJ .prentic hall.

- Rahman, M. & Al Amri, K. (2011) : Effect of Outlier on Coefficien of Determination. International Journal of Education Research Academic Journal, 1(6). P 9.

- Seo, M.S. (2006) : A Review and Comparison of Methods for Detecting Outliers in University Data Sets. Unpublished Master Thesis, University of Pittsburgh.

- Nasrollahzadeh, S., & Koramaz, T. K. (2020). Residential satisfaction and mobility in Göktürk peripheral neighbourhood. *socialspacejournal.eu, 20*(2), 51-84. http://socialspacejournal.eu/Social%20Space%20Journal%2022020(20).pdf#page=51
- Ottuh, P. O. O. (2020). Xenophobia in Africa: origins and manifestations. *socialspacejournal.eu,* *20*(2), 29-50. http://socialspacejournal.eu/Social%20Space%20Journal%2022020(20).pdf#page=29
- Rygielska, M. (2020). Migrująca teranga. O współczesnych przemianach afrykańskiego systemu świadczeń całościowych we Włoszech. *socialspacejournal.eu,* *19*(1), 71-88. http://mail.socialspacejournal.eu/Social%20Space%20Journal%2012020(19).pdf#page=71