Students with Aberrant Responses in TIMSS2019 Mathematics Test in Sultanate of Oman: Comparison of Between Squared Residual Index by Content, Cognitive Process, Item Type, and Difficulty

Nawal Ali Al Yahyai[1], Rashid Saif Al Mehrizi[2], Ihab Mohammed Amara[3]

[1,2,3]Sultan Qaboos University, Sultanate of Oman

Abstract:

This study aimed to detect the students with misfit responses for eight grade students in the Sultanate of Oman in 2019, who amounting 6,745 students, by using the Between Squared Residual Index over sections of TIMSS2019 Mathematics test (content, cognitive process, item format and item difficulty). This study also aimed to compare the results of the Between Squared Residual Index over Mathematics test sections with the results of the Squared Residual Index over item level. The results of the study indicated that the percentage of students with misfit responses ranged between (2.16% & 3.22%) and between (2.13% & 3.88%) according to the (unweighted and weighted) Between Squared Residual indices respectively, while their percentage amounted to only 0.30% & 0.95% according to the (unweighted & weighed) Squared Residual indices respectively. The results showed that most of students' abilities with misfit responses in the Mathematics Test in TIMSS2019 ranged between low and medium. The study recommended conducting studies to determine the reasons for the presence of misfit responses in Mathematics tests in TIMSS among students of the Sultanate. It also recommended directing teachers and those in charge of applying TIMSS tests to train students in the basic mathematical skills they need to reduce behaviors that lead to misfit responses.

Keywords: Misfit Responses, Between Squared Residual Index, Squared Residual Index, TIMSS2019 Mathematics Test

Preamble

Given the importance of education as a basis for growth and competitiveness between countries, different countries are keen to develop their education systems, by comparing their performance with that of different educational systems. This is achieved by international assessments that provide a wealth of data that contributes to improving educational policies. As such, the Ministry of Education in the Sultanate of Oman was keen to participate in international tests, including the Trends of the International Mathematics and Science Studies (TIMSS) tests, which serve as one of the indices for evaluating the education's quality in the subjects of Science and Mathematics. The purpose of the science and Mathematics TIMSS tests is to obtain full information on the concepts and points of view that fourth and eighth grade students gained in these subjects. These tests aim to assess the achievement and provide information to improve teaching and learning of Mathematics and science in participating countries. This study is being conducted under the supervision of the International Association for the Evaluation of Educational Achievement (IEA), which is located in Amsterdam, Netherlands. This organization works to monitor the results of this test, in which more than 60 countries participate, and it is held periodically every 4 years (Zaki, 2011).

Accurate measurement of student characteristics is necessary for the educational community to make the right decisions. This helps policymakers decide on important educational issues. In contrast, inaccurate measurement of test performance can result in negative consequences. Falsely high-test degrees can undermine assessment of student learning progress and curriculum planning efforts (Karabatoss. 2003).

Since the psychometrics movement is found, measurement scientists have been interested in methods that would achieve the highest level of objectivity in measurement tools. The scientists' efforts have resulted in some recent trends in measurement and assessment. One such trend is Item Response Theory (IRT), which introduced many methods to achieve high accuracy scores while evaluating results of measuring tools.

One of the methods created by the Item Response Theory (IRT) is the Person Fit Statistic (PFS) index, which helps identify misfit or aberrant response patterns. The responses of persons are misfit when their response pattern differs significantly from the expected response pattern. Thereafter, their responses have outliers according to the used Person Fit Statistic index, thus identifying their response pattern as a misfit response to the IRT model used (Ferrando & Chico, 2001).

Misfit responses occur as a result of various factors that affect an individual’s performance, as Wright (1977) stated, laziness as a result of boredom, which leads to the individual answering inaccurately, especially in the last questions of the test, and confusion that occurs due to the form of the test, which leads to not answering the first questions accurately, slowness in answering the questions, which leads to missing the solution to the last questions, guessing and cheating. Misfit responses also occur due to the extraordinary creativity of some students in answering the questions. Some of individuals also show misfit responses, such as being of high ability, or receiving education in a language other than their native tongue (Bracey et al., 1992).

Many studies have been interested in investigating the reasons for the emergence of misfit responses of individuals in various tests, including Bani Atta's Study (2019), found that guessing, cheating, laziness, and extraordinary creativity were the primary reasons of misfit response patterns in the Otis-Lennon test, which was standardized for the Jordanian environment, Birenbaum's Study (1986) also found that the anxiety affected individual's performance in an aptitude test, and Chen's Study (2004) found that the time pressure and anxiety affected individual's performance with medium and high abilities on an English language test given to Chinese students. In addition, Brown & Villareal’s Study (2007) found that cheating played a role in the existence of misfit responses among low-achieving students in an adaptive test in Mathematics.

Person Fit Statistic Indices

Person Fit Statistic can be used to identify misfit patterns, such as a student with low ability answering questions of a high level, or vice versa. According to Meijer (2003), it is important to match the individual in order to determine if the participant's response is according to the essential trait being measured or it is affected by other factors (Steinkamp, 2017).

Person Fit Statistics indices come in two types: (1) over item level and (2) for test sections level. There are many indices over item level that have been used by specialists in psychological and educational measurement, it is numbered 36 indices in the Karabatoss's Study (2003). Over item level, some examples of indices are Wright's index (1977) based on residuals and Drasgow et al. L_z's index (1985), which is based on the maximum likelihood, and Almehrizi (2003) squared residual index, which is also based on the residuals.

There are two indices at the level of the test sections: the Wright's index (1977) and the Between Squared Residual Index, which Almehrizi (2019) developed and has good statistical properties, as stated in the results of his study in which he studied the statistical properties of both Between Squared Residual Index (unweighted & weighed) and the Wright's index (1977) (unweighted & weighed).

Person Fit Statistic Indices over item level can be differentiated from those at the test section level in that the first is calculated by the value from scores of each test item that the student obtains. The index's value in indices at the test sections level is calculated from the student's scores on each section of the test. indices at the test section level examine the existence of misfit responses in the sub-scores of the test sections by comparing them with the expected sub-scores of the test sections according to the Item Response Theory used.

Between Squared Residual Index

Between Squared Residual index is one of the modern indices at the test sections level and it is launched from Squared Residual Index over item level.

The Squared Residual Index over item level is one of the indices that use the residual approach between the individual's observed response and the probability of the correct answer to the item, and it is in two versions: unweighted ($\mathrm{USR_a}$) and weighted ($\mathrm{WSR_a}$).

The unweighted is calculated as follows:

$$USR_a = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{(y_{ai}-p_{ai})^2-p_{ai}q_{ai}}{\sqrt{p_{ai}q_{ai}(p_{ai}-q_{ai})^2}} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{(p_{ai}-y_{ai})(p_{ai}-q_{ai})}{\sqrt{p_{ai}q_{ai}(p_{ai}-q_{ai})^2}}$$

Whereas, y_{ai} represents individual's response a to the test item i with the number n of items, p_{ai} represents the probability of individual's answer a to the item i, $\mathrm{q_{ai}}$ represents $(1-p_{ai})$.

As for the weighted version, it is calculated as follows:

$$WSR_a = \frac{\sum_{i=1}^{n}(y_{ai}-p_{ai})^2-\sum_{i=1}^{n}p_{ai}q_{ai}}{\sqrt{\sum_{i=1}^{n}p_{ai}q_{ai}(p_{ai}-q_{ai})^2}} = \frac{\sum_{i=1}^{n}(p_{ai}-y_{ai})(p_{ai}-q_{ai})}{\sqrt{\sum_{i=1}^{n}p_{ai}q_{ai}(p_{ai}-q_{ai})^2}}$$

Almehrizi (2010) states that if the data fit the mathematical model of the Item Response Theory, both WSR_a and USR_a follow the normal distribution with mean of zero and one standard deviation. The values of WSR_a and USR_a indices will be large and positive to show misfit values in response patterns, which are values above 1.645 and corresponding to a significance level of 0.05 (Odeh, 2019).

Almehrizi's Study (2010) is one of the studies that dealt with the Squared Residual Index over item level, in which it compared the statistical properties of the Squared Residual Indices (unweighted & weighed) with the statistical properties of Wright's indices (1977) (unweighted & weighed) by using data generated consisting of twelve sets of data that resulting from two tests of different length, three different levels of difficulty, and two different levels of discrimination. The results showed that the Squared Residual Indices had better statistical properties under different test conditions. Al Shaqsi's Study (2019) also aimed to examine the effectiveness of three person fit indices matching (Weighted Wright, Drasgow and Weighted Squared Residuals) in item response models when the strength of the local dependence between items varies. The results directly indicated that the percentage of misfit individuals increases with the increase in the local dependence between items of the three indices: (Weighted Wright, Drasgow and Weighted Squared Residuals). and the descriptive statistics, skewness and kurtosis for the indices indicated that the values of the weighted Squared Residual index follow the normal distribution in the case of local dependence 0.0, and this result is consistent with what was stated in Almehrizi (2010) study, in which he indicated that if the data matched the model used, the results of the index followed the normal distribution with mean of zero and one standard deviation, while in cases of local dependence 0.3, 0.6 and 0.9, the results moved away from the normal distribution and maintained the form of distribution.

Odeh's Study (2019) sought to compare the effectiveness of five Person Fit Statistic Indices over the item level: the weighted & unweighted Squared Residual Index (WSR & USR), the unweighted USR, Drasgow (Lz) index, and the weighted & unweighted Wright indices (WT & UT) using numerical ability test data in the Gulf Mental Ability Scale. The sample size was 4,206 students from the Gulf countries. The value of the Squared Residual Index (unweighted & weighed) is one of this study's results, which has an arithmetic mean that is not equal to zero and is distributed normally. This runs counter to what Almehrizi (2010) found in its study, which is showed that both the WSR and US indices follow a normal distribution with a mean of 0 and a standard deviation of 1 if the data match the used model. While Almehrizi used produced data in its analysis, the researcher used actual data. Additionally, the study found that, in terms of determining the number of misfits responds, the weighted Wright and Drasgow indices ranked first and second, the weighted Squared Residual Index ranked third, and the unweighted Squared Residual Index ranked fourth, whereas the unweighted Wright index ranked last. The Drasgow index and the Squared Residual index were the two most reliable indicators out of the five, according to a study result on index validation. According to the study, cheating is the most widespread misfit patterns and is more common in males than in females. All of these studies indicated that the Squared Residual Index over the item level has good statistical properties.

Between Squared Residual index depends on the amount of congruence in the results of the individual on each section of the test and examining the possibility of interpreting this match according to the IRT model used. This index also has two versions: weighted and unweighted (WBSR & UBSR). The two versions of the index use to estimate the amount of match $BSR_{\alpha j}$ in scores of student α in each section j by calculating the square of the difference of remainders:

$$BSR_{\alpha j} = \left(X_{\alpha j} - \sum_{i=1}^{n_j} p_{\alpha i}\right)^2$$

Whereas, $X_{\alpha j}$ represents the sum of scores of α student in j section and n_j represents the number of items in j section.

$BSR_{\alpha j}$ takes any continuous value between zero and n_j^2 for any student depending on their ability levels. To rule the fit estimate and use it to classify students' responses as appropriate to the model or inappropriate, calibration is required by calculating the mean and standard deviation of this estimate from the $BSR_{\alpha j}$. Under the assumption of local independence of item responses using IRT, the mean of $BSR_{\alpha j}$ for each section is the value of the known variance for the overall score of each section by test (given the ability):

$$E\left(BSR_{\alpha j}\right) = \sum_{X_{j=0}}^{n_j} \left(X_j - \sum_{j=1}^{n_j} p_{aj}\right)^2 f\left(X_j|\Theta_a\right) = \sum_{i=1}^{n_j} p_{ai} q_{ai}$$

Similarly, the variance of the difference of the Squared Residual, $BSR_{\alpha j}$ for each section can be obtained by the typical variance equation:

$$Var\left(BSR_{aj}\right) = \sum_{X_j=0}^{n_j}\left[\left(X_j - \sum_{i=1}^{n_j} p_{ai}\right)^2 - \sum_{i=1}^{n_j} p_{ai}q_{ai}\right]^2 f\left(X_j|\Theta_a\right)$$
$$= \sum_{X_j=0}^{n_j}\left(X_j - \sum_{i=1}^{n_j} p_{ai}\right)^4 f\left(X_j|\Theta_a\right) - \left[\sum_{i=1}^{n_j} p_{ai}q_{ai}\right]^2$$

Whereas, X_j are all possible scores on each set of j and $f\left(X_j|\Theta_a\right)$ is the probability density function of X_j scores when the examiner's ability is a. This function can be obtained through the recursion formula Lord and Wingersky (1984) used by Almehrizi (2019).

To get $f\left(X_j|\Theta_a\right)$ for each section, defining X_j as a random variable for the raw scores on the first items in the test section requires. Subsequently $f\left(X_j = x_j|\Theta_a\right)$ represents the probability density function for each $X_j$when it is equal to $X_j$in test section j consisting of i items. The periodic formula is applied to each test section and we start by assuming that it consists of one item, and i = 1 is entered or inserted into the formula:

$$f\left(X_{j1} = 0|\Theta_a\right) = 1 - p_{a1}$$
$$f\left(X_{j1} = 1|\Theta_a\right) = p_{a1}$$

Then we enter the rest of the items in that section so that the formula for $i > 1$ is as follows:

$$f\left(X_{ji} = x_i|\Theta_a\right) = f\left(X_{j(i-1)} = x_j|\Theta_a\right)(1 - p_{ai})$$
$$+f\left(X_{ji} = x_j - 1|\Theta_a\right)p_{ai} \quad \text{for} \;\; x_{ji} = 0,1, \dots . i$$

To use the periodic formula, the items enter the periodic formula in any order, starting from i = 1, and apply the formula repeatedly by incrementing i in each iteration. The process stops after i =n$_j$, which gives the required formula $f\left(X_j|\Theta_a\right)$ wherein;

$$f\left(X_j|\Theta_a\right) = f\left(X_{jn_j} = x_{n_j}|\Theta_a\right)$$

The standard unified version of the squared difference of the residuals across all sections (the unweighted index) can be obtained in two methods. The first method requires standardizing the BSR_{aj}in each section, then summing all the sections and finally dividing by the square root of the number of sections (J) as follows:

$$UBSR_a = \frac{1}{\sqrt{J}}\sum_{j=1}^{J}\frac{BSR_{\text{a}j} - E(BSR_{\text{a}j})}{\sqrt{Var(BSR_{\text{a}j})}}$$

The other method to form a standardized version of the squared difference of the residuals across all sections is (the weighted index) by summing the extent to which a person's scores match across all groups of items. Then calculating the standardized version using the positional independence of the test sections.

$$WBSR_a = \frac{BSR_{\text{a}} - E(BSR_{\text{a}})}{\sqrt{Var(BSR_{\text{a}})}}$$

Where:

$$BSR_a = \sum_{j=1}^{J} BSR_{aj}$$

$$E(BSR_a) = \sum_{j=1}^{J} E\left(BSR_{aj}\right) = \sum_{i=1}^{n} p_{ai}q_{ai}$$

$$Var(BSR_a) = \sum_{j=1}^{J} Var\left(BSR_{aj}\right)$$

Assuming the IRT model matches the test data, both (the weighted and the unweighted) indices theoretically follow a moderately normal distribution. Extreme scores of the Person fit index at the section level match the right end of the moderate normal distribution and are deemed misfit patterns of response, which are values that exceed the value 1.645 at a significance level of 0.05 (Almehrizi, 2019).

Almehrizi (2019) compered four person fit indices using the residuals on the test sections: the weighted and unweighted Between Squared Residual indices, and the weighted and unweighted indices (Wright 1977). The study used Dichotomous Item Models. It also verified the statistical properties of these indices by applying them to hypothetical generated data (by controlling the number of test sections and the number of items in each section), and other data generated from real features of the verbal ability test in the Gulf Scale of Mental Abilities, in which abilities are distributed at specific levels (-3, -2, -1, 0,1,2,3) so that it follows a normal distribution. The study concluded that the two indices of the Between Squared Residual had distinct statistical properties, in contrast to what appeared when (Wright 1977) two indices were applied, where the arithmetic mean values were very close to zero for all abilities within all of the total data, even those with the smallest number of sections and items within. Each section, as well as the two Wright indices, had arithmetic mean values that differed from zero at all ability levels in the entire data set. Furthermore, the standard deviation values in the two indices of the Between Squared Residual were close to the theoretical value (one) for all abilities, while the standard deviation values in the two Wright indices deviated from one and were affected more by the number of sections than by the number of items in each. Type I error rates were inflated in the Wright indices at all levels of ability in all data, and the error rates were affected by the number of test sections, with more sections improving the Wright indices' capacity to limit Type I error rates. In contrast, the Type I error rates for the two Between Squared Residual indices were about 0.05 across all datasets. In terms of detecting non-matching replies, The Between Squared Residual indices were able to detect misfit responses even at the ability levels at which the Wright index had inflated error rates of the first type. The increase in the number of items per section also improved the strength of this index.

Study Problem:

Developed countries rely heavily and primarily on multiple educational studies and research to develop their educational systems, and the results of TIMSS tests are regarded as one of the sources that can be relied on in the development of science and mathematics education systems and practices in participating countries. Those who follow the research conducted on the results of TIMSS in the participating countries note their focus on Multiple factors and variables, as well as the endeavor to determine their impact on students' achievement in these two subjects. The Gulf countries, especially Sultanate, should prioritize these results in terms of study and research, rather than simply reviewing the recently disclosed results. Alternatively, providing a descriptive report on the country's position, or knowing the percentages of students who answered particular test items, since this type of report does not help to benefit from the results of the study in a deep way that can be reflected in the development policies of the educational system and its various practices, especially at the school and classroom levels. Thus, many studies and analytical research should be conducted to look deeply at the results of TIMSS in order to reach scientific evidence that leads to the development of educational policies and practices (AlShamrani et al., 2016).

Since 2007, Sultanate has participated in TIMSS testing cycles, knowing that these cycles are held once every four years. In the mathematics and science tests for the eighth grade, Sultanate obtained low results during the previous sessions, as it obtained 372, 366, 403, and 411 points for the years 2007, 2011, 2015, and 2019, respectively, in the mathematics test. In the science test, Sultanate obtained 423, 420, 455, and 457 points, respectively, knowing that the general average is 500 points. In mathematics, Sultanate ranked 41, and in science, it ranked 36 out of 48 participating countries in 2007. In addition, Sultanate ranked 41 in mathematics and science, with a rank of 36 out of 42 in the 2011 session, a rank of 32 in mathematics and 29 in science, out of 39 in the 2015 session, and a rank of 35 in mathematics and 30 in science out of 39 in 2019 session. (TIMSS & PIRLS, n.d).

The decline in these results is due to the fact that a percentage of students' responses did not match the TIMSS test questions, emphasizing the importance of carefully studying these results and recognizing that they are the result of expected responses according to the students' levels, or of misfit responses resulting from various factors such as anxiety, guesswork, and others, by employing Person Fit Statistic indices that have been proven effective. Determining the causes of misfit replies in international tests such as the TIMSS test allows specialists to reassess educational practices, resulting in better international outcomes. By reviewing previous studies that dealt with the results of TIMSS, it is found that they focused on the reasons for the decline in student results in that they are due to various factors such as teachers' practices, the attitudes of high-achieving students, and others. However, few of them paid attention to the issue of ensuring that these results represent the students' real responses. Among the studies that examined the results of TIMSS were Odaibat's study (2019), which aimed to compare the practices of eighth-grade science teachers in Jordan and Singapore in light of the results of TIMSS2015, to reveal the contribution of these practices to the discrepancy in student results between the two countries. A questionnaire was used for science teachers in Jordan and Singapore participating in TIMSS 2015, and among the results of the study there were statistically significant differences in teaching practices and practices related to the teacher's sense of confidence while teaching in favor of Singapore. Al Alawi's study (2017) aimed to reveal the extent to which science books for grades 5–8 in Sultanate of Oman included topics in the TIMSS 2015 tests. A content analysis card was prepared in light of these topics after they were translated and presented to specialists. The results showed that the content of science books included topics from the TIMSS 2015 tests in varying proportions. In addition, the study recommended the necessity of informing the authors of science books about TIMSS topics, including them in accordance with the proportions stipulated, and taking advantage of the participating countries that achieved advanced positions and how their science books included these topics.

The study by Shehadeh and Al Qaramiti (2016) also aimed to identify the reasons for the low level of achievement of students in Kingdom of Saudi Arabia in the TIMSS results from the point of view of teachers and supervisors by preparing a questionnaire consisting of 46 items distributed over four fields: the curriculum, the teacher, the learner, and the educational environment. The study reached several results, including students' lack of seriousness in answering test questions because they are not included in their academic results. Furthermore, Abu Aish's study (2015) sought to identify the study habits and personal characteristics of students who obtained high results in TIMSS tests compared to others who obtained low results. The study was based on the results of students in Kingdom of Saudi Arabia in TIMSS 2003 and on the questionnaires that they filled out as part of the TIMSS 2003 tools, which dealt with the students' family background, their attitudes and ambitions, the classroom and extracurricular practices of mathematics and science teachers, and their use of computers inside and outside the school. The study reached several results, including that the attitudes of students who obtained high results toward science and mathematics were more positive than those who obtained low results.

One study that looked into identifying non-matching student responses on TIMSS tests was that conducted by Al Jarrah (2020). Which aimed to identify response patterns in the TIMSS 2015 test in the science and mathematics subjects for eighth grade students in Kingdom of Saudi Arabia and eighth grade students in the state of Singapore using person fit indices. Developed by (Huang 2011), based on the Within Ability Index (W) and (B) Beyond Ability Index, which were developed by (D' Costa, 1993a 1993b), Huang's indices are (w_i^1, B_i^1, W_i^1, B_i^0) or Capability Index, Guessing Index, Carelessness Index, and Misconception Index, and these indices are calculated on the Guttman matrix based on the order of the examinees' abilities from highest to lowest and the order of items from easiest to difficult. The results of the analysis of Singapore data for science and mathematics showed that 92% and 94%, respectively, were classified as normal response patterns and 8% and 6% were classified as abnormal response patterns, while the results of the analysis for Saudi Arabia for science and mathematics were 85% and 86%, respectively, normal responses and 15% and 14%, abnormal responses. The reasons for these abnormal responses from students in both countries were due to guesswork and carelessness.

The importance of Person Fit Statistic indices at the test section level is that they help detect misfit responses that cannot be detected using Person Fit Statistic indices at the single item level. The study by (Felt et al., 2017) provided an example of this, demonstrating that the use of conformity indices over the item level led to not detecting misfit responses among respondents to one of the questionnaires applied for the purposes of a health study. Although there were eighteen misfit responses indicated by the conformity indices at the test section level, it would have been impossible to fully comprehend the issues these respondents faced if these responses hadn't been disclosed.

The Between Squared Residual Index is considered one of the new indices that appeared recently at the level of the test sections. It's regarded as an extension of the approach taken by Almehrizi (2004, 2010) to the individual suitability index at the individual level, which is based on the residuals. The new indicator depends on the amount of correspondence in the individual's results on each section of the test and examines the possibility of interpreting this correspondence according to the Item Response Theory (IRT) model used or whether it indicates the presence of aberrant results for the test sections (Almehrizi, 2019).

The test items can be categorized based on the nature of the trait that it measures in terms of content areas and axes that make up the psychological trait or the cognitive trait. In accordance with Bloom's definition of educational outcomes in terms of academic achievement, the test items can also be divided into sections according to the cognitive processes targeted by the test. Indices for examining matching responses across test sections help judge the extent to which student responses match different content areas or cognitive processes. It is expected that a student's actual performance will differ from his expected performance in these sections, depending on the student's level of ability. Conversely, the type of items used in the test and their levels of difficulty can be used to divide the test and be used to compare the student's actual performance with his expected performance in these sections and to detect students with misfit responses.

Study Questions:

In evaluating the responses of eighth-grade students in Sultanate of Oman, the current study aims to use the diversity of the mathematics test items in TIMSS2019 in terms of the content areas they measure, the cognitive processes, the type of items used, and their levels of difficulty. It also aims to determine the extent of the prevalence of aberrant or misfit responses using two indices of the square of the residuals. The following queries can be used to summarize the study problem:

1. What are the descriptive statistics of the values of the Between Squared Residual indices (weighted

and unweighted) in the eighth-grade mathematics test at TIMSS2019 in the Sultanate of Oman using the four test sections: content, cognitive process, item format and item difficulty?

2. What is the percentage of individuals with misfit responses in the eighth-grade mathematics test at TIMSS2019 in the Sultanate of Oman according to the two Between Squared Residual indices using the four sections of the test: content, cognitive process, item format, and item difficulty?
3. Does the percentage of individuals with misfit responses in the eighth-grade mathematics test at TIMSS2019 in the Sultanate of Oman differ when using the Squared Residual at the level of the item and when using the Between squared Residual at the level of the four test sections: content, cognitive process, item format, and item difficulty?

Importance of Study:

The importance of this study comes from the fact that it is one of the few studies that deals with detecting misfit responses in students' results in TIMSS tests, which may help specialists in educational evaluation make better decisions to improve the level of students' performance. This study helps in detecting misfit responses of students using individual conformity indices at the individual level and at the test section level. Detecting the presence of this type of response and knowing their percentage helps in better identifying the problem of Sultanate's low results in the TIMSS tests. If the percentages are high, this requires for research into the reasons behind these misfit responses in order to deal with them in a way that reduces them in the future and thus improves the results in the upcoming sessions of the TIMSS tests. However, if the percentages of misfit responses are low, meaning that the highest percentage of students' responses are normal, then this indicates the presence of a real weakness among the students of Sultanate. Subsequently it is necessary to study the possible causes of this weakness and develop appropriate plans to treat this weakness.

The study's theoretical significance is further demonstrated by the way it concentrates attention on Person Fit Statistic indices at the test section level, which had previously been neglected because of the Person Fit Statistic indices' poor distribution characteristics over all test sections. This is accomplished by shedding light on one of the modern indices based on residuals, which is the Between Squared Residual Index. It is necessary not to neglect indices at the test section level, as they reveal misfit responses that indices at the individual level cannot detect. This study is considered the first in which the Between Squared Residual Index is used on field data.

Terminology of Study:

Person Fit Statistic Index: It was stated in (Lopez & Montesinos 2005) and (Meijer & Sijtsma 2001), which was mentioned in Hamadnah (2015), that it is a statistical index that determines the range or distance between the actual data represented by individuals' responses and the values expected through the model used by comparing the values of these. Indicators with a critical value identify matching response patterns and misfit response patterns.

The Between Squared Residual Index: An index to detect misfit responses of an individual at the level of test sections. It is considered an extension of the Squared Residual index over the item level, which was developed by (Almehrizi 2004, 2010) and which is based on the residuals approach. The square of the residuals index depends on the sum of the individual's scores on each section of the test and not on the scores of each item (Almehrizi 2019).

Limits of Study:

- The sample data are drawn from the results of eighth grade students only on the TIMSS2019 mathematics test.
- The study is limited to using the three-parameter model based on the concepts of Item Response Theory.
- The accuracy of the results depends on the accuracy of the software used in estimating the parameters of individuals and items and that used in estimating misfit detection indices

Study Approach and procedures

Study Approach:

The current study relies on the descriptive approach by describing and analyzing the data collected from the study sample and treating it in light of statistical methods. The data of eighth grade students in Sultanate of Oman in the mathematics test in TIMSS 2019 was analyzed using Item Response Theory, then the number and percentages of misfit responses for these students were calculated using the weighted and unweighted Almehrizi (2019) indices of the Between Squared Residual across the test sections and compared to the number and percentages of misfit responses by using the Squared Residual indices (Almehrizi, 2010), weighted and unweighted, over the item level.

Study Population and Sample:

The study used archival data for the responses of eighth-grade students participating in the TIMSS 2019 mathematics tests applied to them in the academic year 2018/2019, who numbered 6745 where is 3341 related to male and 3404 related to female students distributed across various governorates and regions of Sultanate of Oman.

Study Tool:

The study used the TIMSS 2019 mathematics test, which consists of 14 equivalent booklets, a booklet for each student to complete. The International Association for the Evaluation of Educational Achievement (IEA) prepares TIMSS test booklets using the Matrix Sampling technique, in which questions are divided into blocks or groups of questions, so that each booklet contains two sets of these blocks or groups of questions.

The number of questions in each group of question groups ranges between 12 and 18 questions and each question appears in two booklets. For example, the first booklet consists of question groups that begin with the symbol MP01 & MP02. The second booklet consists of question groups that begin with the symbol MP02 & MP03 and so on, which provides a mechanism to link students' responses from different booklets together when taking data from all booklets together (Mullis et al., 2017).

The questions in the mathematics test are divided into two dimensions: the content dimension, which is divided into four areas: numbers (30%), algebra (30%), geometry (20%), data and probability (20%), and cognitive processes which consists of three levels: knowledge (35%), application (40%), and reasoning (25%) (Mullis et al., 2017).

In this study, two-response questions were relied upon because the Between Squared Residual index is appropriate for this type of response and cannot be applied to multi-response questions. The two-response questions were divided into two parts according to the item formats: multiple choice and short question. Scores on multiple choice questions and two-response questions were converted to zero and one. The test items were also divided into three sections according to their level of difficulty using the difficulty parameter in the three-parameters model: (0-0.75), (greater than 0.75) and (less than 0), and this division was adopted since most of the questions in the TIMSS2019 math test range from very difficult to medium difficulty. The two-response questions were distributed to the different sections of the test (content, cognitive process, item format, and item difficulty)in order to enter them into the R program to perform the required analyzes, and their distribution was as in Table (1), which shows the test sections and the number of items in them according to the four divisions in each booklets of the mathematics test:

Table (1)

Number of questions distributed throughout the test section's fields in the mathematics booklets, as well as the total number of students per booklet.

		Booklets														
Sections	Sections	1	2	3	4	5	6	7	8	9	10	11	12	13	14	Total
No. of Students		486	474	488	480	476	476	483	479	478	482	489	486	483	485	
Total Questions		30	28	27	28	26	26	30	27	26	27	27	26	29	33	390
Content	Numbers	8	7	8	10	10	9	9	8	8	7	8	9	9	10	120
	Algebra	10	9	9	9	6	5	9	10	9	9	9	8	8	10	120
	Geometry	6	6	4	3	4	5	5	5	6	6	5	5	7	7	74
	Probabilities	6	6	6	6	6	7	7	4	3	5	5	4	5	6	76
Cognitive processes	Knowledge	7	9	9	11	10	9	10	8	7	6	9	10	10	9	124
	Application	15	12	12	14	14	13	15	13	11	12	12	13	14	16	186
	Reasoning	8	7	6	3	2	4	5	6	8	9	6	3	5	8	80
Type of questions	MCQ	15	12	12	12	15	10	14	11	10	11	13	12	11	16	174
	Short answer question	15	16	15	16	11	16	16	16	16	16	14	14	18	17	216
Difficulty	0>	6	6	6	6	5	6	5	2	5	6	2	1	2	5	63
	0-0.75	16	16	11	13	13	12	16	15	11	12	17	17	18	18	202
	0.75<	8	8	10	9	8	8	9	10	10	9	8	8	9	10	125

Table (1) shows that in the content sections, the number of questions in the numbers and algebra fields constitutes the largest number of questions, followed by the probability field and finally geometry. In the cognitive process's sections, the application field constitutes the largest number of questions, followed by the knowledge field, and finally the reasoning field. In the type of question sections, the number of short questions is greater than the number of multiple-choice questions MCQs. Table (1) also indicates that the number of questions is greater in (0-0.75), then (>0.75), and finally (less than 0).

Psychometric properties of the TIMSS2019 mathematics test

Validity:

The validity indices of the TIMSS2019 mathematics test results were verified by calculating the correlation using the Pearson correlation coefficient between the students' scores on the item of each dimension and the total score for this dimension (numbers, algebra, geometry, probability). The correlation coefficients in all booklets ranged between 0.14 and 0.72 and were statistically significant at 0.01, which indicates the internal consistency of the test.

Reliability:

The Cronbach alpha reliability coefficient, the split-half method using the SPSS software, and the beta index (Almehrizi, 2021) using a R language software called Galpha Beta that were developed by the researchers were used to calculate the reliability indices for each booklet. The values of the alpha reliability coefficients ranged between 0.80 and 0.88, between 0.76 and 0.88 using split-half method, and between 0.76 and 0.86 using the beta index, which represent high reliability values.

Verifying Item Response Theory Assumptions

The item response theory assumptions were verified in the results of the TIMSS2019 mathematics test on the study sample, and the results were as follows:

First: Unidimensionality Assumption

The unidimensionality assumption for each booklet in mathematics was verified by the exploratory factor analysis using the SPSS software. It has been found that the total variance explained by the first factor in all booklets ranged between (18.51% and 24.50%), which are higher percentages than the rest of the factors. Furthermore, the value of the latent root of the first factor in all the booklets is higher than the rest of the factors, and by dividing the latent root of the first factor by the latent root of the second factor, the result was greater than 2 in all the booklets, and this is an indication of unidimensionality, as Glorfeld (1995) stated. Accordingly, the unidimensionality assumption was verified.

Second: Local Independence

According to item response theory, the test items are assumed to be locally independent, which means that the subject's response to one item is not affected positively or negatively by the subject's response to another item. If the assumption of local independence between items is violated, the local correlation between items will appear. Yen (1984) proposed a Q3 index to detect the local correlation between items, which is expressed as the relationship between the residuals of a pair of items after adjusting the estimated attributes. The local correlation of the items may be detected by calculating the correlations between the residuals of any two items at a certain capability level. If this relationship is greater than zero, this indicates the presence of a local correlation between the items. Accordingly, the values of these correlations approaching zero are considered an indication of local independence verification (AlNuaimi, 2006). The correlation coefficient values in all booklets ranged between 0.03 and 0.12. The average values of the Q3 index for each test booklet ranged between 0.05 and 0.07, which are very small values. Accordingly, we conclude that the questions in the mathematics booklets have met the local independence assumption.

Third: Model fit

Regarding the three-parameter model fit to the set of questions in the booklets, this was verified using the standardized residuals index, where the capabilities were estimated using the multilog software, and the residuals and the standardized residuals index were calculated and compared with the chi-square as an index of good fit of the items to the model. Table (2) shows the general index of conformity of each booklet with the three-parameter model. Table (2) shows that the fit of the three-parameters model was achieved with the booklet items, as the values of the standardized residuals index were not significant with a chi-square value of 3.84 at a significance level of 0.05 and a degree of freedom (1).

Table (2)
General index of the model fit to the TIMSS2019 mathematics test items

Booklet No.	General index	Booklet No.	General index
1	0.76	8	0.67
2	0.78	9	0.69
3	0.79	10	0.77
4	0.78	11	0.81
5	0.85	12	0.77
6	0.89	13	0.76
7	0.73	14	0.78

Estimates of TIMSS2019 mathematics test item parameters

The study used the TIMSS2019 mathematics test item parameters of (difficulty, discrimination, and guessing) that were rated by the International Association for the Evaluation of Educational Achievement (IEA). The values of the item discrimination parameters ranged between 0.54 as the lowest value, which was in the 21st and 7th items in the 10th and 11th booklets, respectively, and 2.29 as the highest value, which was in the 18th item in the 11th booklet. Regarding the difficulty parameters, the values ranged between -0.84 as the lowest value, which was in the 16th and 1st items in the 8th and 9th booklets, respectively, and 2.10 as the highest value, which was in the 27th and 14th items in the 4th and 5th booklets, respectively. Regarding guessing, the values ranged between 0 as the lowest value in 218 (55.90%) of the items and 0.48 as the highest value, which was in the 24th and 11th items in the 11th and 12th booklets, respectively.

Statistical processing:

To answer question (1): What are the descriptive statistics of the values of the Between Squared Residual indices (weighted and unweighted) in the eighth-grade mathematics test at TIMSS2019 in the Sultanate of Oman using the four test sections: content, cognitive process, item format and item difficulty? The descriptive statistics for WBSR and UBSR indices were calculated by R software using a code called Section SRes developed by the researchers, where the response file and question parameters file for each booklet are entered each time to calculate the values of the indices and their descriptive statistics and the weighted average through the TIMSS2019 mathematics test booklets (considering the difference in the number of items in the booklets) for both the means and standard deviations for the two indices.

To answer question (2): What is the percentage of individuals with misfit responses in the eighth-grade mathematics test at TIMSS2019 in the Sultanate of Oman according to the two Between Squared Residual indices using the four sections of the test: content, cognitive process, item format, and item difficulty? The question was answered through three statistical analyses:

Analysis (1): Calculating the number and percentage of individuals with misfit responses in the TIMSS2019 mathematics test for each index.

Analysis (2): Comparing the number and percentage of individuals with frequent misfit responses according to the number of test sections for each index.

Analysis (3): Studying the distribution of capability estimates for individuals with misfit responses for each index.

Regarding Analysis (1), For the first analysis, the number and percentage of individuals with misfit responses were calculated according to the WBSR and unweighted UBSR indices in the four sections of the test, and the individual's responses are considered misfit if the index value is greater than the value of 1.645, which is the critical score corresponding to the significance level 0.05 in the single-tailed normal distribution, as the Between squared Residual index follows the normal distribution.

To answer question (3): Does the percentage of individuals with misfit responses in the eighth-grade mathematics test at TIMSS2019 in the Sultanate of Oman differ when using the Squared Residual at the level of the item and when using the Between squared Residual at the level of the four test sections: content, cognitive process, item format, and item difficulty?

R language code called Item SRes, developed by the researchers, was used where the responses file is entered each time to calculate the value of the indices, to calculate the number of individuals with misfit responses according to the Squared Residual indices at the individual level, and then to calculate their percentage in relation to the total number of responses in each booklet. The response is classified as misfit if the index value is greater than 1.645, which is the critical value corresponding to the significance level of 0.05 in the one-tailed normal distribution.

Study Findings

Question (1): What are the descriptive statistics of the values of the Between Squared Residual indices (weighted and unweighted) in the eighth-grade mathematics test at TIMSS2019 in the Sultanate of Oman using the four test sections: content, cognitive process, item format and item difficulty?

Table (3) shows the weighted average values in the TIMSS2019 mathematics test booklets, considering the difference in the number of items in the booklets, for both the means and standard deviations for the WBSR and UBSR indices. Considering the general average of the means in all TIMSS2019 mathematics test booklets shows that they ranged between -0.30 and -0.23 and between -0.29 and -0.24 for the UBSR and WBSR indices, respectively, and that the lowest value of the weighted average was in the cognitive process's sections, where it reached -0.3 for the UBSR index and -0.29 for the WBSR index, while the largest value was in the question type sections, where it reached -0.23 for the UBSR index and 0.24 for the WBSR index.

Table (3)
Descriptive statistics for WBSR and UBSR indices for total mathematics test booklets

	UBSR			WBSR	
Sections	Mean	Standard Deviation		Mean	Standard Deviation
Content	-0.27	0.83		-0.25	0.84
Cognitive process	-0.30	0.77		-0.29	0.76
Item Format	-0.23	0.75		-0.24	0.78
Item Difficulty	-0.26	0.78		-0.26	0.73

The general average of the standard deviations in all TIMSS2019 mathematics test booklets for the UBSR and WBSR indices, respectively, shows that they ranged between 0.75 and 0.83 and between 0.73 and

0.84, and that the largest value was in the content section, where it reached 0.83, 0.84 for the UBSR and WBSR indices respectively. Whereas the lowest value was in Item Format in UBSR index, where it reached 0.75, and the lowest value was in Item Difficulty in WBSR index, where it reached 0.73.

Question (2): What is the percentage of individuals with misfit responses in the eighth- grade mathematics test at TIMSS2019 in the Sultanate of Oman according to the two Between Squared Residual indices using the four sections of the test: content, cognitive process, item format, and item difficulty?

Table (4) shows the total number and percentage of individuals with misfit responses according to the WBSR and UBSR indices for the TIMSS2019 mathematics test sections. The findings show that the multiplicity of sections contributed to recognizing a greater number of individuals showing misfit responses. Considering the total number of individuals with misfit responses in all booklets, it's noted that the number of individuals with misfit responses in the Item format section is greater than the rest of the sections, followed by the content sections, then the difficulty sections, and finally the cognitive processes sections.

Table (4)

Total number and percentage of individuals with misfit responses according to the WBSR and UBSR indices for the TIMSS2019 mathematics test sections

Sections	UBSR	WBSR
Content	(%3.16) 213	(%3.28) 221
Cognitive process	(%2.16)146	(%2.13) 144
Item Format	(%3.22) 217	(%3.88) 262
Item Difficulty	(%2.42) 163	(%2.51) 169

Regarding the second analysis, which determined the number and percentage of people who consistently provided misfit responses according to the number of test sections in which the misfit was detected in their responses, The ability of students, it turned out, had their misfit responses identified in only one of the four sections (content, cognitive process, item format, or item difficulty) without the rest of the sections. That is, individuals who show misfit responses according to content sections, for example, are different from those who show misfit responses according to other sections, as well as those who show misfit responses according to cognitive process sections, as well as those who show misfit responses according to other sections. and so on for the rest of the sections. The number of these individuals represents 651 students, representing 94% of the total students with misfit responses according to the UBSR index, and 652 students, representing 90%, according to the WBSR index, as for the students whose responses appear to be inconsistent according to two of the four sections, they represent 44 students, representing 6% according to the UBSR index, and 59 students, representing 8%, according to the WBSR index. The results also showed that there were no individuals with misfit responses that were repeated in three or four sections according to the UBSR index, while 12 students, representing 5%, appeared in three sections according to the WBSR index.

As for the third analysis, which is related to the ability levels of students who show misfit responses, the results showed that their abilities ranged between -2.04 and 1.64 according to the two indices, The abilities of individuals with misfit responses according to the UBSR index were more widely distributed in levels between (-1, 1) in the sections of content, cognitive process, and item difficulty, They were more widely distributed at levels (below -1) in item format sections in most booklets. While abilities for individuals with misfit responses according to the WBSR index were more widely distributed at levels between (-1.1) in content sections and cognitive processes in most booklets, while it was more widely distributed in the levels less than (-1) in the item format sections, while it was distributed in the levels between (-1, 1) and (less than - 1) in the difficulty sections. The results revealed that individuals with abilities less than (-1) ranged between 16% and 54% according to the UBSR index across the books, and between 17% and 65% according to the WBSR index, while average abilities (between -1 and 1) were between 44% and 82% according to the UBSR index, and 33% and 81% according to the WBSR index, High abilities (more than +1) varied between 2% and 5% according to the UBSR index, and 2% to 4% according to the WBSR index. These results reveal that individuals exhibiting misfit responses according to the two indices are of low and medium ability, and only a handful of high-ability students exhibiting misfit responses.

Question (3): Does the percentage of individuals with misfit responses in the eighth- grade mathematics test at TIMSS2019 in the Sultanate of Oman differ when using the Squared Residual at the level of the item and when using the Between squared Residual at the level of the four test sections: content, cognitive process, item format, and item difficulty?

Table (5) summarizes the number and percentage of participants who provided misfit responses according to the USR and WSR indices in the TIMSS2019 mathematics test booklets. The results show that the number of misfit responses according to the two squared residual indices constitutes a very small fraction of the total responses. The number of students reached 20, representing only 0.30% according to the USR index, and 64 students, representing 0.95% according to the WSR index. The results also show that there are no individuals with misfit responses in some booklets according to the USR and WSR indicators (Booklet 10 in the WSR and Booklets 9, 10 and 13 in the USR).

Table (5)

Number and percentage of individuals with misfit responses according to the USR and WSR indicators in the TIMSS 2019 mathematics test booklets

Summary	USR			WSR	
	Number And Percentage	Booklets		Number And Percentage	Booklets
Smallest Number of Students	(%0) 0	9،10،13		(%0) 0	10
Largest Number of Students	4 (0.82%)	4		14 (%2.88)	12
Total Students	20 (0.30%)	-		(0.95%) 64	-

The percentages of individuals with misfit responses according to the Squared Residual indices at the item level, WSR and USR appear small in comparison to the results of the Between Squared Residual indices over item level, UBSR and WBSR, where the percentages of misfit responses ranged between 2.16% and 3.22% according to the UBSR index, 3.88 according to the WBSR index; Which means that there is a difference between the two methods in the percentage of students with misfit responses in the test.

By studying the distribution of abilities of students who show misfit responses according to the USR and WSR indicators as appeared in table (5), the results showed that they are low abilities ranging between -4 and -2.10 according to the USR index, and between -4 and -1.94 according to the WSR index, with a small number of individuals with medium abilities ranging between -0.24 and 0.94 according to the USR index, numbering 8 individuals, and between -0.62 and 0.24 according to the WSR index, numbering 4 students, and according to the two indices, there are no individuals with high abilities (greater than 1). The results showed that the highest percentage of abilities for students who show misfit responses were low abilities (less than -1), representing 70% of the abilities according to the USR index, and 97% according to the WSR index, while the average abilities (between -1 and +1) reached 30% according to the USR index, and only 3% according to the WSR index, and there are no higher abilities among these abilities according to both indices.

It can be said that the Squared Residual and Between Squared Residual indices agreed that the misfit responses in the TIMSS2019 math test were among students with low and medium abilities, but they differed in the extent of these abilities, as it is found that the lowest ability of the Between Squared Residual index was -2.04 and the largest was 1.67, while the lowest ability of the Squared Residual index was -4 and the largest was 0.94.

Discussion of Results

The study aimed to determine the degree of prevalence of misfit or aberrant responses among eighth graders in the Sultanate of Oman in the TIMSS2019 math testing. through the application of Between Squared Residual index across the test sections associated with the cognitive structure of the test (content areas and levels of cognitive processes) and the associated sections test items (format and difficulty of items). The study's results demonstrated that there is a small deviation from zero in the arithmetic mean values of the weighted and unweighted indices and that its standard deviation departs from one. These results contradict those of Almehrizi's (2019) investigation. For every ability in the data, the two indices' arithmetic averages were extremely near to zero and the standard deviation values were nearly equal to the theoretical value of one, The study (Almehrizi, 2019) in which the two indices were applied to virtual generated data and others derived from field parameters to test verbal ability, in which skills are dispersed at certain values (-3, -2, -1, 0, 1, 2, 3) means that this is not regarded as a difference such that it follows a normal distribution. This study used field data, as students' abilities in mathematics testing in TIMSS2019 are distributed over a wide range of abilities to students in the 14 booklets, which distribute in a positive twisted distribution, i.e. the low abilities of students prevail.

According to the sections of the TIMSS2019 mathematics test, the study's results indicate that there are differences in the proportion of misfit responses from the Sultanate of Oman's total sample. Additionally, the data demonstrated that students who exhibit misfit responses are dispersed across the different sections of test, and do not participate in all sections but participating in a relatively small percentage of two sections and an even smaller percentage of three. This indicates the value of segmenting the test into multiple sections, since the variety of the sections identified instances that would not have been identified if one section was relied upon.

The results indicated that the item format portions had the highest percentage of misfit people, which was subsequently followed by the content sections then difficulty sections, and cognitive process sections. We can explain the higher percentage of misfit responses in the item format sections (short answer and multiple-choice questions) by pointing out that students are more likely to engage in guessing behavior in multiple choice questions than in short answer questions. Sonas et al. (2000, et al.) indicated that guessing is one of the main factors in the emergence of misfit response patterns (Bani Atta, 2019). When a student transfers the answer from a colleague more easily, it could also be the result of cheating because objective questions, such as multiple-choice questions, make this conceivable. It is also conceivable that the student was cheating because it is simpler for him to copy the answer from a coworker on objective questions, such as multiple-choice questions.

The results of the study also demonstrated that the values of the indices and their arithmetic averages in all booklets tend towards low values. Much of the responses in the mathematics test of the TIMSS2019 are normal, as evidenced by the low percentage of misfit responses across all booklets, which ranged between 2.16% and 3.22% according to the UBSR index and between 2.13% and 3.88% according to the WBSR index of the Sultanate of Oman's total sample.

The ability of the students who displayed misfit responses were found to have medium and low abilities. This may be because the ability of the items was from levels of application and reasoning, i.e., which are higher ability questions where low-level students typically turn to guesswork, as previously mentioned. Also, students with lower ability abilities displayed a higher number of misfit responses mostly in the sections containing item format (multiple choice and short answer), In which the possibility of the student resorting to guessing or cheating is high, especially if he is a low achiever. This is consistent with the results of Chen’s study (2004), which indicated that low achiever’s resort to guessing more than high achievers in an English language test administered to Chinese students. These results are also consistent with the research conducted by Brown & Villareal (2007), which demonstrated that low-achieving students show more misfit responses than high-achieving students in an adaptive test in mathematics.

The study found that there was a difference in the number and percentage of misfit responses according to the Squared Residual indices at the level of the test item and that at the level of the test sections, as their percentage was low according to the two Squared Residual indices at the level of the test item compared to those according to that at the level of the test sections. So, it can be concluded from this the importance of using indices at the level of sections for their ability to detect responses that the indices at the level of the individual cannot detect.

According to the Squared Residual indices at the item level, the students' abilities that displayed misfit responses were classified as low to medium. The students' abilities, which varied between low and medium according to Between Squared Residual over item level, also displayed misfit responses. The values of these abilities differed between the two methods, that is, the individuals whose responses were detected that are misfit according to the two methods were different. Thus, it can be said that the process of detecting misfit responses requires the use of indices at the level of the item and others at the level of the test sections, as the two methods help in detecting different individuals with misfit responses.

Recommendations and Proposals:

In light of the results reached, the current study recommends the following:

1. Using the Between Squared Residual index over item level to detect misfit responses in other tests such as Nationals tests and Pirls tests to compare its statistical properties with what was found in this study.
2. Using TIMSS2019 data in science for eight grade and TIMSS2019 data in math and science for fourth grade to study the misfit responses of students in the Sultanate of Oman and compare them with what was reached in this study.

3. Using other test sections to identify individuals with misfit responses keep in mind that these parts are meaningfully divided, and the items are not dispersed randomly among them.
4. Guiding teachers, both male and female, and those in charge of administering the TIMSS 2019 exams to training pupils on the fundamental mathematical abilities required, therefore minimizing the likelihood that they would revert to actions that result misfit responses.
5. Comparative research on the percentage of misfit responses according to squared residual index should be conducted among the various nations taking part in the TIMSS 2019 exams. This will help determine the degree of similarity and difference and establish a connection between it and the fluctuations in students' science and math proficiency.
6. Conducting studies, the causes of the misfit responses that students in the Sultanate of Oman in TIMSS 2019 tests.
7. Conducting further studies on the effect of the number of test sections and the number of items in each section on the statistical properties of the Between Squared Residual index.

References

1. Al Alawi, Sultan bin Nasser bin Saif. (2017). *The extent to which the topics of the International Trends in Mathematics and Science Study Test in Science Textbooks for Grades (5-8) in the Sultanate of Oman included* [Unpublished Master Thesis]. Sultan Qaboos University.
2. Abu Aish, Bassina Rashad Ben Ali. (2015). Personal factors and academic habits related to the variation in the achievement of students and second-grade intermediate students in mathematics and science in the Kingdom of Saudi Arabia in light of the results of the study of international trends in mathematics and science TIMSS2003. *Arab Journal of Social Sciences, Arab Foundation for Scientific Consulting and Human Resources Development, Egypt,* 16(50), 1-48.
3. Al Jarah, Bandar Nawaf. (2020). Detection of response patterns in science and mathematics tests for TIMSS 2015 data among a sample of Saudi and Singapore students using person fit indices. *Journal of Educational Sciences, King Saud University*, 23(2), 299-320.
4. Almehrizi, R. S. (2021). Coefficient beta as extension of KR-21 reliability for summed and scale scores. *Applied Measurment in Education, 34,* 139-149. DOI: 10.1080/08957347.2021.1890740
5. Almehrizi, R. S. (2019). Residual-based person fit statistics over test sections. *Journal of Educational and Psychological Studies*, *Sultan Qaboos University, 13*(4),687-702 .
6. Almehrizi, R. S. (2010). Comparison among new residual-based person fit indices and Wright's indices for dichotomous three-parameter IRT model with standardized tests. *Journal of Educational and Psychological Studies, Sultan Qaboos University*, 4(2), 14-26.
7. Almehrizi, R. S. (2004). *Investigating a new modification of the residual-based person fit index and its relationship with other indices in dichotomous item response theory* [Unpublished PhD Dissertation]. University of Iowa.
8. Al Nuaimi, Izz Al-Din Abdullah Awad. (2006). *The effect of violation of localized dependence on different estimates of item response theory* [unpublished doctoral thesis]. Yarmouk University.
9. AlShamrani, Saleh; AlShamrani, Said; Al-Barsan, Ismail; and Al-Darwani, Bakil. (2016). Highlights on the results of the Gulf countries in international trends in science and mathematics TIMSS 2015. *Education Evaluation Commission, King Saud University.*
10. Bani Atta, Zayed Saleh Ibrahim. (2019). Abnormal response patterns in the standardized Otis - Lennon test for the Jordanian environment and their impact on the accuracy of estimates of individual ability

and information function. *Journal of Educational and Psychological Studies, Sultan Qaboos University, 13(1), 27-45.*

11. Belov, D. I., & Armstrong, R. D. (2010) Automatic detection of answer copying via Kullback- Leibler divergence and K- Index. *Applied Psychological Measurement*. 34(6), 379-392.

12. Biernbaum, M. (1986). Effect of dissimulation motivation and anxiety on response pattern appropriateness measures. *Applied Psychological Measurement*, 10, 167-174.
13. Bracey, Gerald & Rudner, Lawrence M. (1992). Person-fit statistics: high potential and many unanswered questions. *Practical Assessment, Research & Evaluation,* 3(7), 1-6.
14. Brown, R., & Villareal, J. (2007). Correcting for person misfit in aggregated score reporting. *International Journal of Testing*, 7(1),1-25.
15. Chen, J. (2004). *Effect of test anxiety, time pressure, ability and gender on response aberrance*. [Doctoral dissertation]. The Ohio State University.
16. https://etd.ohiolink.edu/apexprod/rws_etd/send_file/send?accession=osu1092840837&disposition=inline
17. D'Costa, A. (1993a, April). Extending the Sato caution index to define the within and beyond ability caution indexes. Paper presented at convention of *National Council for Measurement in Education*, Atlanta, GA.
18. D'Costa, A. (1993b, April). The validity of the W, B and Sato Caution indexes. Paper presented at the *Seventh International Objective Measurement Conference*, Atlanta, GA.
19. Drasgow, F., Levine, M. V., & Williams, E. A. (1985). Appropriateness measurement with polytomous item response models and standardized indices. *British Journal of Mathematical and statistical psychology*, 38, 67-86.
20. Felt,J. M., Castaneda, R., TiemensmaJ., & Depaoli, S. (2017). Using person fit statistics to detect outliers in survey research. *Frontires Psychol*., 8, 1-9.
21. Ferrando, Pere J.; Chico, Elise. (2001). Detecting dissimulation in personality test scores: A comparison between person-fit indices and detecting scales. *Journal of Educational and psychological Measurement.* Rovira Virgili University, 61(6), 997-1012.
22. Glorfeld, L. W. (1995). An improvement on Horn's parallel Analysis methodology for selecting the correct number of factors to retain. *Educational and Psychological Measurements*, 55, 377-393.
23. Hamadnah, Iyad Mohammed. (2005). Investigation of the effectiveness of the Lz^{new} statistic in detecting the non-conforming response of the examined according to the item response theory. *Journal of Psychological and Educational Sciences, Al al-Bayt University, 16(3), 565-593.*
24. Huang, T. (2011). Robustness of BW aberrance indices against test length. Knowledge Management & E-Learning: *An International Journal*, 3 (3), 310
25. Iasonas, C., Bill, B. & David, W. (2000). *The consistency of examinee misfit across tests on the same subject and across subject: the case of the KS2 mathematics and science National Curriculum tests in England*. Retrieved from: http//www.man.edu.uk
26. Karabatosos, G. (2003). Comparing the aberrant response detection performance of thirty-six person-fit statistics, *Applied Measurement in Education,* 16(4), 277-298.
27. Mullis, I. V. S., & Martin, M. O. (Eds.). (2017). *TIMSS 2019 Assessment Frameworks. Retrieved from Boston College, TIMSS & PIRLS International Study Center website:*
28. http://timssandpirls.bc.edu/timss2019/frameworks/

29. Odaibat, Tasneem Numan. (2019). *Teaching Practices for Science Teachers for the Eighth Grade in Jordan and Singapore in Light of the Results of the International Trends in Mathematics and Science Study (TIMSS-2015)* [Unpublished Master's Thesis]. Yarmouk University.
30. Odeh, Amal Ahmed Hamad (2018). *Comparing the effectiveness of person fit indices according to the item response models of the dual-response item using the Gulf Scale for Multiple Mental Abilities*. [Unpublished Master Thesis], Sultan Qaboos University.
31. Shaqsi, Ya'qub; Abu Shandi, Yusuf; and Mehrezi, Rashid. (2019). The effectiveness of person fit indices in item response models when the strength of the local dependence between the items and the type of model features differ. *Journal of Educational and Psychological Studies, Sultan Qaboos University, 14(1), 41-53.*
32. Shehadeh, Fawz Hassan Ibrahim; Al qaramiti, Abu Alfotoh Mokhtar. (2016). The level of achievement of students in the Kingdom of Saudi Arabia in mathematics and science according to the results of international studies TIMSS compared to other countries from the point of view of teachers and supervisors: causes, solutions and treatment, methods of development. *Journal of Education, Al-Azhar University, 169, 326-370.*
33. Steinkamp, S. (2017). *Identifying aberrant responding: Use of multiple multiple measures*.[Unpublished doctoral dissertation]. University of Minnesota http:// .http://hdl.handle.net/11299/188885
34. TIMSS & PIRLS (n.d). *International Study Centre*. Boston College. Retrieved from: https://timssandpirls.bc.edu/index.html
35. Wright, B. D. (1977). Solving measurement problems with the Rasch model. *Journal of Education Measurement*, 114, 96-115.
36. Zaki, Medhat. (2011). TIMSS International Tests. Retrieved from: https://taalmnashet.ahlamontada.net/t314-topic