IEJME
OPEN ACCESS

INTERNATIONAL ELECTRONIC JOURNAL OF MATHEMATICS EDUCATION
e-ISSN: 1306-3030. 2019, Vol. 14, No. 3, 535-546
https://doi.org/10.29333/iejme/5755

Reliability and Validity Analysis of Statistical Reasoning Test Survey Instrument using the Rasch Measurement Model

Siti Shahirah Saidi [1], Nyet Moi Siew [1*]

[1] *Faculty of Psychology and Education, Universiti Malaysia Sabah, Sabah, MALAYSIA*

* CORRESPONDENCE: snyetmoi@yahoo.com

ABSTRACT

This study is an assessment of the reliability and validity analysis of Statistical Reasoning Test Survey (SRTS) instrument using the Rasch Measurement Model. The SRTS instrument was developed by the researchers to assess students' statistical reasoning in descriptive statistics among Tenth Grade science- stream students in rural schools. SRTS was a combination of a subjective test and an open-ended format questionnaire which contained of 12 items. The respondents' statistical reasoning was assessed based on these four constructs: Describing Data, Organizing Data, Representing Data and Analyzing and Interpreting Data. The sample comprised of 115 (76%) girls and 36 (24%) boys aged 15-16 years old from a rural district in Sabah, Malaysia. Overall, the SRTS instrument was found to have a high reliability with a Cronbach's alpha value (KR-20) of 0.81. Results also showed that SRTS has an excellent item reliability and high item separation value of 0.99 and 9.57 respectively. SRTS also has a good person reliability and person separation value of 0.81 and 2.04 respectively. Meanwhile, the validity of the SRTS instrument was appropriately established through the item fit, person fit, variable map, and unidimensionality. In conclusion, this study indicates that the SRTS is a reliable and valid instrument for measuring the statistical reasoning of science-stream students from rural secondary schools.

Keywords: statistical reasoning test survey, Rasch measurement model, tenth graders, rural schools

INTRODUCTION

Statistical reasoning, along with statistical literacy and statistical thinking are at the focus of interest and is one of the pertinent goals of learning outcomes in statistics education. Statistical reasoning as defined by Garfield and Chance (2000) is the way people reason with statistical ideas and make sense of statistical information. Studies related to statistical reasoning have been carried out extensively in other countries (Karatoprak, Karagöz & Börkan, 2014; Martin, 2013; Tempelaar, 2004; Ulusoy & Altay, 2017; Wang, Wang, & Chen, 2009) but in Malaysia, studies related to this field is progressively new in this decade. The current literature reveals that the level of Malaysian students' statistical reasoning is still poor and unsatisfactory (Chan & Ismail, 2013; Foo, Idris, Mohamed, & Foo, 2014; Ismail & Chan, 2015; "Misconceptions in Inferential Statistics", 2018; Zaidan *et al.,* 2012).

At the secondary school level, Chan and Ismail (2014) constructed an instrument which was modelled on the technology-based Geogebra software to assess the level of students' statistical reasoning in descriptive statistics. This instrument was developed based on the statistical reasoning construct proposed by Jones, Thornton, Langrall, Mooney, Perry and Putt (2000) and Mooney (2002), while the model of statistical reasoning by Garfield and Chance (2000) namely the Idiosyncratic, Verbal, Transitional, Procedural and

Article History: Received 7 February 2019 ◆ Revised 13 April 2019 ◆ Accepted 20 April 2019

Integrated Process was used to determine the level of students' statistical reasoning. This instrument is useful to assess the level of students' statistical reasoning in task-based interviews and small number of samples however it is not suitable for many samples, particularly in the study that utilizes a survey research method.

According to Garfield (1998), although one-to-one communication such as interviews or observations or examination of students' work such as statistical projects may be the best to assess students' statistical reasoning, a carefully designed paper-and-pencil instrument can also be employed to obtain information regarding students' statistical reasoning. Meanwhile, Karatoprak, Karagöz and Börkan (2015) asserted that qualitative methods are not practical for large groups of people. A survey method on the other hand is a more practical and systematic way to collect data and easier to administer and score. Besides that, it also provides an opportunity for researchers to gain more widely and comprehensive feedback from the respondents. Thus, due to these reasons, this study attempts to develop a survey research instrument by using the model of statistical reasoning proposed by Jones, Langrall, Mooney and Thornton (2004). The instrument which is known as the Statistical Reasoning Test Survey (SRTS), was specifically developed by the researchers to assess Malaysian Tenth Grade science-stream students' statistical reasoning in rural areas (Saidi & Siew, 2019).

In relevance to the Malaysian national mathematics achievement, the World Bank in 2010 reported that there was a gap in mathematics achievement between students in the urban and rural schools predominantly in poorer states like Sabah, where the urban school students achieved better results in mathematics than those in rural areas (Marwan, Sumintono & Mislan, 2012). Meanwhile, with regards to the assessment in the statistical learning conducted by Saidi and Siew (2019), the rural secondary school students in one of the district in Sabah were found to have a low level of understanding regarding the properties of measures of central tendency concept, where the students were unable to understand the concept of outliers in the data, as well as failed to understand which measures of central tendency could be used quantitatively and qualitatively. Besides that, the students were also found to have a difficulty in understanding the concept of representativeness in the measures of central tendency, since majority of the students were unable to provide which type of averages (mean, median, or mode) is the best to represent the data, either the data contained outliers or not. Since the students had a very poor understanding regarding the idea of representativeness and outliers in the measures of central tendency concept, majority of them failed to give the correct reasoning or justifications for the reasons why they chose a particular type of averages to best represent the data. These findings provided an early indication of students' poor level in statistical reasoning, particularly among the rural secondary school students in Sabah, Malaysia.

The Rasch Model is a psychometric technique that was developed to improve the precision of a constructed instrument, to monitor the quality of an instrument and compute the performances of respondents (Boone, 2016). It is the simplest model in the Item Response Theory (IRT) as it is a probabilistic model that assesses an item's difficulty and person's ability in such a way that they can be scored on the same continuous scale (Deane *et al.,* 2016). The Rasch Model estimates the probability of a person in choosing a particular item or category (Mahmud & Porter, 2015). The item difficulty and person ability in the Rasch Model are measured in a logit scale (Runnels, 2012).

The analysis from the Rasch Model can inform the researcher about the person and item reliability, item and person separation, as well as Cronbach's alpha value. Meanwhile, the construct validity of an instrument can be assessed through the item and fit, variable map and un-dimensionality. Thus, the abovementioned key concepts will be used by the researchers to establish the reliability and validity evidence of the SRTS instrument using the Rasch analysis.

RESEARCH METHODOLOGY

Instrumentation

The SRTS instrument is a combination of a subjective test and an open-ended format questionnaire which contains 12 items. It is developed by researchers based on the constructs of cognitive models of development proposed by Jones *et al.* (2004) to assess students' statistical reasoning. According to Jones *et al.* (2004), students' statistical reasoning can be assessed based on these four constructs, which are Describing Data, Organizing Data, Representing Data and Analyzing and Interpreting Data. Describing Data is related to the explicit reading of raw data or data presented in tables, charts, or graphical representations, while Organizing Data is related to arranging, categorizing, or consolidating data into a summary form. Representing Data is

Table 1. Distribution of Items in the SRST Instrument

Construct	Sub-process	Code	Item numbering	Task	Question
Describing Data	Showing awareness of display features	DD1	3a*	3	Examine the graphs carefully. What information do you get from the graphs?
	Identifying units of data values	DD2	2a*	2	Which quarter shows the highest value of the services imported to Netherlands from Malaysia? Please explain how to get the answer.
Organizing Data	Grouping or organizing data	OD1	1a***	1	Based on the data above, organize the data into the table below. Can you organize the data in different ways? Explain what you will do.
	Summarizing data in terms of measures of central tendency	OD2	2b*	2	What is the mean for the value of the services imported to Netherlands from Malaysia? Please explain how you determine the mean.
			2c*	2	What is the median for the value of the services exported to Malaysia from Netherlands? Please explain how you determine the median.
			2d*	2	What is the mode for the value of the services exported to Malaysia from Netherlands? Please explain how you determine the mode.
	Summarizing data in terms of measures of spread	OD3	3b*	3	What is the range of the number of books read by Form 2A students? Please explain how you determine the range.
			3c***	3	What is the standard deviation of the number of books read by 2B students in March? Explain how you determine the standard deviation.
Representing Data	Constructing a data display for a given data set	RD1	1b***	1	Based on the table in 1a, construct a histogram and frequency polygon graph in the graph paper provided at the last page using a scale of 2 cm to 8 gram amount of protein on the horizontal axis and 2 cm to 2 fast food sandwiches in the vertical axis. Explain how.
	Evaluating the effectiveness of data displays in representing data	RD2	1c**	1	In your opinion, which graph do you think represents the data better, the histogram or frequency polygon? Explain why.
Analyzing and Interpreting Data	Reading between data	AI1	3d**	3	Compare the distribution of the two graphs. Explain your answer(s).
	Reading beyond data	AI2	2e**	2	In your opinion, which type of average (mean, median, and mode) is the most suitable to be used to represent both sets of data? Explain why.

* *Item adapted from Mooney (2002)*
** *Item adapted from Chan and Ismail (2014)*
*** *New Item*

related to displaying data in a graphical form, while Analyzing and Interpreting Data is related to recognizing patterns and trends in the data and making inferences and predictions from data. These four constructs contain several sub-processes which guide educators and researchers to assess students' statistical reasoning.

The SRTS instrument aims to assess students' statistical reasoning among Tenth Grade science- stream students. The researchers adapted some of the items from Mooney (2002) and Chan and Ismail (2014) and at the same time constructed new items. The researchers' purpose for adapting items from these researchers is because of their suitability in the context of Malaysian Tenth Grade secondary school students. Mooney (2002)'s study assessed middle school students' statistical reasoning hence some of the items in the study would not be suitable for the upper secondary school level. Chan and Ismail (2014) provided more suitable items for the context of Malaysian upper secondary school students. However, the items in their study are technology-based which is not compatible with survey research. In spite of this, some of the items in the construct of Representing Data and Analyzing and Interpreting Data in Chan and Ismail (2014) could be applied for the current study which used a survey research method. **Table 1** shows the distribution of items in the SRTS instrument. This instrument has evidence of content validity as verified by an expert from a university.

There were three tasks in the SRTS instrument: Task 1, Task 2, and Task 3. Task 1 required the students to organize or group data from the raw data given (Item 1a) and expected the students to construct data displays from the grouped data created (Item 1b). The students' reasoning regarding which data displays were the best to represent the data was also assessed (Item 1c). The raw data in Task 1 were obtained from Chan and Ismail's (2014) instrument. Item 1a was a new item created by the researcher to assess the statistical reasoning in the sub-process of 'grouping or organizing data' in the Organizing Data construct. In order to identify whether the students' statistical reasoning in the 'grouping or organizing data' could extend to the Analytical level, a question was forwarded to the students as to whether they could organize the data in different ways. This item was familiar to the students and suitable for the context and level of Malaysian upper secondary school students. Item 1b was also created by the researcher for the same purpose (suitability of the context), as the Tenth-Grade secondary school students had prior knowledge on constructing or drawing a histogram and frequency polygon on a graph paper. Meanwhile, Item 1c was adapted from Chan and Ismail's (2014) study (e.g. Which graph do you think represents the data better, the histogram or the boxplot? Explain why).

Task 2 required the students to reduce the data using measures of central tendency (mean, median, and mode) from two groups of data, where one of the groups contained a significant outlier (Items 2b, 2c, and 2d). Besides that, it also required the students to identify the unit of data value based on the data given (Item 2a). Students' reasoning regarding which type of average was the best to represent both data was also assessed in Task 2 (Item 2e). Item 2a was adapted from Mooney's (2002) study (e.g. Which country won the most gold medals? How can you tell?). Item 2b, 2c and 2d were similarly adapted from Mooney (2002) which assessed the reasoning in the sub-process 'summarizing data in terms of measures of central tendency, in the Organizing Data Construct (e.g. What is the typical salary for the actress? How did you determine the typical salary?). Mooney (2002) used the word 'typical' in the item because middle school students might not be familiar with the word 'average'. Since the Tenth-Grade science-stream students in this study already knew the term 'average', thus, the terms mean, median and mode were used for Items 2b, 2c, and 2d respectively. Meanwhile, Item 2e was adapted from Chan and Ismail (2014) (e.g. Which measures of center is the most suitable to represent the score obtained by students? Explain why).

Task 3 required the students to reduce the data using measures of spread from a data display (Items 3b and 3c). Besides that, students were also required to make comparisons about the distribution of the two data displays (Item 3d). Students' awareness of display features was assessed in Task 3 (Item 3a). The data (in bar graphs), displayed different distributions where one was normal while the other not normal (skewed to the right) and was constructed by the researcher. Items 3a and 3b were adapted from Mooney (2002) (e.g. Examine the bar graph. What information did you get from the graph? How can you tell? What is the range of pets sold? How can you tell?). Meanwhile Item 3d was adapted from Chan and Ismail's (2014) study (e.g. Compare the distribution of both box plots with respect to shape, center, and variability). Item 3c is a new item created by the researcher to measure the reasoning in the sub-process 'Summarizing the data in terms of spread' in the Organizing Data construct, since the concept of standard deviation is taught to Tenth Grade science-stream students in the Additional Mathematics subject.

Corresponds to the four levels of cognitive thinking identified in the SOLO taxonomy model are the Prestructural, Unistructural, Multistructural, and Relational levels. Jones *et al.* (2004) formulated four level of students' statistical reasoning, namely Idiosyncratic (Level 1), Transitional (Level 2), Quantitative (Level 3), and Analytical (Level 4). The previous work by Jones *et al.* (2000) and Mooney (2002) also used the same level to characterize the level of children and middle school students' statistical thinking respectively. The Idiosyncratic level corresponds to the Prestructural level, where students are engaged in the task but could be distracted or misled by irrelevant aspects. The Transitional level corresponds to the Unistructural level, where students would only focus on a single relevant aspect. Next, the Quantitative level corresponds to the Multistructural level, where students focus on more than one relevant aspect of the task. Lastly, the Analytical level corresponds to the Relational level, where students can make links between relevant parts of the domain. Based on these features, this study formulated an initial framework to assess the level of students' statistical reasoning for each of the items in the SRTS instrument (**Table 2**).

Table 2. Initial Framework of the Statistical Reasoning Test Survey (SRTS)

Construct	Sub process	Idiosyncratic	Transitional	Quantitative	Analytical
Describing Data	Showing awareness of display features.	Shows no awareness to the displayed features (e.g. title and axis labels)	Shows only single awareness to the displayed features	Shows some awareness to the displayed features	Shows complete awareness to the displayed features
	Identifying units of data values	Unable to identify units of data values	Identifies only single unit of data values	Identifies some units of data values	Identifies units of general data values completely
Organizing Data	Grouping or organizing data	Unable to group or organize data into the classes	Group or organize the data into classes that are not consistent (have some flaws)	Groups or organize data into classes without flaws	Groups or organize data into classes without flaws and can group the data in more than one way
	Summarizing data in terms of center	Unable to summarize the data in terms of measures of central tendency	Have an idea about the measures of central tendency but unable to summarize using the valid measures	Summarizes the data using a measures of central tendency but have some flaws in the procedure	Summarizes the data using valid and correct measures of central tendency
	Summarizing data in terms of measures of spread	Unable to summarize the data in terms of measures of spread	Have an idea about the measures of spread but unable to summarize using the valid measures	Summarizes the data using measures of spread but have some flaws in the procedure	Summarizes the data using valid and correct measures of spread
Representing Data	Constructing a data display for a given data set	Unable to construct a data display for a given data set	Constructs a data display that is partially complete for a given data set	Constructs a data display for a given data set completely, but the display may have a few minor flaws	Constructs a data display for a given data set completely, with no flaws
	Evaluating the effectiveness of data displays in representing data	Provide the effectiveness of two different data displays for the same data set by using irrelevant features or reasons	Provide only single effectiveness of two different data displays for the same data set	Provide more than one effectiveness of two different data displays for the same data set	Provide a comprehensive effectiveness of two different data displays for the same data set data
Analyzing and Interpreting Data	Reading between data	Provides no or incorrect comparisons within and between data display or set	Provides only a single comparison within and between data display or set correctly	Provides local comparisons (e.g. shape/ center/ spread) within and between data display or set correctly	Provides global comparisons (e.g. shape, center, and spread) within and between data display or set correctly
	Reading beyond data	Provides inferences that are not based on the data or based on the irrelevant issues	Provides inferences that are partially based on the data. Some inferences may be only partially reasonable	Provides inferences primarily based on the data. Some inferences may be only partially reasonable	Provides reasonable inferences based on data and the context

Sample

The Rasch analysis was conducted based on the data collected from a pilot study with a total number of 151 Tenth Grade science-stream students from eight secondary schools in a rural district of Sabah, Malaysia. The students comprised of 115 (76%) girls and 36 (24%) boys aged 15 to 16 years old. In the Malaysian schooling system, the upper secondary school students who are academically inclined can choose between two main streams, either Science or Arts. Evidently, science stream students are more exposed to the statistical contents and mathematics related subjects

Procedure for Analyzing the Data

The items were analyzed using WINSTEPS version 3.73. Polytomous Rasch Model was used because the data for the SRTS instrument was in the form of polytomous data, where there are four possible scores of responses in all the items measuring the constructs in the SRTS instrument. They are "1" for Idiosyncratic, "2" for Transitional, "3" for Quantitative, and "4" for Analytical. Sumintono and Widhiarso (2015) stated that there are three fit indices criteria (**Table 3**) for establishing the reliability from the Rasch Model which are Cronbach's alpha, item and person reliability, and item and person separation.

Table 3. Reliability in Rasch Analysis

Statistics	Fit Indices	Interpretation
Cronbach's alpha (KR-20)	<0.5	Low
	0.5 – 0.6	Moderate
	0.6 – 0.7	Good
	0.7 – 0.8	High
	>0.8	Very High
Item and Person Reliability	<0.67	Low
	0.67 – 0.80	Sufficient
	0.81 – 0.90	Good
	0.91 – 0.94	Very Good
	>0.94	Excellent
Item and Person Separation		High separation value indicates that the instrument has a good quality since it can identify the group of item and respondent.

Source: Sumintono and Widhiarso (2015)

Table 4. Fit Indices for Item Fit

Statistics	Fit Indices
Outfit mean square values (MNSQ)	0.50 – 1.50
Outfit z-standardized values (ZSTD)	-2.00 – 2.00
Point Measure Correlation (PTMEA-CORR)	0.40 – 0.85

Source: Boone *et al.* (2014)

Meanwhile, the validity of the SRTS instrument using the Rasch Model can be established based on the analysis from the misfit order of the items. The logit which is produced from the Rasch analysis can give an indicator of the ability of a respondent in answering the items based on the item's difficulty (Olsen, 2003). According to Sumintono and Widhiarso (2015), item fit can inform the researcher whether the item is functioning normally in performing the supposed measurements, as well as to assess the suitability of the item. Moreover, it is indicated that the respondents had a misconception regarding the item if the item shows misfit. Boone, Staver and Yale (2014) and Bond and Fox (2015) suggested three criteria to be used for assessing the item fit, which are Outfit Mean Square Values (MNSQ), Outfit Z-Standardized Values (ZSTD), and Point Measure Correlation (PTMEA-CORR).

According to Bond and Fox (2007), Outfit MNSQ can inform the researcher about the suitability of the item in measuring the validity, while PTMEA-CORR informs the extent to which the development of the constructs has achieved its goals. A positive PTMEA-CORR value indicates that the item measured the construct to be measured, while a negative PTMEA-CORR value indicates otherwise. On the other hand, ZSTD are *t*-tests of the hypothesis which can inform the researcher whether the data perfectly fits the model. Any item that fails to fulfill these three criteria (**Table 4**) needs to be improved or modified to ensure the quality and suitability of the item (Sumintono & Widhiarso, 2015).

Besides that, the Rasch analysis also provides the researcher information of the person fit. Boone (2016) stated that the Rasch Model can identify a person fit based on the unusual response pattern. For instance, the unusual patterns that are detected by Rasch analysis suggests that the student may guess wildly, cheat, or is careless when answering the items. The criteria for assessing person misfit are based on the 'MEASURE', Outfit MNSQ, and Outfit ZSTD (Edwards & Alcock, 2019; Nevin *et al.,* 2015). According to Nevin *et al.* (2015), a high Outfit ZSTD value (> 2.0) coupled with a high MEASURE may indicate that a student with a high ability answered incorrectly on an 'easy' item. Meanwhile, a high Outfit ZSTD value (> 2.0) coupled with a low MEASURE may indicate that a student with a low ability answered correctly a 'difficult' item but incorrectly for the rest of items. According to Mohd Rahim and Norliza (2015), removing the misfit person from the Rasch analysis may improve the Rasch measurement scale such as its reliability.

In addition to the item fit and person fit, Variable Map (also called as Wright Map or Item-Person Map) which demonstrates the distribution of students' ability and item difficulty on a same logit scale -allows the researcher to identify if the items match the ability of the students (Bond & Fox, 2007). In the variable map, the item difficulty is listed on the right side of the map with the most difficult item placed on the top and the easiest item is placed at the bottom. Meanwhile, the person ability is listed on the left side of the map with the lower part for individuals with a low ability and the top is for individuals with a high ability. In other words, higher logits indicate persons with higher ability and more difficult items and vice versa (Iramaneerat, Smith & Smith, 2008).

Table 5. The Value for Person Reliability, Item Reliability, Person Separation, Item Separation and Cronbach's Alpha (KR-20) Value of the SRTS Instrument

Statistics	Value
Cronbach's alpha (KR-20)	0.81
Person Reliability	0.81
Item Reliability	0.99
Person Separation	2.04
Item Separation	9.57

Other than that, it is important to evaluate an instrument's unidimensionality to ensure whether it measures what it is supposed to measure (Abdul Aziz, Jusoh, Omar, Amlus, & Awang Salleh, 2014; Sumintono & Widhiarso, 2015), which is in this case, the construct of statistical reasoning. According to Ariffin, Omara, Isaa and Sharif (2010), the items which have been developed should test constructs which measures a single dimension only. The Rasch analysis uses the Principal Component Analysis (PCA) of the standardized residuals to measure to what extent the instrument's diversity measured what it is meant to measure. Sumintono and Widhiarso (2015) provided the criteria of unidimensionality based on the 'raw variance explained by measures' from the standardized residual variance. The value of 'raw variance explained by measures' which is higher than 20% is acceptable, higher than 40% is good, while higher than 60% is excellent. Meanwhile, the ideal value for the 'unexplained variance' should not exceed 15%.

FINDINGS

Reliability, Item and Person Separation

Table 5 shows the value for person reliability, item reliability, person separation, item separation and Cronbach's alpha (KR-20) value of the SRTS instrument based on the Rasch analysis in WINSTEPS. The value for person reliability is 0.81 with the person separation value of 2.04. Sumintono and Widhiarso (2015) stated that when the value of person reliability is higher than 0.80, it is 'good', while Bond and Fox (2007) stated that when the person reliability is higher than 0.80, this indicates a good and consistent response from the respondent. For the person separation, the value of 2.04 is interpreted as 'good', and this is supported by Linacre (2003) which stated that a good separation value of item difficulty is appropriate if the person separation value is higher than 2.00. Meanwhile, Krishnan and Idris (2014) stated that the person separation must be more than 1.00 to warrant that the students are measured across the spread.

In this study, the value for item reliability is 0.99 with an item separation value of 9.57. Sumintono and Widhiarso (2015) stated that an item reliability which is higher than 0.94 is interpreted as 'excellent'. Meanwhile, Bond and Fox (2007) stated that an item reliability value which is higher than 0.80 has a good value and is strongly acceptable, while a value less than 0.80 is less acceptable. As for the item separation value, the value of 9.57 is interpreted as high and fulfills the condition mentioned by Linacre (2003). Linacre (2003) asserted that an item separation value which is higher than 2.00 is interpreted as good. Meanwhile, Krishnan and Idris (2014) stated that an item separation value which is higher than 1.00 concludes that the items have enough spread.

Moreover, the Cronbach's alpha (KR-20) value which is 0.81 indicates that the SRTS instrument has a very high reliability of internal consistency (Sumintono & Widhiarso, 2015). Meanwhile, Bond and Fox (2007) stated that the value of Cronbach's alpha (which is based on the Rasch analysis approach) that ranges from 0.71 until 0.99 is acceptable as it is at the best level. Thus, this indicates that the SRTS instrument is highly suitable for the the actual research.

Item Fit

Table 6 presented the misfit order of the items based on the value of Outfit MNSQ, Outfit ZSTD and PT-MEASURE CORR. The bold figures indicate that the items failed to fulfill the criteria suggested by Boone *et al.* (2014). It was discovered that the item which was placed at the top (OD2c) tends to be misfit. Thus, this item is considered for change or removal. However, based on the three criteria to identify misfit items suggested by Boone *et al.* (2014), item OD2c fulfilled all the criteria for Outfit MNSQ (1.31), Outfit ZSTD (1.4), and PTMEA-CORR (0.48). Thus, item OD2c is retained and unchanged. Meanwhile, four items (OD1, OD2b, OD3a, and RD2) fulfilled at least one of the three criteria suggested by Boone *et al.* (2014), while the rest fulfilled all the criteria. According to Sumintono and Widhiarso (2015), the items which fulfilled at least one

Table 6. Misfit Order of the Items in SRTS

Item	MEASURE	Outfit MNSQ (0.50-1.50)	Outfit ZSTD (-2.0-2.0)	PTMEA-CORR (0.40-0.85)
OD2c	1.29	1.32	1.4	0.48
OD2a	-2.12	1.20	1.3	0.74
OD2b	-0.83	1.30	**2.4**	0.66
OD1	-1.74	1.36	**2.5**	**0.25**
AI2	1.96	0.89	-0.3	0.45
RD1	-1.74	0.97	-0.2	0.67
DD2	-0.90	0.94	-0.5	0.58
OD3b	0.93	0.71	-1.7	0.61
RD2	1.17	0.87	-0.6	**0.38**
AI1	1.21	0.64	-1.9	0.56
DD1	-0.05	0.59	**-3.6**	0.68
OD3a	0.83	0.50	**-3.5**	0.65

Table 7. Misfit Order of the Persons in SRTS

Person	Total Score (/48)	MEASURE	Outfit MNSQ (0.50-1.50)	Outfit ZSTD (-2.0-2.0)
F099	33	0.55	2.58	2.5
F082	25	-0.77	2.21	2.2
F085	27	-0.43	2.12	2.2
F004	29	-0.10	0.30	-2.2

of the criteria should be retained. Meanwhile, Abdul Aziz *et al.,* (2014) stated that the item is misfit if all the three criteria are out of the fit range. Thus, no items were changed and removed from the instrument.

Person Fit

Table 7 shows the person (which is the student in this case) whose response was most misfit with the Rasch analysis; or in other words, their response was different from the estimation given by the Rasch model. The students in the sample were coded accordingly - the F in F099 refers to the female while 099 was the student's number. The students were ordered according to the highest value of Outfit ZSTD. Based on **Table 7**, three students (F099, F082, and F085) scored an Outfit ZSTD value higher than 2.0 while one student (F004) had an Outfit ZSTD value lower than 2.0. The remaining students have an Outfit ZSTD value within the acceptable range (from -2.0 to +2.0). This indicates that in the pilot study, the items were suitable for almost all the students (97.35%) and the analysis conducted on those students showed quality findings for the assessment using the Rasch analysis.

A considerably high total score and MEASURE as performed by student F099 indicates that the individual most likely answered easy items incorrectly. This was indeed the case since for item RD1, students F099 scored only "2" while in fact item RD1 is regarded as an easy-to-answer item based on the Rasch analysis (- 1.74 logit). Meanwhile, student F082 and F085 have a low MEASURE but have an Outfit ZSTD value higher than 2.0 which may indicate that they answered a difficult item correctly, but incorrectly for other items. This is true since for student F082, she scored "3" for a quite difficult item DD1 (-0.05 logit), while student F085 scored "3" for a difficult item AI1 (1.21 logit). Furthermore, a large negative Outfit ZSTD value for student F004 (-2.2) is to be viewed as "too predictable" (Linacre, 2002).

Variable Map

Figure 1 presented the variable map which shows the distribution of persons (students) and items in a logit measurement scale. The variable map provides useful information on how the spread of item difficulty matches to the person ability (Sumintono & Widhiarso, 2015). Based on the right side of the variable map, item DD1 is calculated as being at the mean of the item difficulty estimates with a value of 0.00 logit. Six items spread above item DD1, while five items spread below it. It was realized that item AI2 was the most difficult item among the items in the SRTS instrument with a value of +1.96 logit, while item OD2a was the easiest item to be answered by the students in the pilot study with a value of -2.12 logit. This result was not improbable since item AI2 assessed the students' statistical reasoning in Analyzing and Interpreting Data (reading between data) – a question which is not usually presented in the statistical assessment within the Malaysian Mathematics syllabus. In contrast, item OD2a which is related to the mean concept was exposed

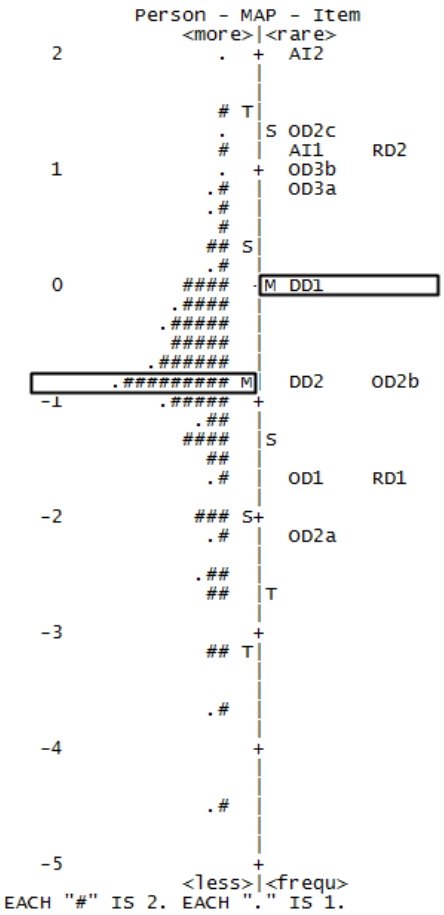

Figure 1. Variable Map of Person and Item

to students as early as fifth grade, which make it easier for Tenth Grade science stream students to solve this question.

The left side of the variable map shows the ability of students. On average (denoted by M in the line), the students were measured to have an ability below the 0.0 logit, which is -0.90 logit to be exact. Besides that, one student (F079), recorded the highest ability with the value of +1.94 logit, but exceeds the T (Two standard deviations) upper boundary, which indicates that this student has a different higher ability compared to the rest. Incidentally, six students exceeded the T lower boundary with the lowest three (F086, F0087, and F091) having recorded the value of -4.42 logit, which indicates that these three possessed the lowest ability among the rest of the students.

Based on the analysis from the variable map, it can be said that student F079 with the highest ability scored higher for all the items in the SRTS instrument. This is because student F079 has a +1.94-logit value, which almost matched the +1.96-logit value for the most difficult item in SRTS instrument, that is, item AI2. Contrarily, students F086, F0087, and F091 were unable to answer all the items since their ability (-4.42 logit) was still far below the easiest item in SRTS instrument, that is, item OD2a (-2.12 logit). Nonetheless, based on the spread of student ability and the spread of the item difficulty, some of the items (items placed above the 0.0 logit) are considered to be quite difficult by the students. Thus, actions will be taken by the researcher to reduce the difficulty of the items so that the items in the SRTS instrument are well targeted for the students in the study.

Unidimensionality

Based on **Figure 2**, the value for the 'raw variance explained by measures' is 61.9%. According to Sumintono and Widhiarso (2015), a value which is higher than 60% is 'excellent' and it indicates that the SRTS instrument has a strong evidence of unidimensionality, that is, the instrument undoubtedly measured

```
INPUT: 151 Person  12 Item  REPORTED: 151 Person  12 Item  4 CATS  WINSTEPS 3.73
-------------------------------------------------------------------------------

     Table of STANDARDIZED RESIDUAL variance (in Eigenvalue units)
                                              -- Empirical --     Modeled
Total raw variance in observations     =       31.5 100.0%          100.0%
  Raw variance explained by measures   =       19.5  61.9%           63.0%
    Raw variance explained by persons  =        5.3  17.0%           17.3%
    Raw variance explained by items    =       14.1  44.9%           45.7%
  Raw unexplained variance (total)     =       12.0  38.1% 100.0%    37.0%
    Unexplned variance in 1st contrast =        2.0   6.3%  16.6%
    Unexplned variance in 2nd contrast =        1.7   5.3%  13.9%
    Unexplned variance in 3rd contrast =        1.6   5.1%  13.5%
    Unexplned variance in 4th contrast =        1.4   4.6%  12.0%
    Unexplned variance in 5th contrast =        1.1   3.4%   8.8%
```

Figure 2. Standardized Residual Variance

the construct of statistical reasoning. Other than that, the unexplained variance for the 1st until 5th contrast is less than 10%, which falls in the ideal range value of less than 15%.

DISCUSSION AND CONCLUSION

Overall, the SRTS instrument has both a very high Cronbach's alpha (KR-20), and item and person reliability based on the analysis from the Rasch Model. This indicates that the SRTS instrument is an extremely reliable instrument for assessing students' statistical reasoning among the Tenth Grade science-stream students in rural schools, particularly in Sabah, Malaysia. The high item separation value indicates that the SRTS instrument has a greater spread of items (Klooster, Taal & Laar, 2008). Meanwhile, the high person separation value indicates that the students in the study can be well distinguished into three different abilities that is, high, medium, and low ability. Whether this is also the case for students in an urban school remains undiscovered, but it is suggested that the Rasch analysis on the SRTS instrument be conducted on a sample of students from urban schools.

In terms of validity, the researcher decided to preserve all the items since the items fulfilled at least one of the fit criteria for Outfit MNSQ, Outfit ZSTD, and PTMEA-CORR. Moreover, all of the items have a positive PTMEA-CORR value which indicates that the items move in one direction (Bond & Fox, 2015). On top of that, all the items have an Outfit MNSQ value within the acceptable range which indicates that the items are consistent with the item measurement. Bond and Fox (2007) stated that the value of Outfit MNSQ which is in the acceptable range is considered as good and productive for item measurement. For the person fit, only four students showed misfit, which indicates that the rest of the students provided a meaningful response for the Rasch analysis. Other than that, the SRTS instrument has a strong evidence of unidimensionality based on the result from the Standardized Residual Variance, and thus was an appropriate and legitimate choice of study on the students in the study.

ACKNOWLEDGEMENTS

The research was supported by the Universiti Malaysia Sabah (UMS), Malaysia under Grant No. FRG0462-2017. Any opinions, viewpoints, findings, conclusions, suggests, or recommendations expressed are the authors and do not necessarily reflect the views of the Universiti Malaysia Sabah, Malaysia.

Disclosure statement

No potential conflict of interest was reported by the authors.

Notes on contributors

Siti Shahirah Saidi – Faculty of Psychology and Education, Universiti Malaysia Sabah, Sabah, Malaysia.

Nyet Moi Siew – Faculty of Psychology and Education, Universiti Malaysia Sabah, Sabah, Malaysia.

REFERENCES

Abdul Aziz, A., Jusoh, M.S., Omar, A.R., Amlus, M.H., & Awang Salleh, T.S. (2014). Construct Validity: A Rasch Measurement Model Approaches. *J. Appl. Sci. & Agric., 9*(12), 7-12.

Ariffin, S. R., Omara, B., Isaa, A., & Sharif, S. (2010). Validity and Reliability Multiple Intelligent Item Using Rasch Measurement Model. *Procedia Social and Behavioral Sciences, 9*, 729-733. https://doi.org/10.1016/j.sbspro.2010.12.225

Bond, T. G., & Fox, C. M. (2007). *Applying The Rasch Model: Fundamental Measurement in the Human Science* (2nd Ed). New Jersey: Lawrence Erlbaum.

Bond, T. G., & Fox, C. M. (2015). *Applying The Rasch Model: Fundamental Measurement in the Human Science* (3nd Ed). New Jersey: Lawrence Erlbaum. https://doi.org/10.4324/9781315814698

Boone, W. J. (2016). Rasch Analysis for Instrument Development: Why, When, and How? *CBE Life Sciences Education, 15*(4), rm4. https://doi.org/10.1187/cbe.16-04-0148

Boone, W. J., Staver, J. R., & Yale, M. S. (2014). *Rasch Analysis in the Human Sciences*. Dordrecht, Netherlands: Springer. https://doi.org/10.1007/978-94-007-6857-4

Chan, S. W, & Ismail, Z. (2014). A Technology-Based Statistical Reasoning Assessment Tool in Descriptive Statistics for Secondary School Students. *The Turkish Online Journal of Educational Technology, 13*(1), 29-46. https://doi.org/10.1016/j.sbspro.2013.10.067

Chan, S. W. & Ismail, Z. (2013). Assessing misconceptions in reasoning about variability among high school students. *Procedia - Social and Behavioral Sciences, 93*, 1478-1483. https://doi.org/10.1016/j.sbspro.2014.03.658

Chan, S. W., Ismail, Z., & Sumintono, B. (2014). A Rasch Model Analysis on Secondary Students' Statistical Reasoning Ability in Descriptive Statistics. *Procedia - Social and Behavioral Sciences 129*, 133-139.

Deane, T., Nomme, K., Jeffery, E., Pollock, C., & Birol, G. (2016). Development of the Statistical Reasoning in Biology Concept Inventory (SRBCI). *CBE— Life Sciences Education*, *15*, ar5. https://doi.org/10.1187/cbe.15-06-0131

Edwards, A., & Alcock, L. (2010). Using Rasch analysis to identify uncharacteristic responses to undergraduate assessments. *Teaching Mathematics and its Applications, 29*(4), 165-175. https://doi.org/10.1093/teamat/hrq008

Foo, K. K., Idris, N., Mohamed, I., & Foo, S. L. (2014). A multiple regression model of statistical reasoning: A Malaysian context. *OIDA International Journal of Sustainable Development*, *9*(10), 59-70.

Garfield, J. B. (1998). The Statistical Reasoning Assessment: Development and Validation of a Research Tool. In: L. Pereira-Mendoza, L. Seu Kea, T. Wee Kee, & W.K. Wong (eds.), *Proceedings of the Fifth International Conference on Teaching Statistics*, 781- 786. Singapore: International Statistical Institute.

Garfield, J., & Chance, B. (2000). Assessment in Statistics Education: Issues and Challenges. *Mathematics Thinking and Learning*, *2*(1&2), 99-125. https://doi.org/10.1207/S15327833MTL0202_5

Iramaneerat, C., Smith, Jr. E. V., & Smith, R.M. (2008). An introduction to Rasch measurement. In J.W. Osborn (Ed.). *Best practices in quantitative methods* (pp. 50-70). Thousand Qaks, California: Sage Publications, Inc. https://doi.org/10.4135/9781412995627.d6

Ismail, Z. & Chan, S. W. (2015). Malaysian Students' Misconceptions about Measures of Central Tendency: An Error Analysis. *AIP Conference Proceedings*, 1643, 93. https://doi.org/10.1063/1.4907430

Jones, G. A., Thornton, C. A., Langrall, C.W., Mooney, E. S., Perry, B., & Putt, I. A. (2000). A framework for characterizing children's statistical thinking. *Mathematical Thinking and Learning*, *2*, 269–307. https://doi.org/10.1207/S15327833MTL0204_3

Jones, G.A., Langrall, C.W., Mooney, E.S. & Thornton, C.A. (2004). Models of development in statistical reasoning. In D. Ben-Zvi and J. Garfield (Eds). *The Challenge of Developing Statistical Literacy, Reasoning, and Thinking,* pp. 97–117, Dordrecht: Kluwer Academic. https://doi.org/10.1007/1-4020-2278-6_5

Karatoprak, R. Karagöz, G. & Börkan, B. (2015). Prospective elementary and secondary school mathematics teachers' statistical reasoning. *International Electronic Journal of Elementary Education, 7*(2), 107-124.

Klooster, P. M., Taal, E., & Laar, M. A. F. J. (2008). Rasch Analysis of the Dutch Health Assessment Questionnaire Disability Index and the Health Assessment Questionnaire II in Patients with Rheumatoid Arthritis. *Arthritis & Rheumatism (Arthritis Care & Research), 59*(12), 1721-1728. https://doi.org/10.1002/art.24065

Krishnan, S. & Idris, N. (2014). Investigating Reliability and Validity for the Construct of Inferential Statistics. *International Journal of Learning, Teaching and Educational Research, 4*(1), 51-60.

Linacre, J. M. (2002). Understanding Rasch Measurement: Optimizing rating scale category effectiveness. *Journal of Applied Measurement, 3*(1), 85-106.

Linacre, J. M. (2003). Dimensionality: Contrasts & variances. *Help for Winsteps Rasch Measurement Software*. Retrieved from http://www.winsteps.com/winman/principalcomponents.htm

Mahmud, Z., & Potter, A. L. (2015). Using Rasch analysis to explore what students learn about probability concept. *Indonesian Mathematical Society Journal on Mathematics Education*, *6*(1), 1-11.

Martin, N. (2013). *Gender Differences in Statistical Reasoning: A Multipronged Approach* (Unpublished PhD Thesis), University of Waterloo.

Marwan, A., Sumintono, B., & Mislan, N. (2012). Revitalizing Rural Schools: A Challenge for Malaysia. In: Educational Issues, Research and Policies (172-188). RMC-UTM Press: Skudai, Johor Bahru,

"Misconceptions in Inferential Statistics." (2018). Malaysian College Students Misconceptions in Inferential Statistics. *International Journal of Pure and Applied Mathematics, 118*(24), 1-14. Retrieved on 10 November 2018 from https://acadpubl.eu/hub/2018-118-24/1/22.pdf

Mooney, E. S. (2002). A framework for characterizing middle school students' statistical thinking. *Mathematical Thinking and Learning*, *4*(1), 23-63. https://doi.org/10.1207/S15327833MTL0401_2

Nevin, E., Behan, A., Duffy, G., Farrel, S., Harding, R., Howard, R., Raighne, A., & Bowe, B. (2015). *Assessing the validity and reliability of dichotomous test results using Item Response Theory on a group of first year engineering students*. The 6th Research in Engineering Education Symposium (REES 2015), Dublin, Ireland, July 13-15.

Olsen, L. W. (2003). *Essays on George Rasch and his Contributions to Statistics*. Unpublished PhD Thesis, University of Copenhagen.

Runnels, J. (2012). Using the Rasch model to validate a multiple choice English achievement test. *International Journal of Language Studies (IJLS)*, *6*(4), 141-153

Saidi, S. S., & Siew, N. M. (2019). Assessing Students' Understanding of Measures of Central Tendency and Attitude towards Statistics in Rural Secondary Schools. *International of Electronic Journal of Mathematics Education, 14*(1), 73-86. https://doi.org/10.12973/iejme/3968

Sumintono, B., & Widhiarso, W. (2015). *Aplikasi Pemodelan Rasch Pada Assessment Pendidikan*. Cimahi: Trim Komunikata Publishing House.

Tempelaar, D. (2004). Statistical Reasoning Assessment: An Analysis of the SRA Instrument. Proceedings of the ARTIST Roundtable Conference on Assessment in Statistics, Lawrence University.

Ulusoy, A., & Altay, K. (2017). Analyzing the statistical reasoning levels of pre-service elementary school teachers in the context of a model eliciting activity. *International Journal of Research in Education and Science (IJRES), 3*(1), 20-30. https://doi.org/10.21890/ijres.267363

Wang, W., Wang, X., & Chen, G. (2009). Survey and Analysis of the Statistical Reasoning among High School Students in China and Dutch. *Journal of Mathematics Education, 2*(1), 15-26.

Zaidan, A., Ismail, Z., Mohamad Yusof, Y., & Kashefi, H. (2012). Misconceptions in descriptive statistics among postgraduates in social sciences, *Procedia-Social and Behavioral Sciences*, *46*, 3535–3540. https://doi.org/10.1016/j.sbspro.2012.06.100