Malaysian Journal of Social Sciences and Humanities (MJSSH) (e-ISSN: 2504-8562)
2022, Volume 7, Issue 3, e001387
DOI: https://doi.org/10.47405/mjssh.v7i3.1387

Rasch Analysis for Thinking Levels of Sustainability Thinking Skills Instrument Among Secondary School Students

Almasiah Hassan[1*], Kamisah Osman[2]

[1]Faculty of Education, Universiti Kebangsaan Malaysia (UKM), 43600, Bangi, Selangor, Malaysia.
Email: almasiahhassan@gmail.com
[2]Faculty of Education, Universiti Kebangsaan Malaysia (UKM), 43600, Bangi, Selangor, Malaysia.
Email: kamisah@ukm.edu.my

CORRESPONDING AUTHOR (*):
Almasiah Hassan
(almasiahhassan@gmail.com)

KEYWORDS:
Sustainability thinking
Secondary school students
STEM education
Validity
Reliability Rasch Measurement Model

CITATION:
Almasiah Hassan & Kamisah Osman. (2022). Rasch Analysis for Thinking Levels of Sustainability Thinking Skills Instrument Among Secondary School Students. *Malaysian Journal of Social Sciences and Humanities (MJSSH), 7*(3), e001387. https://doi.org/10.47405/mjssh.v7i3.1387

ABSTRACT

An Instrument of sustainability thinking was developed to measure students' level towards sustainability thinking. Instrument of Sustainability Thinking Skills (STSI) was developed to measure the level of sustainability thinking towards STEM and non-STEM education among secondary school students. STSI consisted of 51 items measuring the four constructs of curiosity towards sustainability, namely sustainability thinking, green technology in sustainability, sustainable behavior (environmental value) and sustainable behavior (general ecological behavior). A total of 471 form 4 students aged 16 to 17 years enrolled in six urban and non-urban schools in Malaysia, were involved as research sample Rasch Measurement Model was applied to determine the validity and reliability of STSI. The results of the validity analysis found that the polarity of the items through the PTMEA-CORR values showed that all 51 items were >0.00 (+). Through item fit analysis, all items were retained because items meet the requirements of the range in MNSQ outfit, ZSTD outfit and PTMEA-CORR. Results also showed that STSI has an excellent item reliability and moderate high item separation value of 0.99 and 7.53 respectively. STSI also has a high person reliability and person separation value of 0.98 and 5.72 respectively. In conclusion, STSI has good validity and high reliability in measuring curiosity towards STEM Education among form 4 students in secondary schools.

Contribution/Originality: This study originates new formula to demonstrate analysis, using rasch analysis to exhibit and explain in detail about the sustainability thinking skills instrument of secondary school students in Malaysia. This study is one of very few studies which have investigated the sustainability thinking skills of secondary school students in Malaysia in detail and accurately. The paper contributes the first logical analysis using Rasch measurement capital.

1. Introduction

Education is seen as the key to more sustainable and sustainable development. In parallel, the United Nations introduced Education for Sustainable Development also known as Education for Sustainable Development (ESD) in 1983 (United Nations, 2020). Education for Sustainable Development (ESD) emphasizes sustainable development issues in the teaching and learning process including climate change, disaster risk reduction, biodiversity, poverty reduction, and sustainable consumption. Information related to these sustainable issues should be channeled at the school level to produce a generation that is more grateful and loves the benefits of the environment that is preserved for the survival of future generations. In line with that, Sustainability Education through green technology has been introduced in the national curriculum because of the ability of sustainability in acting as a medium that can provide a mold of innovative and creative thinking students who in turn can produce an inventive society.

One of the factors that can influence a student to become an active individual is the nature of thinking. Literally, the nature of thinking refers to the nature of the individual who wants to seek, study, and acquire knowledge. Raharja et al. (2018) define the nature of thinking as the desire to fill the mind with new information without expecting appreciation or even extrinsic factors. The nature of thinking arises as a result of conflict or uncertainty that occurs (Reio & Petrosko, 2013) and acts as a motivation that encourages exploratory behavior (Berlyne, 1960; Day, 1968; Loewenstein, 1994). In the context of this study, the nature of thinking is defined as a positive emotion-motivation system that encourages students to explore more actively and receive information better (Kashdan et al., 2004, 2009).

Without thought, the process of scientific activity, exploration and innovation will not take place (Kashdan et al., 2009). Through this concept highlighted, clearly shows the relationship between thinking and innovation that can be created through STEM Education (Science, Technology, Engineering and Mathematics). This is because the thinking will encourage students to learn the technical and design skills emphasized in STEM Education (Committee on STEM Education, 2018; Jin & Bierma 2013; Kennedy & Odell, 2014; McDonald, 2016; Zollman, 2012) as well as the creation of innovations that combine all four STEM fields (Bybee, 2010; Foster et al., 2010).

Sustainability thinking is a way for the process of sustainability learning and lifelong development (Hurst & Ludwig, 2017). In the context of this study, sustainability thinking acts as a desire that motivates an individual to gain new knowledge and experience of exploring the STEM field on their own. This view is supported by Kashdan et al. (2009) who stated that sustainability thinking is measured based on two aspects namely the aspects of exploration and acceptance. The exploratory construct refers to the search for new knowledge and experience while the acceptance construct is a person's willingness to accept the originality, uncertainty or unexpectedness that occurs in daily life. Figure 1 shows the contribution to the literature as an objective shout out and intention of the journey for this instrument.

Figure 1: Contribution to the Literature

- To raise awareness to the disposition level of Sustainability Thinking.
- The authors have clarified the concept "sustainability thinking", in the aspect of disposition framework
- The study revealed the need to develop an instrument to measure students' thinking to sustainability.
- The authors have developed an instrument with good psychometric properties to measure students' thinking level of sustainability.
- This study has the potential to generate more knowledge and literature on students' STSI. There are very few empirical studies in this regard.

Various studies have been made on thinking and its significance in the context of STEM teaching and learning. Sustainability thinking in students encourages students to explore opportunities and challenges and encourages active involvement in STEM fields (Garrosa et al., 2017). In fact, sustainability thinking should be developed because of its ability in resolving a STEM issue that needs a solution. In addition, sustainability thinking can also train students' minds to be more active, make students an active observer, open up new worlds and attract students to learn more deeply (Baumgarten, 2001). This is in line with the context of STEM Education where students with high curiosity will be encouraged to explore the four areas of STEM more comprehensively.

The need to inculcate sustainability thinking towards STEM among students is very significant. The study of Ludwig et al. (2019) proved that sustainability thinking can develop students 'STEM literacy in order to face the challenges of daily life related to the STEM field (Axon & James, 2018). Furthermore, sustainability thinking that encourages students to seek knowledge through self -access and explore experiences will help provide a comprehensive learning medium for a student (Tseng et al., 2013). In addition, students with a sustainability mindset will also be motivated to pursue science and explore comprehensive STEM knowledge and skills (Committee on STEM Education, 2018; McDonald, 2016).

The Malaysia Education Blueprint Report 2013-2025 focuses on the implementation of STEM as the basis in providing sufficiently trained STEM graduates to meet the job market that drives the Malaysian economy. Therefore, the emphasis on sustainability thinking towards STEM among students must be implemented explicitly starting at the lowest level to produce human capital with STEM sustainability skills that are able to solve global problems, make decisions and create for the benefit of future society. This clearly proves that there is a need to study students 'sustainability thinking towards STEM and an instrument that measures students' sustainability thinking skills towards STEM among secondary school students needs to be developed.

2. Literature Review

Based on the fact of the importance of providing critical thinking human capital and being able to create new ideas, the secondary school sustainable curriculum has set its main objective to stimulate students' sustainable thinking and develop interest in the world around them (Curriculum Development Division, 2016a). Essentially, sustainability thinking is a positive emotional experience (Hurst, 2015). In sustainable education, sustainability thinking makes a learning more meaningful and able to achieve its objectives (Ludwig, 2017).

According to Hurst (2015), sustainability thinking should be started at the school level again through a medium that can drive the knowledge and skills to create students that is sustainable education. Exposure to sustainability thinking at an early stage can: (a) build a learning framework for the development of students 'minds (b) promote critical thinking and reasoning skills; (c) cultivate students' interest in sustainability thinking; (d) develop sustainability thinking; and (e) provide students with experiences about the world around them (Axon & James, 2018; Anastas & Zimmerman, 2018; Hurst, 2015; National Research Council, 2011).

Recognizing the importance of sustainability thinking, the Ministry of Education Malaysia (MOE) has taken pragmatic steps through the emphasis on sustainability thinking in the learning curriculum. Sustainability thinking is one of the Scientific Attitudes and Pure Values in the secondary school Science and Pure Science curriculum and is targeted as one of the profiles of students who want to be born through 21st century learning (Curriculum Development Division, 2016b). Sustainability thinking is so significant that it is included as a major objective in the formulation of the secondary school science curriculum. Students with a sustainability mindset will explore new ideas and information, conduct inquiries, be independent in learning as well as enjoy experiences throughout learning (Kashdan et al., 2004).

Although sustainability thinking is seen to have a significant impact on student learning, motivation, and creativity (Hurst & Ludwig, 2017; Axon & James, 2018) studies on sustainability thinking among school students are not yet comprehensive (Hurst & Ludwig, 2017). A study by Hamid et al. (2017) also proves that the level of sustainability thinking of STEM secondary school students in Malaysia is still low compared to the neighboring country, Brunei. This is quite worrying because based on the PISA 2018 Achievement Report, the average achievement score of science literacy for Malaysian students is 438, slightly higher than Brunei's 431 (Avvisati et al., 2018; Schleicher 2019).

Therefore, Hurst et al. (2019) suggested that specific instruments continue to be developed because previous instruments only measured the nature of sustainability in general. There is no instrument to measure students 'sustainability thinking skills towards STEM Education specifically. Thus, a sustainability thinking skills questionnaire instrument to measure the level of sustainability thinking skills for STEM Education was developed. Accordingly, this study was conducted to obtain empirical evidence on the validity and reliability of the Questionnaire instrument of students' sustainability thinking skills in STEM Education (STSI-STEM) by using the Rasch Measurement Model (MPR).

As a result, the research gap can be resolved by conducting psychometric feature testing on the development of instrument items and applying rigorous empirical analytic techniques such as the Rasch model. The Rasch model has undoubtedly attracted the attention of numerous researchers' both domestically and internationally to validate the item on their instrument's development (Balsamo et al., 2014; Othman et al., 2014). This article examines the validity and reliability of measuring instruments using the Rasch model's three core assumptions, namely item fit, unidimensionality, and local independence.

The primary objective of this research is to ascertain the sustainability thinking that secondary school students exhibit when practicing sustainability in their daily lives. Thus, it is thought that endorsing an item for each STSI construct using the Rasch model is capable of improving item quality measurement.

2.1. Theoretical Framework of Sustainability Thinking

The theory of sustainability thinking founded by Hurst arose from the study of sustainability that affects human behavior and behavior. Hurst (2013) states that sustainability thinking is one of the motivational components that influences the development of opportunity in the individual. In Hurst's Theory of Sustainability (Hurst 2013, 2015), Hurst introduces three ideas about sustainability thinking namely stimuli in triggering sustainability thinking, four dimensions in sustainability thinking and two types of exploratory tendencies.

The first idea introduced by Hurst was the thought of self -sustainability would motivate a person to be proactive and have behaviors that are driven by stimuli from within. It is this internal stimulus that will indicate activities such as renewal, uncertainty, and conflict. A sense of uncertainty arises when an individual experiences something new, surprising, or complex. This condition will cause high stimulation in our central nervous system. The human response when faced with an uncertainty is what is said to be thought. Sustainability thinking will encourage humans to form behaviors that can reduce feelings of uncertainty (Gagné, 1985).

The theoretical framework of the study describes the overall study of IKPK-STEM construction. In this study, the sustainability framework by Hurst (2015), Ludwig's (2017) sustainability framework, Core Science and Pure Science Curriculum and Assessment Standards Document (DSKP), Secondary School Standard Curriculum and other literature materials were used to obtain appropriate constructs for IKPK-STEM. This study refers to works Hurst (2015) and Ludwig (2017) to determine sustainability skills. Hurst conducted a review of the academic literature and concluded the following concepts as core skills of sustainability thinking. The concepts are sustainability thinking, green technology in conserving nature, environmental values, and general ecological behavior. Based on this review, this study summarizes two sustainability thinking skills and two attitudes contribute to sustainability.

According to Ludwig et al. (2019), someone who has a sustainability mindset is likely to be driven by a combination of several of these dimensions. Figure 1 shows a dimensional map based on Hurst's Theory of Sustainability Thinking. The characteristic features of constructs or attitudes towards IKPK-STEM obtained from literature sources and documents are categorized into 4 main constructs. In a recent study, Mahaffy (2019) emphasized students 'learning behaviors including collaboration, perseverance, communication, creativity and playfulness and problem solving, assessed during play-while-learning sessions, while using sustainability to voice, question and investigate students 'understanding of the world of technology. This construct explores how students use the innovative thinking they think to contribute to society through presentation, communication and questioning the digital world. As a result, instead of using existing technology, such as surfing the internet or text, a programmer can create and adapt the technology to solve problems in new ways (Grover & Pea, 2013; Curriculum Development Division, 2015).

Sustainability thinking is an approach to problem solving. In short, this is the process of finding a solution to a problem. According to Holme (2018), A systems thinking approach towards green chemistry education and sustainability (and chemistry education more broadly). This is in line with Jonassen (2000) who stated that the problem-solving process requires cognitive as well as affective and conative elements. Affective and conative elements such as attitudes, feelings, beliefs, motivations, and values, reinforce individual tendencies during the problem -solving process. Further in this study, skills refer to the ability to use knowledge based on sustainability modules i.e., concepts, techniques, and methods to solve problems. Such attitudes refer to character traits, values, feelings, and motivations that are relevant to the problem -solving process based on sustainability thinking.

The Green and Sustainable Chemistry dimension is a holistic thinking about the sustainability of new materials or knowledge. This dimension is more closely related to cognitive traits (thinking) than experience. The combination of the dimensions of Green and Sustainable Chemistry requires an information or knowledge through exploration. The combination of Strategies and Tools dimensions is a more dynamic system for finding information or knowledge, but more focused on preparing for future sustainability -related phenomena that are specific, such as solving an issue or scientific study.

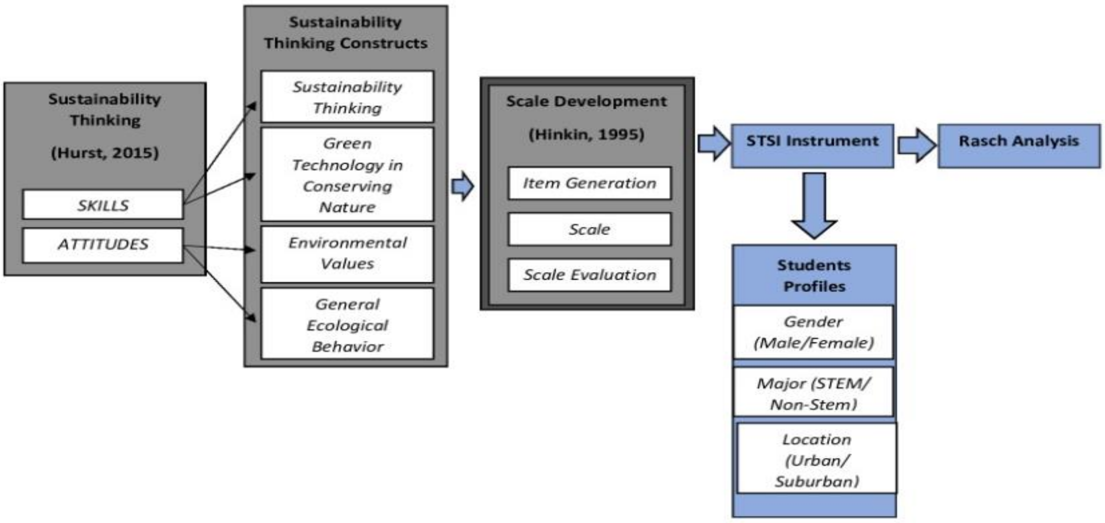
Ludwig et al. (2019) explained that these four dimensions are not exclusive; that is, a person may have all four dimensions at one time depending on emotional and environmental factors. Accordingly, it is not impossible also if an individual is more inclined to one of these dimensions.

This study also suggests sustainable behavior as another element of sustainability thinking attitudes. The definition of behavior comes from many scholars who are psychologists. According to Vaughan (2002) state that behavior is a way of thinking. The views of Vaughan (2002) about this behavior is found to have a relationship with the view of Zohar and Marshall (2000) who define behavior as intelligence to solve problems of meaning and value, to place one's behavior and life in the context of broader meaning and to evaluate actions or ways of life someone so more meaningful than others. To measure the height of a person's behavior, Zohar and Marshall (2000) presented behavioral indicators that are: "the ability to appreciate something, self -awareness, the ability to face and solve problems, the ability to face various difficulties, have values and vision as the quality of inspiration life and the ability to avoid injury.

Completely, the results of this study are an instrument that is proposed to be validated in order to measure the sustainability thinking of upper secondary school students. Profiles based on gender, stream, and school location. Gender (male, female), major STEM and non -STEM specializations and the regional digital divide i.e., school location is among the factors related to the study of sustainability thinking. Gender differences are an old problem in the field of thought. In general, male students were found to have greater confidence in using technology -oriented thinking (Young, 2000). Ultimately, through the appreciation of Hurst's theory of sustainability thinking, sustainability thinking in students will motivate them to explore themselves and receive information openly about questions related to sustainability and STEM in their daily lives.

Therefore, the end of this study is to produce a valid instrument and subsequently produce a profile of sustainability thinking skills, STSI. Researchers construct profiles that aim to showcase norm differences in more detail. This profile covers aspects of gender, stream, and school location. In addition, the psychometric characteristics of the constructed items will be tested through factor analysis and Rasch model to improve the level of reliability of the instrument. Figure 2 is the theoretical framework for this study.

Figure 2: Theoretical framework



3. Methodology

3.1. Research Design

The study took a quantitative approach, focusing on a cross-sectional quantitative survey. The quantitative technique was used for this study because it enables the collection and analysis of data in a numerical framework to explain the phenomena being studied (Gay & Mills, 2018). The data was collected via a self-administered internet survey since it is less expensive, requires no copying of surveys, and requires no coding. Additionally, it is simple to administer and capable of collecting detailed and ordered data (Creswell, 2012; Creswell & Creswell, 2018). Thus, the data are nearly immediately ready for statistical analysis (Hair et al., 2017). The data collection method was an online survey in which participants were required to respond to all items before submitting their responses; this eliminated the potential of missing data.

3.2. Study Sample

In this study, the researchers have employed probability sampling. Sampling is intended to be applied to selected individuals because they have experiences at the centre of the phenomenon (Creswell, 2009). Probability sampling techniques employ some type of random selection and allow for the calculation of sampling error, hence reducing selection bias. Thus, 471 secondary school students were surveyed for this study, with 283 females (60%) and 188 males (40%). They were selected using simple probability sampling from six zones: north (37, 8%), east (87, 14%), west (67, 14%), and south (80, 17%), Sabah (90, 19%) and Sarawak (110, 23%. To proceed with the data collection procedure, authorization from the Ministry of Education is required. Henceforth, the researcher must obtain permission from the principal before meeting with the respondents. Respondents were invited to participate in research until the required sample size was attained.

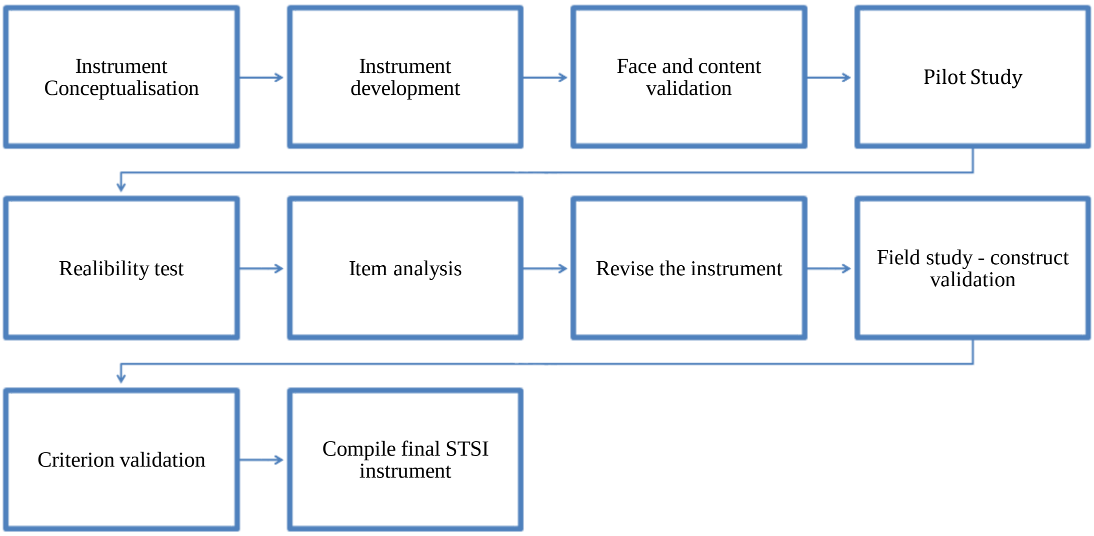
The number of respondents in the field study is sufficient in accordance with Linacre's recommendation (Linacre, 1994), who specified a minimum requirement of 108 respondents for polytomous data with a 99 percent confidence interval and a calibration value of 0.5 logits in order to implement the Rasch measurement model analysis. The Rasch measurement model was used to evaluate the data in this study to determine item fit, polarity, local independence, unidimensionality, item-individual map, reliability, and separation index for both items and respondents. Since this survey participation was done voluntarily, the total return earned was 63.14% (471 from 746 instrument were distribute) acceptable which also met the expectations of previous studies (Carley-Baxter et al., 2009; Mellahi & Harris, 2016). 471 complete instruments were received from respondents and could be processed for further analysis. This value is sufficient to obtain a 99% confidence level for the determination of definite or high importance (Linacre, 1994). In addition, the high reliability values obtained from the analysis indicate that the sample has a large range of people's abilities, from can to not, which is sufficient to perform measurements (Linacre, 2018).

3.3. Instrument Development

To develop the instrument, we adopted Hinkin (1995) scale development guideline, which divided the test development process into 10 steps (Hinkin, 1995). The first step in the instrument which is instrument conceptualization. In this preliminary step, the direction of the study such as the purpose, testing universe, and target audience and test format were defined. On the second step, provides several approaches in constructing the instrument items. The approaches are deductive, inductive, and mixed method approach. The stage provides the foundation for all other development activities. Accordingly, the test plan specifies the characteristics of the test, including an operational definition of the construct and content to be measured, the format of the questions, and the administration and scoring of the test. Then, on the third step is face and content validity which the purpose of face validity is to determine the clarity of the items generated in previous step. Face validity refers to the extent to which the measured variable appears to be a passable measure of the conceptual variable. This step of validity will use experts to assess the items on the instrument whether they had covered the appropriate content. Using panel of experts' review, this will provide constructive feedback about the quality of the newly developed items. It can provide information on the representativeness and clarity for each item.

We then follow up the pilot test with other studies that provide the necessary data for validation and norming. Thus, conducting the pilot test and analyzing its data are an integral part of the test development process. Quantitative item analysis examines how well each test item performs. Subsequently, in revision of the test step, items are dropped based on their consistency, difficulty, discrimination, and bias until a final form of the test is reached. After, the test has been revised, we conduct the reliability test using alpha value that measures internal consistency. Internal consistency describes the extent to which all the items in a test will measure the same concept or construct. Therefore, indirectly, reliability test can be prior prediction of the unidimensionality existence in the instrument (Hinkin, 1995). Items from the pilot study data were analyzed using Rasch model to identify the item fit, unidimensionality, local independence, item polarity, separation index, item person map and differential item functioning. In instrument revising process, any items that did not meet the Rasch model assumption were refined and discussed the decision within focus group discussion. The remaining items that met the assumption were maintained for the next validation process. Again, in step 8 which is field study the item will be assess rigorously using item analysis aforementioned. The next type of validation is criterion validity. The criterion validity occurs when the instrument has an empirical association with some criterion or standard (DeVellis, 2017). In this research, predictive validity testing was conducted to test the empirical associations with some criterion. Based on the newly developed STSI instrument, this study aims to analyse the extent of students' sustainability thinking skills and its effect to country. During the final step, Step 10, the finalized items were arranged as the STSI instrument. Figure 3 depicts the development process.

Figure 3: Instrument development and validation process



Each student self-assessed the instrument in Malay. The instrument is composed of 51 items. Four constructs of STSI are included in the instrument (Sustainability Thinking, Green Technology, Environmental Values and General Ecological Behavior). The instrument is scored on a five-point Likert scale, with 1 indicating "strongly disagree" and 5 indicating "strongly agree." One week is allotted for completing the questionnaires. The raw scores for scales are calculated using the mean score. To begin constructing the scale, a literature review, and interviews with experts (professional and lay) were conducted, and a list of the characteristics that a 'person' must possess was compiled. This five-point rating scale was used because five points Likert items are suitable uses in Rasch Analysis (Fisher, 2006). Therefore, due to this issue, the five-point rating scale was used as the STSI measurement scale. The suitability of this scale was also verified through the Rasch Model analysis.

With 51 items, the scale is intended to assess four primary domains of sustainability. A linguist and two educational professionals reviewed the drafted scale for clarity, language, spelling, and punctuation issues. After making the necessary modifications, an instrument consisting of 51 items was constructed. The following are the findings from the validity and reliability evaluations of the data.

3.4. Rasch Model

The data were analyzed using the Rasch measurement model and suitable to evaluate and assess an instrument's psychometric qualities in terms of validity and reliability. Software WINSTEPS version 3.71.0 (John M. Linacre) Chicago, IL, USA) was used to analyze the following aspects of item functionality: (1) item fit based on infit and outfit values in the range of 0.60 to 1.4 logits (Bond & Fox, 2007); (2) item polarity based on a positive value of point measure correlation (PTMEA CORR); (3) local item dependency analysis; (4) unidimensionality based on the principal component analysis of residuals (PCA); and (5) item polarity based on the positive value of point measure correlation The reliability factor can be calculated using a good internal consistency value (Cronbach's alpha), which is considered acceptable when it surpasses 0.7 (Nunnally & Bernstein, 1994). The item separation index is used to describe a range of item difficulty levels, whereas the individual separation index is used to describe a range of students' ability levels when responding to questionnaires.

4. Result

The psychometric properties of the instrument were determined using the Rasch measurement modal. A total of 51 items were evaluated for item fit, polarity, local independence, unidimensionality, reliability index, and separation index, based on four constructs. Items that fit and contribute to the psychometric features of the instrument were kept, while items that did not fit were submitted for revision or elimination (Linacre, 2010). Additionally, the Rasch measurement model may be used to assess the adequacy of the Likert scale employed in this study using Linacre's six criteria (Linacre, 2002).

Table 1 contains descriptive statistics on the mean and standard deviation of four STSI constructs. The highest degree of STSI was discovered to be environmental value (M=1.31 SD= 0.322), followed by sustainability thinking (M=1.20, SD= 0.270). General ecological behavior (M=1.11, SD=0.231), green technology on the other hand, exhibits the lowest level with (M=1.09, SD=0.223). The highest mean score on the environmental value dimension reflects respondents' excitement, interest, awareness, and empowerment to think and apply sustainability in daily life. Finally, responders' sustainability thinking abilities reveal that they are capable of acquiring sustainability knowledge, value, and perception. The green technology level with the lowest mean score reflects persistence, technology, and self-confidence in practicing STSI in daily life. This score indicates an individual's eagerness to learn more about STSI in depth. Thus, the findings emphasize construct validity, or the degree to which the questions on an instrument correspond to the corresponding theoretical construct (DeVon et al., 2007). Additionally, it contains conclusions regarding the dimensionality of subconstructs and validation of the conceptual framework's structure.

Table 1: Descriptive statistic

	Mean	**Std. Deviation**	**N**
Sustainability Thinking	1.20	0.270	471
Green Technology	1.09	0.223	471
Environmental Value	1.31	0.322	471
General Ecological Behavior	1.11	0.231	471

4.1. Item Fit

Item suitability tests are performed to ensure the items fit the model, which means they provide information to the measurement. Inappropriate items have a tendency for misinformation that will mislead the accuracy of the measurement (Bond & Fox, 2015). Of the 51 items tested, 8 items did not meet the item range set by Fisher (2007) which was in the range of 0.77 to 1.3 logs. This item was removed because the MnSq value does not match the range mentioned above. The remaining 43 items showed that the infit MnSq values ranged from 1.27 to 0.79 log. MnSq clothing values are in the range of 0.78 to 1.28. Standard Error (S.E) values ranged from 0.11 to 0.12 for all 41 items. S.E values are important in indicating accuracy in calculations (Linacre, 2005). This range below the value of 0.25 is considered very good (Fisher, 2007). This indicates that the item has a very small error value that does not interfere with the result. Based on the fit quality criteria mentioned, two items were removed from each construct, namely Sustainability Thinking, Green Technology in Sustaining Nature, Environmental Values, and General Ecological Behavior. Those items are item 1 and item 7 of the sustainability thinking construct. Items 24 and 25 were then removed from the green technology construct in conserving nature. For the Environmental values construct, items 37 and 40 were removed. The last two items, namely items 42 and 45 were removed from the General Ecological Behavior construct.

4.2. Polarity

Additionally, item fit can be determined based on the polarity of the item by calculating the PTMEA CORR value. This value refers to a collection of items that all measure the same construct, assuming that the items all measure the same construct (Bond & Fox, 2007). The PTMEA CORR value achieved in this study was between 0.43 and 0.68, which was within the minimal value of 0.3 (Wu & Adam, 2007). Item polarity is an indicator used to indicate that a measured item will move in one direction within a constructed dimension. Based on Table 2, the polarity values of the items shown are between 0.53 and 0.69. The positive polarity value of the measurement result indicates that the instrument measures one of the same latent variables (Linacre, 2018). Values above +.4 also indicate the polarity consistency of the indicator in the scale (Bond & Fox, 2015). No items were dropped based on the polarity item value decision. The positive PTMEA CORR score indicated that the retained items could contribute to the instrument's psychometric features, allowing it to distinguish sustainability thinking of students. In addition, this indicates that all the items used are parallel to the measurement of STSI. This analysis is the fundamental procedure to gauge the validity of construct used to build and validate the instrument. Table 2 depicts the MNSQ values and polarity values of the items.

Table 2: Fit statistics of measurement items

Item	**Measure**	**Standard Error**	**MNSQ**		**PTMEA**	
			Infit	***Outfit***	**Corr.**	**Exp.**
PK1	0.19	0.11	0.65	0.63	0.6	0.61
PK2	-0.86	0.12	1.17	1.12	0.51	0.57
PK3	-0.37	0.11	1.07	1.05	0.6	0.59
PK4	0.45	0.11	0.78	0.79	0.62	0.61
PK5	0.53	0.11	0.89	0.89	0.61	0.61
PK6	-0.19	0.11	0.94	0.93	0.64	0.6
PK7	-0.56	0.12	0.67	0.61	0.57	0.58
PK8	0.85	0.11	1.14	1.14	0.55	0.62

PK9	0.17	0.11	1.05	1.22	0.58	0.61
PK10	0.54	0.11	1.06	1.05	0.58	0.62
PK11	1.17	0.11	0.75	0.75	0.67	0.63
PK12	1.08	0.11	1.04	1.09	0.58	0.63
PK13	1.78	0.11	0.86	0.85	0.67	0.64
PK14	1.03	0.11	0.9	0.9	0.63	0.63
PK15	0.27	0.11	0.95	0.95	0.63	0.61
PK16	1.02	0.11	1.09	1.1	0.64	0.63
TH1	-0.58	0.12	1.18	1.15	0.61	0.58
TH2	-0.79	0.12	0.83	0.82	0.63	0.57
TH3	-0.7	0.12	0.95	0.89	0.64	0.58
TH4	-0.56	0.12	1.04	1.02	0.58	0.58
TH5	-0.35	0.11	0.88	0.89	0.62	0.59
TH6	-1.02	0.12	0.91	0.89	0.62	0.57
TH7	-1.09	0.12	0.82	0.78	0.64	0.56
TH8	0.03	0.11	0.64	0.64	0.68	0.6
TH9	-0.59	0.12	1.45	1.43	0.44	0.58
TH10	-0.11	0.11	0.78	0.8	0.63	0.6
TH11	-0.48	0.11	0.87	0.88	0.64	0.59
TH12	0.37	0.11	1.25	1.28	0.56	0.61
TH13	-0.2	0.11	1.01	1.01	0.61	0.59
EV1	0.05	0.11	0.81	0.81	0.65	0.6
EV2	-0.86	0.12	0.83	0.82	0.63	0.57
EV3	-1.13	0.12	0.86	0.81	0.62	0.56
EV4	-1.31	0.12	0.95	0.92	0.59	0.55
EV5	-0.6	0.12	0.97	0.95	0.59	0.58
EV6	-0.97	0.12	0.94	0.91	0.62	0.57
EV7	-0.54	0.11	0.87	0.86	0.62	0.58
EV8	2.1	0.11	1.45	1.48	0.46	0.64
EV9	-0.54	0.11	0.79	0.8	0.62	0.58
EV10	-1.2	0.12	0.76	0.79	0.63	0.56
EV11	-1.17	0.12	0.6	0.68	0.63	0.56
EV12	-1.65	0.13	0.74	0.71	0.64	0.54
GEB1	2.02	0.11	1.9	1.93	0.45	0.64
GEB2	0.82	0.11	1.13	1.13	0.6	0.62
GEB3	1.03	0.11	1	1.01	0.6	0.63
GEB4	1.29	0.11	1.42	1.41	0.56	0.63
GEB5	0.69	0.11	0.91	0.9	0.65	0.62
GEB6	-1.12	0.12	1.08	1.04	0.58	0.56
GEB7	0.19	0.11	1.11	1.1	0.6	0.61
GEB8	-0.11	0.11	0.88	0.87	0.65	0.6
GEB9	0.34	0.11	0.9	0.9	0.65	0.61
GEB10	1.65	0.11	1.55	1.56	0.51	0.64

4.3. Local Independence

The following feature of item measurement is local independence analysis. Local independence tests were performed to ensure that the items were not related to each other. The results in Table 3 show 3 pairs of items with standard residual correlation values ranging from .38 to .56. This range meets the recommended correlation value that must be less than 0.70 (Linacre, 2018). This indicates that all items are not related to each other to measure the dimensions together. Local independence is often quantified using the standardized residual correlation value between two items, which should not exceed 0.3 (Balsamo et al., 2014). Likewise, if the correlation between two items is greater than 0.7, only one item is kept and the other is excluded from the model (Linacre, 2005). The retained item will be determined using the MNSQ value, which should be close to or equal to 1.0 (Linacre, 2005; Bond & Fox, 2015), as this value represents the predicted value for model fit (Aziz et al., 2015). This procedure is taken to ensure that retained items do not duplicate existing ones (Matore et al., 2020). The results showed that the highest correlation value between items was only 0.57, far from the minimum correlation value set by Linacre (2018) of 0.7. This means that all items are locally independent in measuring the shared dimension, sustainability thinking skills of students.

Table 3: Standard correlation of residual values

Correlation	Entry Number	Item	Entry Number	Item
.56	39	EV10	40	**EV11**
.43	14	PK14	15	**PK15**
.38	**46**	**GEB5**	**47**	**GEB6**

4.4. Gender Differential Item Functioning (GDIF)

The purpose of this analysis was to determine the presence of Gender Differential Item Functioning (GDIF) in the instrument used. Winstep uses a two-tailed t-test to determine the significance of the difference between two index difficulties when analyzing GDIF. For all DIF analyses, the confidence level is 95 percent, and the level of t critical value is 2.0. Additionally, the GDIF Contrast index is utilized to demonstrate the difference in gap confirmation levels between males and females when males and females are compared. According to Lai and Eton (2002), the Likert scale requires a value of 0.5 logits DIF contrast. Meanwhile, in Pallant and Tenant (2007), Wright and Panchapakesan suggest that GDIFs with a size less than 0.5 logits are regarded inconsequential. A low GDIF Contrast index indicates that the item is more easily affirmed by female respondents. DIF Measurement is the difficulty index of this group while all other variables are kept constant. The DIF contrast results indicate that 10 out of 51 items illustrate the relevance of GDIF in terms of t ≥±2 logit value. However, the GDIF contrast (±0.5 logits) indicates that 10 items do not exhibit significant GDIF, as indicated by the GDIF index being less than 0.5 logit. The GDIF Contrast value ranges between -0.43 and 0.49. As such, it is identified that 51 items remain. Items that passed this GDIF analysis demonstrated that they satisfy the disposition testing element of fairness. The study uses DIF to identify all 51 items that did not exhibit evidence of injustice when a group of students with varying skill levels of the same sex was compared.

4.5. Differential Item Functioning (DIF) based on stream (STEM and Non-STEM)

The purpose of differential item functioning is to identify the biased items between STEM students and non-STEM students. From the analysis, there were only two items that showed potential bias. It is based on the difference in DIF Contrast value in the difficulty of the item between the two groups that is greater than 0.64 and p (| DIF | = <.05 (Linacre, 2018). The items were item 17 (TH 1) and item 39 (EV 10). Item 17 (PK 6) (*I provided questions based on the level of green technology*) showed DIF contrast value of -0.66 and value of p = .0037. This value indicates that STEM students found this question easier to endorse (measure = -0.89 logits) compared to non-STEM students (measure = -0.23 logits). However, another item indicates the opposite situation. As for item 39 (EV 10)(*I provided questions based on the level of environmental values*) showed DIF contrast value of 0.67 and value of p= 0.002. STEM students (measure = -0.05) found it is easier to endorse this item than STEM students (measure = 0.61).

4.6. Differential Item Functioning (DIF) based on school location

The purpose of differential item functioning is to identify the biased items between students in rural and urban areas. From the analysis, there were only two items that showed potential bias. It is based on the difference in DIF Contrast value in the difficulty of the item between the two groups that is greater than 0.64 and p (| DIF | = <.05 (Linacre, 2018). The items were item 6 (PK 6) and item 34 (EV 5). Item 6 (PK 6) (*I provided questions based on the level of thinking skills*) showed DIF contrast value of -0.68 and value of p = .0038. This value indicates that students in the urban schools found this question easier to endorse (measure = -0.87 logits) compared to students in the rural schools (measure = -0.21 logits). However, another item indicates the opposite situation. As for item 34 (EV 5) (*I provided questions based on the level of environmental values*) showed DIF contrast value of 0.69 and value of p= 0.003. Students in the rural schools (measure = -0.05) found it is easier to endorse this item than students in the urban schools (measure = 0.63).

4.7. Unidimensionality

Unidimensionality tests were performed to ensure the instrument only measured one dimension, which in this study, was students 'sustainability thinking skills. In the Rasch model, in addition to proving that the instrument measures one dimension and emphasizes that there is no second dimension. There are four assumptions in the unidimensionality test namely the crude variance explained by the measure value, the unexplained variance in the contrast value 1, the variance ratio and the Eigen value for the first unexplained contrast. First, the findings in the Standardized Analysis of Residual Variance in Table 4 show that 66.1% variance is explained by measurement. This figure meets the minimum requirement of good by Rasch Scholar Fisher (2007), this value meets the value agreed by Azrilah et al. (2017) and Conrad et al. (2012) which is widely accepted by instrument developers in Malaysia. The 66.1% variation described by the measure meets the minimum 40% of Rasch dimensional measurements (Azrilah et al., 2017). Values above 40% are also considered a strong measurement dimension by Conrad et al. (2012).

Table 4: Standardized residual variance (in eigenvalue units)

	Eigen Value	**Empirical**		**Modeled**
Total raw variance in observations	76.1	100.0%		100.0%
Raw variance explained by measures	45.1	66.1%		66.3%
Raw variance explained by persons	17.9	23.5%		23.6%
Raw Variance explained by items	17.2	22.6%		22.7%
Raw unexplained variance (total)	41.0	53.9%	100.0%	
Unexplained variance in 1st contrast	3.7	3.7%	8.0%	
Unexplained variance in 2nd contrast	3.3	4.3%	8.0%	
Unexplained variance in 3rd contrast	2.8	3.7%	6.9%	
Unexplained variance in 4th contrast	2.3	3.1%	5.7%	
Unexplained variance in 5th contrast	1.7	2.3%	4.2%	

Second, the degree of interference was measured by identifying unexplained variances on the first contrast. The value recorded was 3.9% and was in the excellent and adequate category (Eakman, 2012; Fisher, 2007). Next, the ratio of variance between the variance explained by the measure (66.1%) and the unexplained variance from the first contrast (3.9%) was 9.4: 1 which exceeded the recommended minimum value of 3: 1 (Conrad et al., 2012; Linacre, 2018). The unexplained eigenvalue of variance in the first contrast was 3.7 also confirming that no other dimensions exist in this instrument (Linacre, 2005a). Dimensionality tests were determined using four methods to ensure that the instrument measured only one dimension, namely students 'sustainability thinking skills. First, values are described by measures that exceed the minimum level of 40%. Second, the level of interference in the first contrast range is very good, which means that the instrument does not show the presence of a second dimension. Also, the variance ratio (9.4: 1) exceeding the minimum value of the 3: 1 variance ratio indicates that the dimension measured is the dimension that dominates the measurement. The eigenvalue for the first unexplained variance of 3.7 also confirms unidimensionality which means that this instrument robustly measures students 'sustainability thinking skills only. Overall, the items in STSI meet the non-dimensional assumptions in the Rasch Model.

4.8. Reliability Index

The interpretation of person reliability is equivalent to Alpha Cronbach or KR20 (Wright & Master, 1982). Cronbach's alpha is 0.98, and item reliability is 0.99, both of which are considered excellent values (Nunnally & Bernstein, 1994). In this study, the reliability index for respondents is 0.98, which is an acceptable range (Pallant & Tennant, 2007) which could expect consistency level personal situation arrangement in the log scale if this sample answers different set item, but to measure that the same construct (Wright & Masters, 1982).

4.9. Separation Index

The instrument items' reliability was determined using the index of person separation, which is comparable to Cronbach's alpha. The term "person separation" refers to the process of classifying persons and estimating how well a measure can separate individuals on a construct. The presence of a high degree of person separation or stratification (two distinct levels of performance, i.e., high, and low, that can be separated based on test scores, person reliability of 0.7) indicates that the measure may be sensitive to distinguishing between high and low performers. Separation of items is used to validate the item hierarchy.

The presence of a high degree of item separation or stratification (three items representing three distinct levels of difficulty, namely high, medium, and low; item reliability of 0.9) indicates that the person sample is sufficiently large to corroborate the item difficulty hierarchy (Linacre, 2017). Individual separation index is recorded at 5.72, as per in Table 5, which means there were 5 ability level of respondent's ability level and is regarded adequate when it exceeds the value of 2 (Linacre, 2012; Fox & Jones, 1998). Meanwhile, Table 6 represents the item separation index was 7.53 which is considered acceptable. This means the scale can be statistically differentiated into 7 difficulty levels. Increased item separation index values imply a more effective separation of items of varied difficulty. Separation is dependent on item reliability (Wright & Masters, 1982). This outcome confirms Linacre's (2005) assertion that separation indexes of two and above indicate greater reliability. As a result, the instrument has a widespread when it comes to determining the level of STSI. Additionally, it demonstrates that the tool is measuring what it is designed to measure, thereby establishing its validity.

Malaysian Journal of Social Sciences and Humanities (MJSSH) (e-ISSN : 2504-8562)

Table 5: Statistical summary for person

	Raw Score	Count	Measure	Model Error	*Infit*		*Outfit*	
					MNSQ	ZSTD	MNSQ	ZSTD
Mean	155.9	51.0	2.41	0.26	1.01	-0.20	0.99	-0.2
Standard Deviation	20.9	0.0	1.80	0.02	0.43	2.1	0.42	2.1
Max	194.0	51.0	8.17	0.37	4.07	8.9	4.02	8.8
Min	80.0	51.0	-2.07	0.24	0.29	-53	0.28	-5.2
Real RMSE	0.28	True S. D	1.32	Separation	5.72	Person Reliability		0.98
Model RMSE	0.26	True S. D	1.33	Separation	5.09	Person Reliability		0.98

Person Raw Score-To-Measure Correlation = .1.00
Cronbach's Alpha (KR-20) Person Raw Score Reliability = .98

Table 6: Statistical summary for item

	Raw Score	Count	Measure	Model Error	*Infit*		*Outfit*	
					MNSQ	ZSTD	MNSQ	ZSTD
Mean	1448.8	471.0	0.12	0.11	0.95	-0.6	0.95	0.5
Standard Deviation	73.8	0.0	0.76	0.00	0.12	1.13	0.13	1.4
Max	1740.0	471.0	1.78	0.12	1.25	2.8	1.28	3.0
Min	1390.0	471.0	-1.31	0.011	0.78	-2.8	0.78	-2.6
Real RMSE	0.12	True S. D	0.75	Separation	7.53	Item Reliability		0.99
Model RMSE	0.11	True S. D	0.75	Separation	6.74	Item Reliability		0.99

UMean = 0.000 UScale = 1.000
Item Raw Score-To-Measure Correlation = 0.99

4.10. Item-individual Map

The item-individual map depicts the distribution mapping of the items and respondents in this study on a similar logits scale following the calibration process. To explore students 'levels of sustainability thinking skills, these levels of sustainability thinking skills were assessed visually through item-people maps. The item-person map is a visual map to identify the position and items of the respondents on the same scale. Item PK7 (I think economic investment should require assurance of life and property environment) was the most difficult item for respondents to support while the easiest item to support was item EV2 (Industrial smoke emissions from chimneys make me angry).

The maximum respondent ability value was 4.70 logs, and the minimum value was -2.70 logs, while the maximum item difficulty value was 1.77 logs, and the minimum value was -1.41 logs. The ability range of the respondents was 7.78 logs while the item difficulty range was 3.07 logs. The difficulty of the mean value item is lower than the mean value of the individual ability. This shows that students' overall sustainability thinking skills exceed the difficulty level of IKPK-STEM items. The value of the item spread measure located between 1.77 logs to -1.41 logs meets a satisfactory range from +3.00 logs to -3.00 logs to indicate its stability (Embretson & Reise, 2000; Linacre, 1994). The number of items was insufficient at the highest and lowest levels of difficulty to ensure a high level of measurement accuracy for the highest and least prepared respondents. The addition of items between 1.80 logs to 4.79 and -1.33 to -2.87 logs could help the instrument to differentiate respondents with the highest and least levels of students 'sustainability thinking skills.

Several items in the same construct are on the same level. For example, TH9 and TH10 items are on the same level for green technology constructs in conserving nature. Even so, because of the importance of acquiring different interests and motivations, researchers have decided to defend it. It can also be seen that most of the items for the Sustainability Thinking construct are above the mean value of those items. This suggests that almost all items of Sustainability Thinking are difficult for students to agree on. All items in the Environmental values construct were below the mean value of the item, indicating that students were more prepared in this construct.

The deduction value for students 'sustainability thinking skills was calculated based on the position above the mean value of the item. Mean values were used as reference values to differentiate respondents 'abilities based on previous researcher practices (Moeini et al., 2016; Nazlinda et al., 2017). In this study, the mean of individual abilities (μ = 1.67) was higher than the mean of item difficulty (μ = - 0.11) indicating that many respondents agreed easily with the item. This explains that STSI items are easily supported and the probability of students 'sustainability thinking skills is high.

As a visual map, this map successfully shows the level of students' sustainability thinking skills, distribution of respondents and items as well as identifying item gaps that need to be added to improve the quality of STSI. It also provides information on difficult items for students to be additional information to stakeholders in providing appropriate training later. In conclusion, it can be concluded that most students are willing to implement sustainability.

4.11. Scale Review

The Rasch measurement model can be used to determine the efficacy of a scale used in an instrument based on six specified criteria (Linacre, 2002). The first condition is that each concept contain at least ten observations, which was met in this study. A scale review was performed to test the effectiveness of the measurement scale in Rasch measurements. Discussions on scale weighing are made based on criteria outlined by Linacre (2002) and Bond and Fox (2015). This consideration was made to ensure that the five-likert scale used was appropriate for STSI. For this study, it can be referred to the number observed which shows the response value of more than ten. With regards, second, cross -category observations for all indicators show a unimodal distribution with peaks in the middle category as illustrated in Figure 4.

There are six assumptions to consider. First, the category frequency must have more than 10 responses in each category. Third, the mean measurement function of each category must increase in line with the increase in the category scale. It can be seen in the average which shows an increase from -1.43 <0.15 <1.65 <3.23, across the 1,2,3,4,5 rating scale response categories. Fourth, the MnSq value of the garment must be less than two to indicate no noise in the measurement. The results showed that the MnSq value of the outfit between 1.25 to 0.95 indicates that no excessive randomization is a threat to the measurement system. Next is the increase in the threshold value. Based on the structural calibration the threshold value increased from -2.87 < - 0.14 <2.98 (Table 7). Lastly, a test for advanced step difficulty. The value required to ensure a category does not need to be removed or added is between 1.5 and 5 logs. It was found that the difficulty of the steps between the categories were all within a range that reinforced no need to discard, or additional scales had to be made. A check of each restriction as per Table 8, revealed that the difference between each scale category exceeded one and fell within the range of five, as follow.

Figure 4: Threshold for scale review

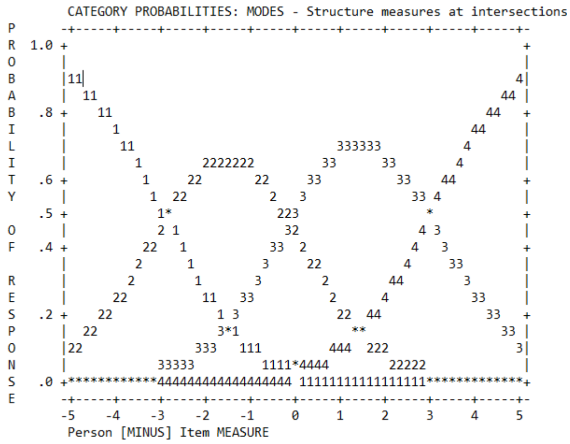


Table 7: Mean category of measurement: Observed average

Category		Observed		Observed average	Sample expected	MNSQ		Structure calibration	Category measurement
label	score	count	%			*infit*	*outfit*		
1	1	288	2	-1.42	-1.72	1.22	1.24	NONE	(-4.01)
2	2	2432	18	0.14	0.16	0.99	1	-2.87	-1.51
3	3	701	51	1.64	1.67	0.99	0.95	-0.14	1.44
4	4	4083	30	3.22	3.18	0.97	0.97	2.98	4.13
5	5	2037	15	1.78	1.97	0.99	0.99	1.98	2.13

Table 8: Revision scale check

Scale	Gaps Calculation	Range of acceptance	Decision
$S_1 - S_2$	0.00 – (-2.87)	1.00 < 2.87 < 5.00	Accepted
$S_2 - S_3$	-0.14 – (-2.87)	1.00 < 2.73< 5.00	Accepted
$S_3 - S_4$	2.98– (-0.14)	1.00 < 3.12 < 5.00	Accepted
$S_4 - S_5$	2.98 – 1.98	1.00 < 1.00< 5.00	Accepted

A scale review was performed to test the effectiveness of the measurement scale on STSI. There were six hypothesis tests conducted and they gave good results. First, each category had more than 10 observations. This indicates that no data is expected to be problematic. Second, there is a clear upper probability in each category that shows a uniform probability distribution.

Third, the mean measurement values that increased along the scale of each category indicated the correct response from the sample. Fourth, clothing values of less than two indicate no noise level that could cause misinformation in the observations. Next, increase the threshold value with the respective scale evaluation categories. It shows evidence of a satisfactory category scale. Finally, threshold category values ranging from 1.4 to 5 indicate that the scale used is accurate. Based on the compliance of the six assumptions for the scale review, the five-likert scale used in this instrument is appropriate for measuring students 'sustainability thinking skills.

5. Discussion

The study has exemplified that this newly developed instrument has good features in terms of achieving objectives, validity, reliability, and usability. There are two phases involved, namely the development phase and the validation phase comprising 10 steps that have been carried out in this study. We detail the development of a scale to assess secondary school students' disposition toward STSI. Item Response Theory (IRT) method was selected to apply the Rasch measurement model to analyze each test item to determine the validity and reliability of the instrument. To begin, the instrument had good psychometric qualities. The study established strong reliabilities for the construct. During the pilot test, the STSI constructs were empirically validated in Malaysia using EFA and Rasch analysis.

Our findings in this investigation substantiated unidimensionality at the scale level. The summation of raw item scores into an interpretable total scale score is acceptable since each component's items all measure the same latent characteristic. Reliability tests were performed to test the replication and repeatability of the instrument. The advantage of using a Rasch model in testing reliability values is its ability to provide three types of reliability readings. All values of reliability, person reliability, item reliability and alpha cronchbach were found to be in excellent range. Other than it is a sign for adequate number of respondent and items, this instrument has been verified to be replicable for other group of secondary school students in Malaysia (Fisher, 2007).

STSI was developed for use in the context of sustainability education, there is a strong need to clearly differentiate participants at the most self -focused level or other levels. In general, estimates of individual ability and item difficulty are fairly evenly distributed around the logit continuum. However, psychometrically, items on this scale were not sufficient to capture these participants at their high level of ability.

We also explored gender differences in the item differentiation (DIF) function. Briefly DIF occurs when individuals with the same level of ability respond differently to items simply because they belong to different groups. Completely, a DIF item is a question that has been distorted by a specific group of people. All items in this study were DIF -independent, allowing meaningful comparisons between groups. These findings establish the basis for further testing of DIF with more samples. So, researchers should proceed with caution when conducting comparisons around the world using this instrument.

Cronbach's alpha values and People Separation Index (PSI) were within the margin of reliability., Cronbach's alpha was calculated using raw scores while PSI was obtained using logit -modified individual estimates. PSI was equivalent to Cronbach's alpha when the distribution was normal. PSI and alpha values greater than 0.7 are usually considered adequate (Fisher, 1992; Tavakol & Dennick, 2011). The findings have answered all the possibilities designed to study the suitability of the item. The reliability of the item is high, and this means the item is stable. A good instrument is one that could discriminate person and item. STSI is proven to differentiate that. With the ability to separate the item and person to more than five groups indicate that this instrument is excellent (Fisher, 2007). It also a sign that the number of item and the person is sufficient for the analysis (Boone et al., 2014).

6. Limitations and Future Directions

Current research has some limitations, which also provide direction for further research. The main limitation of this study is limited to secondary school students in one country, Malaysia. However, conclusions can be drawn from about STSI in various domains. Completely, the instrument must be used for additional domains. In addition, caution is advised when using this instrument in other situations, and more testing with samples from other cultural groups should be performed. Another important point when extending this instrument to a broader context is the need to investigate the function of differentiation items to make relevant comparisons. In addition, replication in multiple countries will increase the relevance of the study across multiple countries. Lastly, other types of validity, such as convergent validity and discrimination, may also be investigated in future studies, however they are still beyond the scope of this research. Comparing research across different tests can also provide a more holistic range of psychometric assessments of findings from a variety of angles. This analysis will not only influence the subsequent analysis, but it can also improve the psychometric quality of the item. Most critically, the researcher must match the appropriate tendencies of the students in the Malaysian environment. However, this questionnaire does not yet cover all the features listed in the literature, and it is possible that some relevant variables were not included. Future studies could build on this work by examining additional elements of STSI.

7. Conclusion

In summary, the data from each STSI item meet the assumptions of the Rasch model. Each of the 51 items was retained. Each item showed good performance for item suitability, polarity, and local freedom. This work becomes an innovation in the research body on STSI teaching and learning by providing a more comprehensive overview of sustainability tendencies and attitudes, as well as their impact on their willingness to adapt to sustainability in todays and future lives. This is important to adhere to the various concepts of sustainability education across the K-12 curriculum. Rasch’s analysis confirms the applicability of the STSI as an instrument to assess students ’sustainability attitudes about sustainability in daily life, particularly in the context of education and sustainability of future generations.

Acknowledgement

The authors like to express their gratitude to the Ministry of Education, specialists, teachers, and students who assisted in the data gathering and instrument validation processes. Additionally, the authors like to thank and express their gratitude to the Faculty of Education at Universiti Kebangsaan Malaysia (UKM).

Funding

This study received no funding.

Conflict of Interests

The authors declare no conflict of interest in this study.

References

Axon, S., & James, D. (2018). The UN Sustainable Development Goals: How can sustainable chemistry contribute? A view from the chemical industry. *Current Opinion in Green and Sustainable Chemistry*, *13*, 140–145.

Anastas, P. T., & Zimmerman, J. B. (2018). The United Nations sustainability goals: How can sustainable chemistry contribute? *Current Opinion in Green and Sustainable Chemistry*, *13*, 150–153.

Aziz, A.A., Masodi, M.S., & Zaharim, A. (2015). *Rasch measurement model basis: Scale development and measurement structure* (3rd ed.). UKM Press.

Azrilah, A. A., Mohd Saidfudin, M., & Azami, Z. (2015). Asas model pengukuran Rasch: Pembentukan skala & struktur pengukuran. In *Penerbit UKM*.

Avvisati, F., Echazarra, A., Givord, P., & Schwabe, M. (2018). *Programme for International Student Assessment (PISA) - Results from PISA 2018* (pp. 1–10). https://www.oecd.org/pisa/publications/PISA2018_CN_BRN.pdf

Balsamo, M., Giampaglia, G., & Saggino, A. (2014). Building a new Rasch-based self-report inventory of depression. *Neuropsychiatric Disease and Treatment*, *10*,153-165. https://doi.org/10.2147/ndt.s53425

Baumgarten, E. (2001). Curiosity as a moral virtue. *International Journal of Applied Philosophy*, *15*(2), 169–184.

Berlyne, D. E. (1960). Conflict, arousal and curiosity. *British Journal of Psychology*, *45*(3), 180– 191. https://doi.org/10.1037/e417772008-001

Bond, T G, & Fox, C. M. (2007). Applying the Rasch model. Fundamental measurement in theHuman science. New Jersey: Lawrence Earlbaum Associates. *Inc. Publicers*.

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model: fundamental measurement in the human sciences* (3rd ed., Vol. 1). New York: Routledge. https://doi.org/10.1207/S15327574IJT013&4_10

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

Bybee, R. W. (2010). Advancing STEM Education: A 2020 vision. *Technology and Engineering Teacher*, *70*(1), 30.

Carley-Baxter, L. R., Hill, C. A., Roe, D. J., Twiddy, S. E., Baxter, R. K., & Ruppenkamp, J. (2009). Does response rate matter? journal editors use of survey quality measures in manuscript publication decisions. *Survey Practice*, *2*(7), 1–7. https://doi.org/10.1021/jp012662w

Committee on STEM Education. (2018). Charting a Course for Success: America's Strategy for STEM Education. In *National Science and Technology Council*.

Conrad, K. M., Conrad, K. J., Dennis, M. L., & Funk, R. (2012). *Validation of the Self-Help Improvement Scale to the Rasch measurement model GAIN methods report 1.0*. Chicago.

Creswell, J. W. (2009). *Research design qualitative, quantitative, and mixed methods approach* pp. 3rd Edition. United States of America: Sage Publications, Inc.

Creswell, J. W. (2012). Educational research: Planning, conducting, and evaluating quantitative and qualitative research (4th ed.). Pearson Education Inc.

Creswell, J. W., & Creswell, J. D. (2018). Research design: Qualitative, quantitative, and mixed methods approaches (5th ed.). Sage Publications, Inc.

Curriculum Development Division. (2015). *Dokumen standard kurikulum dan pentaksiran Sains Tingkatan 1*. Putrajaya: Kementerian Pendidikan Malaysia.

Curriculum Development Division. (2016a). *Dokumen standard kurikulum dan pentaksiran Sains Tingkatan 4*. Putrajaya: Kementerian Pendidikan Malaysia.

Curriculum Development Division. (2016b). *Panduan pelaksanaan Sains, Teknologi, Kejuruteraan dan Matematik (STEM) dalam pengajaran dan pembelajaran*. Putrajaya: Ministry of Education.

Day, H. Y. (1968). Role of specific curiosity in school achievement. *Journal of Educational Psychology*, *59*(1), 37.

DeVellis, R. (2017). *Scale development theory and applications (*4th ed.). SAGE Publications Inc.

DeVon, H. A., Block, M. E., Moyle-Wright, P., Ernst, D. M., Hayden, S. J., Lazzara, D. J., Savoy, S. M., & Kostas-Polston, E. (2007). A psychometric toolbox for testing validity and reliability. *Journal of Nursing Scholarship*, *39*(2), 155–164. https://doi.org/10.1111/j.1547-5069.2007.00161.x

Eakman, A. M. (2012). Measurement characteristics of the engagement in meaningful cctivities survey in an age-diverse sample. *American Journal of Occupational Therapy*, *66*(2), 20–29.

Embretson, S. E., & Reise, S. P. (2000). Item response theory for psychologists. Lawrence Erlbaum Associates.

Fisher, J. W. P. (2006). Survey design recommendations. *Rasch Measurements Transactions Contents and Archives, 20*(3), 1072–1074.

Fisher, J. W. P. (2007). Rating scale instrument quality criteria. *Rasch Measurement Transactions, 21*(1), 1095.

Fisher, W. (1992). Reliability, separation, strata statistics. *Rasch Measurement Transactions, 6,* 238.

Foster, K. M., Bergin, K. B., McKenna, A. F., Millard, D. L., Perez, L. C., Prival, J. T., Rainey, D. Y., Sevian, H. M., VanderPutten, E. a, & Hamos, J. E. (2010). Partnerships forSTEM education. *Science*, *329*(5994), 906–907.

Fox, C. M., & Jones, J. A. (1998). Uses of Rasch modeling in counseling psychology research. *Journal of Counseling Psychology*, *45*(1), 30–45. https://doi.org/10.1037/0022-0167.45.1.30

Gagné, E. D. (1985). The cognitive psychology of school learning. Boston, MA: Little, Brown and Company.

Garrosa, E., Blanco-Donoso, L. M., Carmona-Cobo, I., & Moreno-Jiménez, B. (2017). How doescuriosity, meaning in life, and search for meaning predict college students' daily emotional exhaustion and engagement? *Journal of Happiness Studies*, *18*(1), 17–40.

Gay, L. R., & Mills, G. E. (2018). Educational research: Competencies for analysis and applications (12th ed.). Merrill Prentice Hall.

Grover, S. & Pea, R. (2013a). Computational thinking in K-12: A review of the state of the field. *Educational Researcher*, *42*(1), 38–43.

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2013). *Multivariate Data Analysis. Vectors* (7th ed.). Pearson Education Ltd. https://doi.org/10.1016/j.ijpharm.2011.02.019

Hamid, S., Ijab, M. T., Sulaiman, H., Md. Anwar, R., & Norman, A. A. (2017). Social media for environmental sustainability awareness in higher education. In *International Journal of Sustainability in Higher Education* (Vol. 18, Issue 4, pp. 474–491).

Hinkin, T. R. (1995). A review of scale development practices in the study of organizations. *Journal of Management*, *21*(5), 967–988.

Holme T.A, Hutchison J.E (2018). A central learning outcome for the central science. J Chem Educ, 95:499–501.

Hurst G.A (2013). Green and smart: hydrogels to facilitate independent practical learning. J Chem Educ 94:1766–1771.

Hurst G.A (2015). The rheological properties of poly (vinyl alcohol) gels from rotational viscometry. J Chem Educ, 92:940–945.

Hurst G.A, Ludwig J.K (2017). The sustainability thinking in rheological properties gels from rotational viscometry. J Chem Educ, 92:940–945.

Hurst G. A (2019). The green formula for international chemistry education. In Dicks AP, Bastin L, Eds., Integrating green and sustainable chemistry principles into education. Elsevier: 205–228.

Jin, G., & Bierma, T. (2013). STEM for non-STEM Majors: Enhancing Science Literacy in Large Classes. *Journal of College Science Teaching*. https://doi.org/10.2505/4/jcst13_042_06_20

Jonassen, D. (2000). Toward a design theory of problem solving. *Educational Technology Research and Development*, *48*(4), 63–85.

Kashdan, T. B., Rose, P., & Fincham, F. D. (2004). Curiosity and Exploration: Facilitating Positive Subjective Experiences and Personal Growth Opportunities. *Journal of Personality Assessment*, *82*(3), 291–305.

Kashdan, T. B., Gallagher, M. W., Silvia, P. J., Winterstein, B. P., Breen, W. E., Terhar, D.,& Steger, M. F. (2009). The curiosity and exploration inventory-II: Development, factorstructure, and psychometrics. *Journal of Research in Personality*, *43*(6), 987–998. https://doi.org/10.1016/j.jrp.2009.04.011

Kennedy, T. J., & Odell, M. R. L. (2014). Engaging Students in STEM Education. *Science Education International*, *25*(3), 246–258.

Lai, J. S., & Eton, D. T. (2002). Clinically meaningful gaps. *Rasch Measurement Transactions, 15*(4), 850.

Linacre, J. M. (1994). Sample size and item calibration (or person measure) stability. *Rasch Measurement Transactions, 7*(4), 328.

Linacre, J. M. (2002). What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions, 16*(2), 878.

Linacre, J. M. (2005). Winstep Rasch measurement computer program. MESA Press.

Linacre, J. M. (2005a). *A user's guide to WINSTEPS: Rasch Model Computer Programs*. Chicago: MESA Press.

Linacre, J. M. (2010). When to stop removing items and persons in Rasch misfit analysis? *Rasch Measurement Transactions, 23*(4), 1241.

Linacre, J.M. (2012). A user's guide to WINSTEPS: Rasch model computer programs. MESA Press.

Linacre, J. M. (2017). Winsteps rasch measurement computer program user's guide. Winsteps.com.

Linacre, J. M. (2018). *A user's guide to WINSTEPS: Rasch Model Computer Programs*.

Loewenstein, G. (1994). The psychology of curiosity: A review and reinterpretation. *Psychological Bulletin*, *116*(1), 75–98. https://doi.org/https://doi.org/10.1037/0033-2909.116.1.75

Ludwig, J. K. (2017). Green Chemistry: An Introductory Text. *Green Chemistry Letters and Reviews*, *10*(1), 30–31.

Ludwig, J.K, Hurst G.A, Orgill M.K (2019). Introduction to systems thinking for the chemistry education community. J Chem Educ, 96:2720–2729.

Mahaffy P.G (2019). Systems thinking for education about the molecular basis of sustain-ability. Nat Sustain 2019, 2:362–370

Matore, M. E. E. M, Zainal, M. A., Noh, M. F. M., Khairani, A. Z., Rahman, N. A., & Idris, H. (2020). Pengujian Psikometrik Item Kecerdasan Menghadapi Cabaran untuk Pelajar Lelaki Kejuruteraan Mekanikal Menggunakan Model Pengukuran Rasch [Psychometric Properties of an Adversity Quotient Items for Mechanical Engineering Male Students using Rasch Measurement Model]. *Malaysia Education Journal, 45*(1), 87-100. https://doi.org/10.17576/jpen-2020-45.01si-11

Mellahi, K., & Harris, L. C. (2016). Response Rates in business and management research: an overview of current practice and suggestions for future direction. *British Journal of Management*, *27*(2), 426–437. https://doi.org/10.1111/1467-8551.12154

McDonald, C. (2016). STEM Education: A review of the contribution of the disciplines of Science, Technology, Engineering and Mathematics. *Science Education International*, *27*(4), 530–569.

Moeini, S., V. Rasmussen, J., Wirenfelt Klausen, T., & Brorson, S. (2016). Rasch analysis of the Western Ontario Osteoarthritis of the Shoulder index – the Danish version. *Patient Related Outcome Measures*, *Volume 7*, 173–181. https://doi.org/10.2147/prom.s87048

National Research Council. (2011). *Successful K-12 STEM Education: Identifying Effective Approaches in Science, Technology, Engineering, and Mathematics*. National Academies Press.

Nazlinda Abdullah, Shereen Noranee, & Mohd Rahim Khamis. (2017). The Use of Rasch Wright Map in Assessing Conceptual Understanding of Electricity. *Pertanika Journal of Social Sciences & Humanities*, *25*, 81–92.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw Hill.

Othman, N., Salleh, S., Hussein, H., & Wahid, H. A. (2014). Assessing construct validity and reliability of competitiveness scale using Rasch model approach. In *WEI International Academic Conference Proceedings* (pp. 113–120). Indonesia.

Pallant, J. F., & Tennant, A. (2007). An introduction to the Rasch measurement model: An example using the Hospital Anxiety and Depression Scale (HADS). *British Journal of Clinical Psychology, 46*(1), 1–18. https://doi.org/10.1348/014466506x96931

Raharja, S., Wibhawa, M. R., & Lukas, S. (2018). Mengukur rasa ingin tahu siswa. *Polyglot: Jurnal Ilmiah*, *14*(2), 151. https://doi.org/10.19166/pji.v14i2.832

Reio Jr, T. G., & Petrosko, J. M. (2013). The measurement and conceptualization of curiosity. *The Journal of Genetic Psychology: Research and Theory on Human Development*,*167*(2), 117–135.

Schleicher, A. (2019). PISA (2020). insights and interpretations. In *OECD*. https://www.oecd.org/pisa/PISA (2020) Insights and Interpretations FINAL PDF.pdf

Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *International Journal of Medical Education*, *2*, 53-55. https://doi.org/10.5116/ijme.4dfb.8dfd

Tseng, K.H., Chang, C.C., Lou, S.J., & Chen, W.P. (2013). Attitudes towards science, technology, engineering, and mathematics (STEM) in a project-based learning (PjBL) environment. *International Journal of Technology and Design Education*, *23*(1), 87–102.

United Nations. https:// sustainabledevelopment.un.org/sdgs (accessed 23.01.2020).

Wright, B. D., & Master, G. N. (1982). Rating scale analysis. Cesa Press.

Wright, Benjamin D., & Panchapakesan, N. (1969). A Procedure for sample-free item analysis. Educational and psychological measurement. https//doi.org/10.1177/001316446902000102.

Wu, M., & Adams, R. (2007). Applying the Rasch model to psycho-social measurement: A practical approach. Educational Measurement Solutions.

Vaughan, F. (2002). What is spiritual intelligence? *Journal of Humanistic Psychology*, *42*(2), 16–33.

Young, B. J. (2000). Gender differences in student attitudes toward computers. *Journal of Research on Computing in Education*, *33*(2), 204–216.

Zohar, Danah; Marshall, Ian (2000). SQ: Connecting with Our Spiritual Intelligence. New York: Bloomsbury Publishing. ISBN 0-7475-4676-2

Zollman, A. (2012). Learning for STEM literacy: STEM literacy for learning. *School Science and Mathematics*, *112*(1), 12–19. https://doi.org/10.1111/j.1949-8594.2012.00101.x