Development and validation of 2p2s instrument by mix methods

Nancy Ng Yut Kuan and Lay Yoon Fah
Faculty of Psychology and Education, Universiti Malaysia Sabah, Malaysia
Email: mainmato@yahoo.com; layyoonfah@yahoo.com.my

Date received: February 18, 2017
Date accepted: April 6, 2017
Date published: December 15, 2017

ABSTRACT

Lack of pedagogical content knowledge in questioning skills and how they were observed, lack of co-operation and not diversified teaching methods that can stimulate the cognitive, psychomotor and affective domains in students, making all of these gaps in the research literature. Thus, the researchers makes the development and validation of instruments for identifying Mathematics teachers perception about knowledge, simulation models, clinical observation and synergy (2P2S) in Professional Learning Community: Lesson Study. The population is consist of primary and secondary schools in three districts as known Beluran, Telupid and Sandakan, in the such eastern part of Sabah. The samples were the 30 primary Mathematics teachers who were non randomly and purposive selected who have been teaching at least more than 1 year. In mix methods, researchers can analyzed and used the model IDCV which contains 10 detailed and interactive phases to develop quantitative instrument optimally. The revised 2P2S instrument validation phase of quantitative analysis involves Rasch measurement model. This instrument has good person reliability at 0.90 and an excellent item realibility at 0.92. This instrument has a Standard Error of Measurement (SE) which is low, ± 0.19 logit. Both Infit MNSQ and the z-std are close to ideal value of 1 and 0 [(Infit MNSQ Person = 0.97; z-std = -0.1), (Infit MNSQ Item = 0.98; z-std = 0.0)] which is you can imagine the suitability of instruments to measure what should be measured based on the underlying theorems.

Keywords: *2P2S, Pedagogy, Clinical, Observation, Simulation, Synergies, Malaysia.*

INTRODUCTION

Mathematics education system in Malaysia has undergone various reforms to improve the quality of curriculum and mathematics teachers as needed from time to time. This reform process should be driven by teachers. One of the educational reforms begun in our country today was the implementation of the National Key Result Areas. It was launched last July 11, 2009 and published on July 27, 2009 (Kementerian Pelajaran Malaysia., 2009), by the Prime Minister, Datuk Seri Najib Tun Razak after 100 days of his administration, which was documented by the state government initiative to realize the Malaysia agenda by Kementerian Pelajaran Malaysia (2009). It focuses on six of the National Key Result Areas which was known as NKRA as well as the Malaysian government's efforts to meet the needs of the people. Government Transformation Programme (GTP) was also launched by the Prime Minister on January 28, 2010 and the education is one of the main targets in NKRA. Often, it is preached that science and mathematics became candlestick extension that illuminates the darkness of life in the world. Even mathematical knowledge is prerequisite to the progress of civilization ever built and decisive insolence (NKRA, 2009).

For decades, schools in Sabah have been urged to engage in improving and increasing the academic's achievement because it always gets last place compared to other states throughout Malaysia. In the 1970s, research on effective schools and academic's achievement in Sabah process is through a change in the mindset of teachers themselves. Lesson study is the approach in PLC in which teachers work together to review the content, direction and how students learn and achieve mutual understanding to improve teaching and learning (T&L) in the classroom (Sato, M., 2003; Sato, M., 2006a). It is recognized by the National Development Staff Council as one of the most powerful designs to build PLC. As reported by the Ministry of Education, as well as many other countries in the world such as the United States, Britain, Australia and New Zealand, the main problem faced by the authorities is the lack of trained teachers to teach mathematics (Fernandez, Clea & Yoshida, Makoto., 2004). In most developed countries, it was found that the mathematics taught by teachers have diverse backgrounds. This situation was coupled with the ratio of teachers to students that is too high and its implications is that teachers who have diverse backgrounds impacted teaching mathematics poorly. Although there among the teachers who have self-adjustment strategies and mechanisms T&L practice in their teaching (Wang-Iverson, Patsy & Yoshida, Makoto (Editors), 2005), but for most other mathematics teachers they need guidance, collaboration, partnership to conduct their own responsibility in PLC: LS to enable them to teach effectively and meaningfully (Lewis, C., 2009). Thus, mathematics teachers are concerned with effective instructional dimensions of effective teaching and learning through PLC: lesson study (Roslee Talip et al., 2011).

In the past research or studies on aspects of knowledge in the field of education, especially among the teachers of mathematics in Sabah, pedagogical knowledge of a teacher gives students an understanding of the lesson content (Noraini Idris., 2010), has received much attention in recent years (Noor Shah Saad., 2002). These studies were conducted from inspiration driven (Shulman, L. S., 1986) which has been identified that teachers' knowledge in questioning skills was a missing paradigm in the studies of education. Researchers believe that aspects of questioning skills among the teachers as defined by Shulman, L. S. (1986), provide a very important contribution to ensuring a good mastery of the subject among students. Making up the pedagogical knowledge and knowing how to provide a better understanding to the students are two factors that can not be separated from each other and often dealt determinant of an effective teaching (Solis, A., 2009). The findings of previous studies found that most teachers of mathematics in Sabah expressed their need to increase knowledge about this subject (Shulman, L. S., 1987). Based on the results obtained by Lewis (2009) and Shulman (1991), the similarities can be concluded as the level of knowledge of mathematics teachers can be attributed to sex, duration of teacher experience, disciplinary specialization and oral communication methods during teaching. Taking into the new decades we've been through, the dynamic changes in mathematics education and curriculum as based on teachers' analysis for the recent findings in Peninsular Malaysia, the researchers felt that it was the right time to conduct a comprehensive study (Chong, A. K.., 1992; Daud Mohamad., 2000 and Noor Shah Saad., 2002) on PLC: LS among mathematics teachers throughout the state in Malaysia especially in Sabah. In PLC: LS, the situation when teachers or students ask questions, the communication occurs actively. However, teachers or students sometimes cannot clearly understand the questions posed in written or oral. More worrisome is if these problems contribute to the weakness of students in mathematics. Among the factors that contribute to it are not skilled teachers pose appropriate questions (Shulman, 1987). Therefore, this study aimed to identify the extent of the level of pedagogical content knowledge (PCK) of mathematics teachers in posing questions (questioning) in mathematics.

This learning method is capable of making students active and able to interact well with peers and other students. The students were able to create a sense of excitement to an activity carried out. All of these will attract understanding to master and achieve the objectives of subjects. Simulation method is also an activity that requires students to use their knowledge and skills in a situation that imitates or simulates real conditions (Joyce, C. R. B., McGee, H. M., & O'Boyle, C. A., 2002). Simulation concept is the same as stated in the study. This whole concept fits well with the concept of teaching in PLC: LS, which is a process that covers the activities of planning, implementation, evaluation and feedback (Johnson, C. C., 2006).

Clinical observation is an important area in teacher education. Teachers should be equipped with knowledge, pedagogical skills and skills to manage observations. This statement coincides with the findings that the education of the present emphasis is on guiding teachers, especially teachers in improving the academic quality of teaching. Improving the quality of teaching through clinical observation can help the performance of student learning. This is because, assuming all experienced teachers can conduct clinical observation is not true. They should receive special training before fully involved. Teachers' experience must be strengthened with the knowledge and skills in teaching observation. It is important because teaching and learning is not just focused on the process of teaching subjects but accounting for other tasks such as curriculum development, teaching establishment, agents of changes, staff development and research-based teaching (Baharudin & Mohd. Yusof, 2008). Clinical observation process is intended to foster and develop cooperation between teachers and administrators in schools to improve teaching and learning processes in the classroom. Furthermore, to develop the potential and capabilities of existing skills among teacher. The statement of findings supported by Goldhammer, R., Anderson, R.H., & Krajewski, R. J. (1981) and Cohen, L., Manion, L., & Morrison, K. (2007) indicatesthat the process of clinical observation has more emphasis to assist and guide the process of teaching to a higher level of quality. Researchers also can feel the importance of this clinical observation in KPP: LS as by lifestyle and technological advances that have changed. Clinical observation process is very important and should be used as an instrument for improving the quality of teaching so much. At the same time, the process of clinical observation can also be implemented by mathematics teachers to diagnose the level of teaching efficiency of their fellow friends and other teachers in the classroom.

Every teacher especially mathematics teachers are basically observable and can practice learning leadership in PLC: LS. Teachers would lead themselves and others around them to achieve shared goals of T&L. Synergies intends to build and drive PLC: LS effectively and efficiently to achieve the right outcomes. The characteristics and practices of reflection in the learning of mathematics teachers in essence, will allow the teachers to have the ability to improve their teaching through the establishment of PLC: LS efficiently. More than that, the ability to explore new ideas that generate processes, materials or innovative methods in T&L, will motivate all teachers of mathematics to implement the PLC: LS. Synergies will stimulate the mind of mathematics teachers to discuss, collaborate and create collaborative reflection that T&L would not be passive or static. Instead, PLC: LS can establish a proactive and dynamic T&L. Synergy can also form mathematics teachers to learn mathematics teaching and learning in the classroom as well as actively involved in the governance process of continuous learning in the classroom. Although sometimes there are differences of opinions and ideas from mathematics teachers, but all this will encourage improvements and changes (Prahalad C. K., & Krishnan M. S., 2008). Synergy is the spirit of mutual cooperation, interdependence and positive social skills, accountability, leadership or sense of responsibility, professionalism, clinical observation (guest) as well as soft skills such as communication, critical problem solving, teamwork skills, continuous learning and moral ethics (Prahalad C. K., & Krishnan M. S., 2008).

METHODOLOGY

To expedite the process of this study, researchers used the study design which involves the combination of quantitative and qualitative approaches (triangulation approach). This study design was chosen because it is compatible with samples that have been selected, the time allocated and the purpose of this study. According to Creswell (2009,2012), the incorporation of qualitative and quantitative approaches allow researchers to obtain more comprehensive data (Chua, Y. P., 2006).

Quantitative method

In quantitative research design, researchers used the questionnaires of 2P2S instrument which wasconstructedfor the respondents. The questionnaire is the heart of the study done by the researchers, in other words, it is the replacement for the researchers themselves between researcher and respondents (Creswell & Plano Clark, 2011). 2P2S instrument is consist of five main dimensions of pedagogical content knowledge (Skills Questioning), Simulation Model, Clinical Observations,

Synergy and Professional Learning Communities: Lesson Study. While for the assessment of the reliability and validity of the questionnaire it was tested using SPSS 16.0 and Rasch measurement model. Rasch model analysis is a very useful tool for the purpose of examining the legality or validity, and reliability of the instrument (persons and items) where it cannot be rivaled by other analysis tools (Depdikna, 2005). More than that, the Rasch model also meets the five principles of measurement which is able to provide a linear scale with the same interval, can make predictions on missing data, can provide more accurate estimate and is able to detect inaccuracies replicable model and size.

Qualitative method

The researchers used the model IDCV or Instrument Development and Construct Validation by Onwuegbuzie, Bustamante and Nelson (2010), to develop the quantitative instrument optimally. The justification for the researchers to use the IDCV model is because, it uses a combination techniques of quantitative and qualitative research methods. Therefore, it is consistent with the design of the study made by the researchers. IDCV model contains 10 detailed and interactive phases in the development and validation of instrument built. Crossover analysis represents a key mechanism in the IDCV Model which involves one or more types of analysis. For example, the use of qualitative data for the analysis of quantitative data and vice versa (Teddlie, C., & Johnson, R. B., 2009). Crossover analysis is also a research paradigm that arises as a result of the response to the current quantitative and qualitative research (Onwuegbuzie, Collins, & Leech, (2007a). Many observers say that quantitative techniques are sufficient for the development of quantitative instrument. However, qualitative techniques can be used to enhance the development of quantitative instrument and vice versa (Creswell, & Plano Clark, 2011). The following is the process IDCV (Instrument Development and Construct Validation) which consists of 10 phases. Some of them are as follows:

i. The conception of what construct to be built by interest

Instrument's construct is defined and conceptualized by the model of PLC: LS, theory and social learning by model of Bandura (in Friedman, Howard and Schustack, 2008). Review of the literature on theories and models related to five dimensions have been studied, pedagogical content knowledge, simulation model, clinical observations, synergy, and Professional Learning: Lesson study (PLC: LS).

ii. Identifying and explaining item's attributes which is fundamental to construct

Items were constructed based on the specifications of each construct. Items usually exceeds the number recommended in the construction of the instrument. In addition, the researchers also created an instrument from adapt and adopt the Ministry of Education, Teachers Management Department as well as Education District Office, interview session and identify factors with Atlas.ti. Each item was constructed with a specific score that reflects the level of perception. Cumulative scores are suitable for testing or psychological instrument where a high score reflects the ability or perception as 'Strongly Agree' high (Creswell, 2012). Each item in 2P2S Instrument is a Likert scale of 5 points which is "1" represents "Strongly Disagree", "2" represents "Disagree", "3" represents "Less Agree", "4" represents "Agree" and "5" represents "Strongly Agree".

iii. Developing early instrument

The initial instrument was developed to identify the perceptions of Mathematics teachers in Professional Learning Community: Lesson study.

iv. First field of test instrument

Field test was conducted after the construction of the instrument (early instrument). These field tests were made among the respondents, math teachers. Researchers conducted field trials for 2P2S instrument built on five teachers of the group that have the same characteristics as the study sample. They were asked to read the questions submitted and give an assessment of the level of readability and comprehension questions in the instrument. Respondents were also required to make recommendations to correct weaknesses in terms of direction by given time. Items should comply with the standard and are capable of psychometric or performance of elicited the desired perception of the respondents.

v. Design and field test instrument revised

After the first field test, the researchers redesigned the instrument that has been reviewed for the purpose of assessing the level of readability and comprehension of questions in the instrument. Second field tests on a pilot study was made after that instrument was redesigned and revised. The pilot study was conducted on 30 teachers of Mathematics randomly chosen in the 3 districts selected for the Beluran (7 respondents), Telupid (10 respondents) and Sandakan (13 respondents). The pilot study was designed to ensure that 2P2S instrument built was understandable, appropriate and relevant in providing necessary information within the scope of this study.

vi. Ratification of the revised instrument: Phase of quantitative analysis

The results from the pilot study involved several purification and stabilization questionnaire sessions. Thus, after a pilot study was made, the researchers determined the validity and reliability of the instrument that have been successfully produced. The revised instrument validation phase of quantitative analysis involved Rasch measurement model with software Bond and Fox (2010). It was used to get the reliability, validity and determine the appropriateness of individual items. The researchers analyzed instrument items after making a pilot study using SPSS 16.0 (key in data) and Rasch model for the reliability, authenticity, clarity and accuracy of the items (how precise and how exact the items are estimates) with person ability (how much/ quantity on target of the latent variables are evident)

vii. Ratification of the revised instrument: Phase of qualitative analysis

The ratification of the revised instrument on qualitative analysis phase involved reviewing items based on comments and input obtained from the panel of experts (four people) who checked the 2P2S instrument. Therefore, items werereviewed by an expert panel consisting of lecturers in institutes of higher learning schools or educational faculty in the field of expertise Mathematics and PLC: LS (Learning Management Community: Lesson study) using the Delphi technique. The expert panel is a panel in Mathematics who reviewed, updated, combined and provided recommendations for improvement items from the item reservoirs into a new instrument. The expert panel also reviewed the items to be clear, easy to understand and not biased item to culture, race and gender. Checking in terms of language wasalso made by linguists to avoid spelling and grammatical errors. The researchers reviewed the items found by the panel of experts as inappropriate in terms of content or does not conform to the construction of psychometric items according to expert opinion. The researchers rewrote the items, repaired the language, vocabulary and grammar. The researchers also dropped items that were found unsuitable according to the views of the experts.

viii. Ratification of the revised instrument: Mixed Phase Analysis - Qualitative Analysis of Dominant Crossover

Qualitative analysis of dominant crossover involving the face and content validity. The instrument went through face validity. Face validity of the assessment process is a visual evaluation made by the appointed panel of expertsFace validity also involved the social acceptance of a technical nature and not as content validity, criterion validity and construct validity as required on systematic statistical analysis. Content validity refers to a measurement tool that can measure all the content areas assessed effectively. Content validity is also a means of measuring the test's ability, skills and behaviour of respondents who took the survey or a specific test. The procedure involved an examination content validity of the test from two perspectives, the relationship of the domain specification and content representation to the specification test.

ix. Ratification of the revised instrument: Mixed Phase Analysis - Quantitative Analysis of Dominant Crossover

Quantitative analysis involved crossover dominant construct validity. It aimed to look at the extent to which an instrument measures what it is supposed to be accurately measured before the test is considered valid. According to Creswell (2012), construct validity is the most complicated because it is evaluated using both statistical and practical procedures. Therefore, before the questionnaires

were administered to the sample of the pilot study, the face validity and content validity had been established first then followed by the construct validity.

x. Instrument development judgement/ Assessment of construct, process and products

It means that the item was on the psychometric aspects in which items that do not meet the criteria of validity and reliability were removed. Then, the final version of the instrument successfully produced at least 5 to 20 items for each construct or dimension. Each item in 2P2S Instrument is a 5-point Likert scale / point of "1" represents "Strongly Disagree", "2" "represents "Disagree", "3" represents "Less Agree", "4" represents "Agree" and "5" represents "Strongly Agree". Issue of the instrument's validity can be handled using the Rasch measurement model. The model is unidimensional model based on the assumption that individuals (latent variable) are capable and knowledgeable, have more probability to answer all items with certain perception as well as, the probability of a simple item can be answered by all respondents with the perceptions of each, when controlled by the difference between the item difficulty (facility) and the ability of respondents (Wright and Stone, 1979; Bond and Fox, 2007). Constructs assessment, processes and products on the instrument development are some matters that cannot be avoided. This is because, it is an essential element before proceeding it to inferential statistics to answer desired research questions and hypotheses.

Population and sampling

The population of this study are mathematics teachers in primary and secondary schools in three districts of Eastern Sabah such as Beluran, Telupid and Sandakan. According to sources obtained from the Sector of School Management in the Sabah State of Education Department in 2015, the number of teachers who teachmathematics is 400 including opsyens and non mathematical opsyens. Sampling technique used is non-probability sampling (non random sampling), known as purposive sampling which is used to select a sample that would provide good cooperation, easily accessible and are selected based on knowledge or past experience (Cohen, L., Manion, L., & Morrison, K., 2007). This technique is widely used (Creswell, J. W., 2012) and the researchers chose it because it coincided with a survey, descriptive and exploratory, aimed at giving a rough idea of the phenomenon that exists in a population study (Creswell, J. W., 2012). The reasons for choosing the sample also is making judgments about the suitability of the respondent groups such as years of service as a teacher who teaches mathematics. The sample size in this study was determined by the method of Cohen, L., Manion, L., & Morrison, K. (2007).

They argue that the determination of the sample size should take into account the significant level and sampling error. In this study, researchers selected a sampling error of 1% and the level of reliability of 99% (significant level = .01) as well as confidence interval of 3%. The population of 400 by 1% sampling error condition and level of reliability (confidence level) of 99% was sufficient to make an analysisit is because, sample is considered to be the mirror of the population where it was found. However, no guarantee that the sample is truly a representative of the population where it came from. One of the reasons that often occur is the sampling error. Inaccuracy statistical sample estimates the population parameters determined by sampling error. Sampling error is the error that occurs when the sample is used to infer the population. Sampling error is when the difference or variation between the mean of a random sample of the population mean is normally distributed. The greater the standard deviation (SD) of the sample, the larger the sampling error exists. The larger the sample size, the smaller the sampling error. Sampling error is working directly with the sample size and the population standard deviation. If the sample size is fixed, the sampling error will increase if the population standard deviation increases and vice versa. Increasing the sample size (n) increases or approaching N, the sampling error decreases, because when n increases, the standard deviation of a population declines. The small sample size n is more prone to sampling error compared to n large. Therefore, the objective of the sampling is to reduce sampling error and sampling error can be reduced by adding n.

Thus, the researchers determined the size of the sample as suggested by Cohen, L., Manion, L., & Morrison, K. (2007) for a population of not more than 400 is 329 respondents. Sample size determination was made based on six main aspects such as the type of research, study design, type of

population, sampling frame, the sample size required by Cohen, L., Manion, L., & Morrison, K. (2007), research questions and objectives (combination techniques of quantitative and qualitative research by IDCV Model) as well as measurement scale (interval data by Rasch Model). This population is heterogeneous, which has several sub samples such as gender, teaching options, served different periods of time, the large sample size is needed. To rank the pilot study, the researchers only selected 30 samples by purposive sampling.

RESULTS AND DISCUSSION

Face validity and content validity by expert panel reviewers

Face validity was used to ensure item clarity, questions posed, sufficient time and most important item to measure what should be measured (Creswell, J. W., 2012). Researcher conducted a pre-test of the instrument with some teachers of mathematics who have same characteristics as the samples of the study. They were asked to read the questions submitted and give an assessment about the level of readability and comprehension questions in the 2P2S instrument. Face validity cannot guarantee whether the test really measures phenomena in the domain (Noraini Idris, 2001; Noraini Idris., 2010). Therefore, researcher had made the validity in advance followed by a review of the legality of the content by a panel of experts. It aims to enhance the development and validation of this 2P2S instrument.

Results of the review panel of experts is consist of four institutions of higher education lecturers who are experts in the field of mathematics and the lesson study, had received feedback on the draft electoral questions, grammar and suitability of routine and non-routine questions, writing sentences too long or short sentences to evaluate two things in one item. The expert panel members also asked to write the steps to resolve the problems that are not listed in the instrument. Thus, the 2P2S instrument has 5-dimensions and is consist of 146 items.

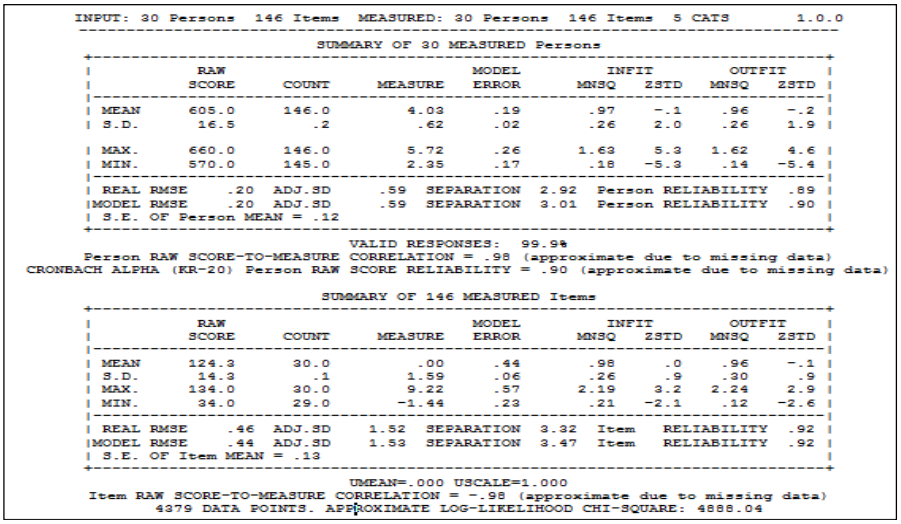
Validity and reliability with Rasch analysis

In this study, researcher referred to the Table of Quality Criteria of Fisher Instrument Rating Scale (Fisher, W. P. Jr., 2007) as shown in Table 1.

Table 1. Quality Criteria of Fisher Instrument Rating Scale

Criteria	Weak	Medium	Good	Very Good	Excellent
Targeting	> 2 errors	1-2 errors	< 1 error	< .5 error	< .25 error
Item Model Fit Mean Square Range Extremes	< .33 – > 3.0	.34 – 2.9	.5 – 2.0	.71 – 1.4	.77 – 1.3
Person And Item Measurement Reliability	< .67	.67 – .80	.81 - .90	.91 - .94	> .94
Person And Item Strata Separated	2 or less	2 – 3	3 – 4	4 – 5	> 5
Ceiling Effect: % maximum extreme scores	> 5%	2% – 5%	1% - 2%	.5% - 1%	< .5%
Floor effect: % minimum extreme scores	> 5%	2% – 5%	1% - 2%	.5% - 1%	< .5%
Variance in data explained by measures	< 50%	50% - 60%	60% - 70%	70% - 80%	> 80%
Unexplained variance in contrasts 1-5 of PCA of residuals	> 15%	10% - 15%	5% - 10%	3% - 5%	< 3%

i. Reliability and Separation Index Analysis For 2P2S Instrument



```
INPUT: 30 Persons  146 Items  MEASURED: 30 Persons  146 Items  5 CATS       1.0.0

                        SUMMARY OF 30 MEASURED Persons

+-------------------------------------------------------------------------------+
|            RAW                          MODEL          INFIT        OUTFIT    |
|           SCORE     COUNT     MEASURE   ERROR      MNSQ   ZSTD   MNSQ   ZSTD  |
|-------------------------------------------------------------------------------|
| MEAN      605.0     146.0        4.03     .19       .97    -.1    .96    -.2  |
| S.D.       16.5        .2         .62     .02       .26    2.0    .26    1.9  |
|                                                                               |
| MAX.      660.0     146.0        5.72     .26      1.63    5.3   1.62    4.6  |
| MIN.      570.0     145.0        2.35     .17       .18   -5.3    .14   -5.4  |
|-------------------------------------------------------------------------------|
| REAL RMSE    .20  ADJ.SD     .59  SEPARATION  2.92  Person RELIABILITY  .89    |
|MODEL RMSE    .20  ADJ.SD     .59  SEPARATION  3.01  Person RELIABILITY  .90    |
| S.E. OF Person MEAN = .12                                                     |
+-------------------------------------------------------------------------------+

                        VALID RESPONSES:  99.9%
    Person RAW SCORE-TO-MEASURE CORRELATION = .98 (approximate due to missing data)
CRONBACH ALPHA (KR-20) Person RAW SCORE RELIABILITY = .90 (approximate due to missing data)

                        SUMMARY OF 146 MEASURED Items

+-------------------------------------------------------------------------------+
|            RAW                          MODEL          INFIT        OUTFIT    |
|           SCORE     COUNT     MEASURE   ERROR      MNSQ   ZSTD   MNSQ   ZSTD  |
|-------------------------------------------------------------------------------|
| MEAN      124.3      30.0         .00     .44       .98     .0    .96    -.1  |
| S.D.       14.3        .1        1.59     .06       .26     .9    .30     .9  |
| MAX.      134.0      30.0        9.22     .57      2.19    3.2   2.24    2.9  |
| MIN.       34.0      29.0       -1.44     .23       .21   -2.1    .12   -2.6  |
|-------------------------------------------------------------------------------|
| REAL RMSE    .46  ADJ.SD    1.52  SEPARATION  3.32  Item   RELIABILITY  .92    |
|MODEL RMSE    .44  ADJ.SD    1.53  SEPARATION  3.47  Item   RELIABILITY  .92    |
| S.E. OF Item MEAN = .13                                                       |
+-------------------------------------------------------------------------------+

                        UMEAN=.000 USCALE=1.000
  Item RAW SCORE-TO-MEASURE CORRELATION = -.98 (approximate due to missing data)
          4379 DATA POINTS. APPROXIMATE LOG-LIKELIHOOD CHI-SQUARE: 4888.04
```

Figure 1. Reliability and separation index analysis for 2P2S instrument

Figure 1, shows the total of 4379 data points resulting from the interaction of the 30 respondents to the 146 items analyzed to produce ready-to-Khi-squared value of 4888.04. Cronbach alpha (KR-20) gross score test showed the level of reliability, which is 0.90, which illustrates the ordinal score marks. Rasch analysis in Figure 2 also shows the high reliability item, that is 0.92 which indicates adequacy of items to measure what should be measured. High quality item that shows he is able to separate individual with a good separation of power, the Person Separation = 2.92. Generally, a math teacher or whether the option is not an option in schools, available practice management Professional Learning Community (KPP) in Lesson Study (LS) are both at the level of μ = + 4:03 logit depicted on the mean person.

This figure also shows the maximum level of measurement item is +9.22 logit (SE: 0.57) compared with the maximum measurement height of +5.72 Person logit (SE: 0.26). Range scale items that can be used also only from +2.35 to +9.22 logit wide is 6.87 logit. There is a significant gap that shows teachers who teach mathematics are free to items, wide 3.5 logit of +5.72 logit to +9.22 logit showing management practices in KPP: LS is high, representing primary and secondary schools. Besides, this figure also shows that there are items that are free Person measurement item showing a much lower minimum at -1.44 logit compared with the math teacher who at least practice LS, which is expected at the level of logit only 2.35. This shows some of the items as wide as 3.79 logit become common practice for teachers to teach mathematics in the classroom at school. This instrument has a Standard Error of Measurement (SE) which is low, ± 0.19 logit. Both Infit MNSQ and the z-std are close to ideal value of 1 and 0 [(Infit MNSQ Person = 0.97; z-std = -0.1), (Infit MNSQ Item = 0.98; z-std = 0.0)] which is you can imagine the suitability of instruments to measure what should be measured based on the underlying theorems.

Table 2. Criterion validity in 2P2S instrument

Criteria	Statistic Info	Result
Item Validity Item = 146	Item Polarity Incompatibility item PCA Of Residuals Respondents' Reliability Items' Reliability	All items show the PTMEA CORR> 0.1 All items showed a mean squared Infit between 0.72 to 1.24 (Level Of Very Good) and Outfit between 0.66 to 1.26 (Level Of Good) (Fisher, W. P. Jr., 2007). There are 52 items showed Misfit value <0.72 and > 1.24 because the Item Infit MNSQ range is 0.72 to 1.24, has 8 items in negative value and has 39 redundant items. Rasch dimensional recorded 60.1% of the variance

		which is equal to 60.1% of the model. Respondents' reliability were 0.89 Items' reliability were 0.92
Individual Distribution	Distribution of teachers answered the 2P2S Instrument questions/items	Over the logits 5 (from 2.35 to 5.72) which is getting closer to the logits 6.0
Item Distribution	Distribution of items answered by teachers in 2P2S Instrument	Over the logits 9 (from -1.44 to 9.22) which is getting closer to the logits 10.0
Validity of Respondents' Response	The Percentage Of Mean Square For Respondents Between 0.5 - 2.0	Infit 6.7% < 0.71 6.7% > 1.23 Outfit 10% < 0.71 6.7% > 1.23
Validity of Items' Response	The Percentage Of Mean Square For Items Between 0.5 - 2.0	Infit 8.9 % < 0.72 15.8 % > 1.24 Outfit 12.3% < 0.72 16.4% > 1.24

ii. Item Suitability

Results of the survey carried out that the mean square Infit item is between 0.72 to 1.24 (Table 3). While the mean square Infit individual is between 0.71 to 1.23 (Table 4).

Table 3. Validity (Item Suitability) items' response (item) on the 2P2S instrument

Mean Square Value (MNSQ)	Infit		Outfit	
	Frequency	Percentage (%)	Frequency	Percentage (%)
Less than 0.72	13	8.9	18	12.3
0.72 to 1.24	110	75.3	104	71.3
More than 1.24	23	15.8	24	16.4
Total	146	100	146	100

Table 4. Validity (Item Suitability) Respondents' Response (Individual) on the 2P2S Instrument

Mean Square Value (MNSQ)	Infit		Outfit	
	Frequency	Percentage (%)	Frequency	Percentage (%)
Less than 0.71	2	6.7	3	10
0.71 to 1.23	26	86.6	25	83.3
More than 1.23	2	6.7	2	6.7
Total	30	100	30	100

ENTRY NUMBER	RAW SCORE	COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PTMEA CORR.		EXACT MATCH OBS%	EXACT MATCH EXP%	Item
8	124	30	.28	.49	2.19	2.6	2.24	2.4	A	.20	66.7	83.8	B2
126	124	30	.28	.49	1.76	1.8	1.79	1.7	B	.23	73.3	83.8	F5
1	34	30	9.22	.52	1.03	.2	1.76	1.2	C	-.11	83.3	86.8	A1
44	130	30	-.85	.39	1.60	3.2	1.69	2.9	D	.17	60.0	67.8	C13
139	121	30	1.14	.57	1.57	1.1	1.65	1.2	E	.22	83.3	89.5	F18
41	126	30	-.15	.45	1.40	1.4	1.64	1.7	F	-.04	76.7	78.7	C10
13	129	30	-.69	.40	1.53	2.5	1.53	2.1	G	.29	63.3	70.5	B7
7	123	30	.54	.52	1.37	.9	1.53	1.2	H	-.07	83.3	86.2	B1
122	123	30	.54	.52	1.40	1.0	1.50	1.1	I	-.13	83.3	86.2	F1
57	131	30	-1.00	.39	1.42	2.6	1.48	2.4	J	.00	56.7	65.1	C26
60	128	30	-.52	.41	1.46	2.0	1.43	1.6	K	.42	63.3	73.3	C29
3	70	30	6.18	.23	1.46	1.9	1.42	1.7	L	.44	26.7	42.6	A3
49	125	30	.05	.47	1.28	.9	1.45	1.2	M	.19	80.0	81.2	C18
107	124	30	.28	.49	1.44	1.2	1.33	.9	N	-.07	80.0	83.8	E13
121	126	30	-.15	.45	1.44	1.5	1.44	1.3	O	-.01	70.0	78.7	E27
106	125	30	.05	.47	1.44	1.3	1.25	.7	P	.02	73.3	81.2	E12
127	125	30	.05	.47	1.32	1.0	1.41	1.1	Q	.22	73.3	81.2	F6
2	40	30	8.20	.34	1.02	.2	1.41	1.1	R	-.17	53.3	70.0	A2
45	131	30	-1.00	.39	1.35	2.3	1.32	1.7	S	.12	63.3	65.1	C14
124	123	30	.54	.52	1.32	.8	1.34	.8	T	.04	83.3	86.2	F3
51	127	30	-.35	.43	1.34	1.3	1.29	1.0	U	.13	73.3	76.0	C20
46	125	30	.05	.47	1.26	.9	1.33	.9	V	.20	80.0	81.2	C15
43	131	30	-1.00	.39	1.22	1.5	1.33	1.7	W	-.16	63.3	65.1	C12
4	94	30	4.88	.25	1.31	1.3	1.32	1.3	X	.32	33.3	47.3	A4
137	124	30	.28	.49	1.27	.8	1.15	.5	Y	.23	80.0	83.8	F16
56	133	30	-1.30	.38	1.26	2.1	1.21	1.4	Z	.26	66.7	60.8	C25
BETTER FITTING OMITTED													
12	128	30	-.52	.41	.82	-.8	.75	-1.0	z	.47	76.7	73.3	B6
140	124	30	.28	.49	.82	-.4	.73	-.6	y	.22	86.7	83.8	F19
22	126	30	-.15	.45	.81	-.6	.79	-.6	x	.37	83.3	78.7	B16
36	129	30	-.69	.40	.81	-1.0	.74	-1.2	w	.31	76.7	70.5	C5
33	125	30	.05	.47	.80	-.5	.75	-.6	v	.32	86.7	81.2	C2
61	129	30	-.69	.40	.80	-1.1	.75	-1.1	u	.51	76.7	70.5	C30
79	126	30	-.15	.45	.80	-.7	.71	-.9	t	.42	83.3	78.7	D11
62	128	30	-.52	.41	.79	-1.0	.72	-1.1	s	.51	76.7	73.3	C31
81	126	30	-.15	.45	.79	-.7	.70	-.9	r	.44	83.3	78.7	D13
14	131	30	-1.00	.39	.78	-1.6	.73	-1.6	q	.58	76.7	65.1	B8
80	126	30	-.15	.45	.78	-.8	.68	-1.0	p	.46	83.3	78.7	D12
101	125	30	.05	.47	.76	-.7	.67	-.9	o	.41	86.7	81.2	E7
64	127	30	-.35	.43	.76	-1.0	.68	-1.2	n	.53	80.0	76.0	C33
138	123	30	.54	.52	.70	-.7	.60	-.9	m	.25	90.0	86.2	F17
100	124	30	.28	.49	.69	-.8	.60	-1.0	l	.41	86.7	83.8	E6
21	124	30	.28	.49	.67	-.9	.61	-1.0	k	.44	86.7	83.8	B15
5	89	30	5.18	.24	.58	-2.0	.65	-1.6	j	.06	70.0	47.4	A5
48	123	30	.54	.52	.62	-.9	.62	-.8	i	.31	90.0	86.2	C17
39	124	30	.28	.49	.60	-1.1	.46	-1.5	h	.59	86.7	83.8	C8
31	122	30	.83	.55	.59	-.9	.57	-.8	g	.10	93.3	88.1	B25
40	123	30	.54	.52	.56	-1.1	.43	-1.5	f	.50	90.0	86.2	C9
24	123	30	.54	.52	.52	-1.3	.37	-1.7	e	.58	90.0	86.2	B18
38	123	30	.54	.52	.52	-1.3	.37	-1.7	d	.58	90.0	86.2	C7
10	122	30	.83	.55	.46	-1.3	.42	-1.3	c	.36	93.3	88.1	B4
19	121	30	1.14	.57	.37	-1.4	.40	-1.3	b	.00	96.7	89.5	B13
9	121	30	1.14	.57	.21	-2.1	.12	-2.6	a	.51	96.7	89.5	B3
MEAN	124.3	30.0	.00	.44	.98	.0	.96	-.1			75.5	75.1	
S.D.	14.3	.1	1.59	.06	.26	.9	.30	.9			10.2	8.5	

Figure 2: Item statistics: Misfit order

To assess the impact of any misfit, respond with suspicion and do not meet the characteristics of the measurement as the MNSQ <1, the z-std is> ± 2, PMC is negative, will be improved or dropped. In this study, it was found that there are 52 items which misfit. 13 of them are maintained and accepted as having value MNSQ Infit items that are in the range of 0.72 to 1.24, the value z-std are within the range of -2 to +2, the PMC positive and non-recurring measure logit (redundant). While 39 items were dropped because PMC for 8 items were negative, which that the items being measured is not in the same direction. In addition, the researchers also reviewed the MNSQ items that are more than 1.24 and less than 0.72, the value z-std more than ± 2 and redundant logit measure. This forecast was reinforced by the high Outfit z-std, t => ± 2. (*Figure 3* and *Figure 4*).

Items piled the items that have the same logit be regarded as redundant items which are measure the same strength would be omitted. This situation is one of the signs of distrust in Rasch analysis and further strengthen an item, then the item is to be dropped. Although the researchers can choose to let only those items like this but it could affect the measurement of the length of time the respondents to complete the survey response and the accuracy of the measurement for more items to be answered, so the more time is needed and many recurring items. Accordingly, a more appropriate way of language should be revised to measure what is desired or what you want to drop.

ENTRY NUMBER	RAW SCORE	COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PTMEA CORR.	EXACT OBS%	MATCH EXP%	Person
7	611	146	4.29	.18	1.63	5.3	1.62	4.6	A .67	58.9	71.3	7
2	618	146	4.51	.18	1.48	4.9	1.45	4.2	B .70	56.8	67.5	2
17	605	146	4.08	.19	1.17	1.5	1.19	1.4	C .58	74.7	74.8	17
10	603	146	4.01	.19	1.12	1.0	1.18	1.3	D .62	76.0	76.0	10
18	610	146	4.25	.18	1.17	1.6	1.18	1.5	E .71	70.5	71.9	18
1	595	146	3.70	.20	1.15	1.0	1.16	1.0	F .75	76.7	80.4	1
21	633	146	4.96	.17	1.06	.9	1.10	1.5	G .70	63.0	61.5	21
15	590	146	3.48	.21	1.09	.6	1.02	.2	H .73	81.5	83.0	15
11	602	146	3.97	.19	1.08	.7	1.07	.6	I .67	74.7	76.6	11
13	598	146	3.82	.20	.98	-.1	1.07	.5	J .72	80.1	78.8	13
30	583	146	3.13	.23	.98	.0	1.06	.3	K .79	85.6	86.1	30
23	610	146	4.25	.18	.98	-.1	1.03	.3	L .70	71.9	71.9	23
25	601	146	3.93	.19	1.01	.1	.96	-.2	M .77	75.3	77.1	25
19	607	146	4.15	.19	1.01	.1	.94	-.5	N .78	72.6	73.6	19
4	660	146	5.72	.17	1.00	.0	.98	-.3	O .72	62.3	63.2	4
8	604	146	4.05	.19	.98	-.1	.86	-1.0	o .77	74.7	75.4	8
9	622	146	4.63	.17	.97	-.3	.98	-.2	n .76	71.2	65.3	9
5	602	146	3.97	.19	.97	-.2	.94	-.3	m .80	76.0	76.6	5
24	602	146	3.97	.19	.93	-.6	.96	-.2	l .70	78.8	76.6	24
29	599	145	4.02	.19	.91	-.7	.91	-.7	k .71	77.9	75.8	29
28	622	146	4.63	.17	.91	-1.1	.82	-2.1	j .70	68.5	65.3	28
22	608	146	4.19	.19	.90	-.9	.91	-.7	i .72	74.7	73.0	22
20	608	146	4.19	.19	.88	-1.1	.87	-1.1	h .73	74.7	73.0	20
14	608	146	4.19	.19	.87	-1.2	.85	-1.2	g .73	74.7	73.0	14
12	610	146	4.25	.18	.86	-1.3	.84	-1.4	f .69	71.9	71.9	12
27	612	146	4.32	.18	.80	-2.1	.76	-2.3	e .77	76.7	70.8	27
16	593	146	3.61	.21	.78	-1.5	.74	-1.6	d .76	84.9	81.5	16
3	589	146	3.43	.22	.72	-1.8	.70	-1.7	c .80	87.7	83.5	3
6	575	146	2.67	.25	.44	-3.1	.54	-2.1	b .86	94.5	88.6	6
26	570	146	2.35	.26	.18	-5.3	.14	-5.4	a .91	98.6	89.6	26
MEAN	605.0	146.0	4.03	.19	.97	-.1	.96	-.2		75.5	75.1	
S.D.	16.5	.2	.62	.02	.26	2.0	.26	1.9		8.9	6.8	

Figure 3. Person statistics: Misfit order

Reliability can be affected negatively as a result of poor response and uncertainties of a misfit Person. The response that causes distortion in the actual measurement should be set aside. Data from respondents can be categorized as such data can not be trusted. To identify patterns of different Person of the ideal, Table 6 shows the Person misfit intended. Just as the item, quantity Infit MNSQ Mean ± SD, 0.97 + 0:26 = 1.23, while the difference Infit MNSQ Mean - SD, 0.97 - 0:26 = 0.71 should be revised. Person Infit range is between 0.71 to 1.23. Respondents who have Point Correlation Measure (PMC) is negative, indicating that perceptions of decision-making or unusual. This is a case of observation to demographic characteristics that may contribute to such behavior. The following display will indicate Person misfit concerned. It was found that there were respondents in the 2nd, 5th, 6th, 7th, 14th, 20th, 22th, 24th and 26th are misfit person (the person who differs from the ideal pattern). The researchers also looked at measure if there is the same logit or repeated (redundant) and greater value than z-std, t> ± 2, which was showed that the respondents were unable to give required rating scale (perception of unforeseeable circumstances).

iii. Unidimensional

Table 5. Variance of standardized residuals (In Eigenvalue Unit)

Table of STANDARDIZED RESIDUAL variance (in Eigenvalue Units)					
		(Observed)	Empirical	(Expected)	Modeled
Total variance in observations	=	365.6	100.0%		100.0%
Variance explained by measures	=	219.6	60.1%		60.1%
Unexplained variance (total)	=	146.0	39.9 %	100.0 %	39.9 %
Unexplained variance in 1st contrast	=	12.2	3.3 %	8.3 %	
Unexplained variance in 2nd contrast	=	12.1	3.3 %	8.3 %	
Unexplained variance in 3rd contrast	=	11.4	3.1 %	7.8 %	
Unexplained variance in 4th contrast	=	10.0	2.7 %	6.8 %	
Unexplained variance in 5th contrast	=	9.0	2.5 %	6.2 %	

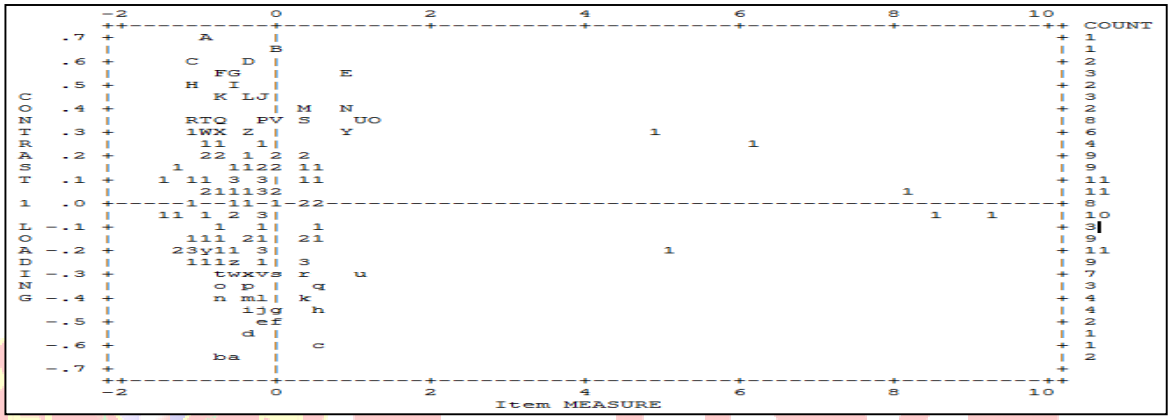


Figure 4. Standardized residual contrast 1 plot

Based on Table 5 above, the researchers found that the variance is 60.1%. Therefore, the variance is greater than 60% and has good level for acceptance. Therefore, in addition to having the reliability of the individual and the well-received item (0.89 and 0.92 respectively), this 2P2S instrument was unidimensional which is describing items by interacting with both attributes. While the raw variance explained by measures is 39.9% for the same empirical and the model of 39.9% as well. The measured noise level is 3.3%. It means that the noise level of the instrument 2P2S are in scale very well (Fisher, W. P. Jr., 2007). According to Figure 5, it shows Standardized Residual Plots Contrast 1, which shows a difference in residuals for respondents. Each letter is one individual (respondent) to a maximum of 30 people, AZ and az. For people who are 53-74, "1" means that there is one person at a location in the plot. "2" means that there are two persons and so on. The Rasch dimension explains 60.1% of the variance in the data is good (Fisher, 2007). The largest secondary dimension, "the first contrast in the residuals" explains 3.3% of the variance which is somewhat greater than around 15% that would be observed in data like these simulated to fit the Rasch model and not over than 15%. The eigenvalue of the first contrast is 12.2. This indicates that it has the strength of about more than 5 items, somewhat bigger than the strength of five items (an eigenvalue of 5), the smallest amount that could be considered a "dimension". Contrast the content of the items at the top and bottom of the plot in Table 5 to identify what this secondary dimension reflects. In addition, there is no correlation between the item that has the standard large residual correlation, exceeding the 0.70 level control. This suggests that respondents see a couple of items relating not same subject matter and the instruments independent of any confusion in terms of the purpose and intent of the survey was conducted (Azrilah et al., 2013).

Reliability index for 2P2S instrument

According to Fisher (2007), as shown in Table 1, the reliability was good at 0.81 and excellent at > 0.94. The analysis showed that both values have proved this 2P2S instrument has good reliability and strong to be used to identify teachers' perception towards mathematics pedagogical content

knowledge, simulation models, clinical observation and synergy in the implementation of the PLC: LS.

Construct reliability and validity for 2P2S instrument

Table 6. Reliability and validity coefficients for each construct in instrument

Five Constructs In 2P2S Instrument	Numbers Of Items	Cronbach's Alpha	Person Raw Score-To-Measure Correlation	Item Raw Score-To-Measure Correlation	S.E of Person Mean	S.E of Item Mean
Pedagogical Content Knowledge (Questioning Skills)	25	.78	.98	- .99	.20	.16
Simulation Models	37	.87	.99	- .97	.19	.10
Clinical Observation	26	.85	.94	- 1.00	.28	. 09
Synergy	27	.62	.95	-1.00	.18	.09
Professional Learning Community: Lesson Study	25	.83	.96	- .99	.23	.10

Overall, the constructs which had positive and negative correlations were significant. Thus, these constructs had the validity of a satisfactory and successfully measured the perception of teachers in the implementation of lesson study. So, these constructs are suitable.

Probability curve

In terms of functional categories for Outfit and Infit matching items, the findings indicated that the value is 1 which is between 0.7 and 1.3. According Fisher, W. P. Jr. (2007), the rating scale and the mean square, Infit and Outfit are good and acceptable on 0.7 to 1.3. However, the response of the 5-points Likert scale was not clear response structure (12345) in Figure 3 below. It shows the scale of category 1 (Strongly Disagree) and scale of category 2 (Disagree) were protected categories under category 3, category 4 and category 5. It shows that all the responses category 3, category 4 and category 5 were function better than category 1 and category 2.

Based on these findings, category 1 (Strongly Disagree) and category 2 (Disagree) sheltered used as category 3 (Poor Agree) and category 4 (agree) to scale 34345 by 3-points scale and the results are shown in Figure 5 below. Figure 6 shows that no category of response which protected and thus all response 34345 of 3-points scale works well. Therefore, this 2P2S instruments work better in a 3-point scale (34345) compared with the 5-points scale (12345).

CATEGORY LABEL	OBSERVED SCORE	OBSERVED COUNT	%	OBSVD AVRGE	SAMPLE EXPECT	INFIT MNSQ	OUTFIT MNSQ	STRUCTURE CALIBRATN	CATEGORY MEASURE	
1	1	78	2	-4.35	-4.51	1.16	1.07	NONE	(-4.36)	1
2	2	40	1	-3.12	-2.81	1.30	1.62	-3.06	-2.63	2
3	3	77	2	1.21	1.01	1.04	.85	-1.98	-1.14	3
4	4	3159	72	4.13	4.15	.97	.97	-.56	2.53	4
5	5	1025	23	4.82	4.78	.98	.95	5.61	(6.71)	5
MISSING		1	0	4.21						

Figure 5. The Structure of the calibration scale rating

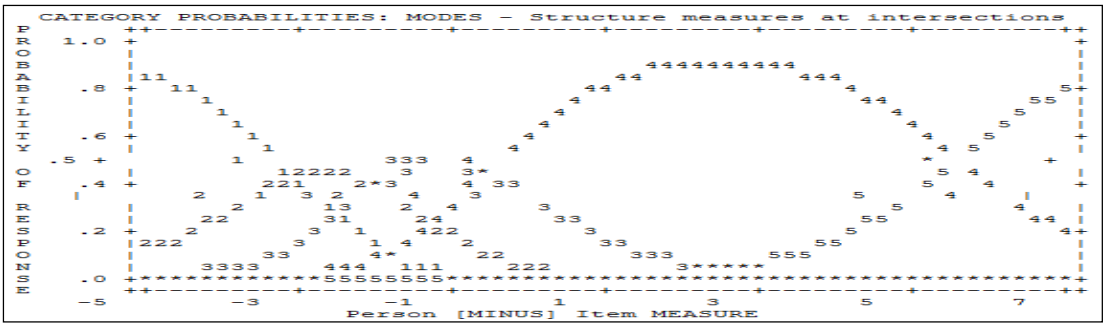


Figure 6. Probability curve In 5-points Likert scale (12345)

Calibration scale is an important element in any system of measurement and data validity. The validity of the scale has been determined whether the data collected are valid for analysis and processing, thus able to generate meaningful information. What is certain that the Rasch model is able to offer is like a simple method to produce more logical instrument calibration using Likert scale. The instrument is not calibrated will publish the data can not be used for purposes analysis. Observed average increased consistently and evenly from -4.35 to 4.82, which shows consistency in response patterns. Rasch - Andrich Threshold is used to enable the formulation of the calibration scale response for each item. This threshold shows the change during an individual's decision-making process of moving from one scale to the scale of the next and so on. This value is expressed in the calibration structure which showed a difference of a subsequent threshold should exceed 1.4 but not exceeding the value 5. If the difference is less than 1.4, then the rating scale was supposed to be collapsed and if the difference is over than 5, the rating is would be separated.

From Figure 7, scale rating of 1 and 2 (threshold = 3.06), scale rating of 2 and 3 (threshold = 1.08), scale 3 and 4 (threshold = 1.42), as well as scale rating of 4 and 5 (the threshold = 6.17). For scale rating of 1 and 2 and 3 and 4, the threshold exceed by 1.4 and less than 5. Therefore, scale 1, scale 3 and scale 4 were maintained. While the scale rating of 2 and 3 (threshold = 1:08) was less than 1.4. Therefore, a rating scale was supposed to be collapsed because scale rating of 2 and 3 were found to be too close to the isolation of less than 1.4 points. Thus, a new scale to be tested can involve a combination of options such as 12245. As scale rating of 4 and 5 (threshold = 6.17) was supposed to be separated. Thus, a new scale to be tested can involve a combination of options such as 12234. Any combination of these calibrations either 12245 or 12234 will be chosen as the best when obtained the smallest Infit MNSQ and SD for each item and person, as well as the biggest value of person separation.

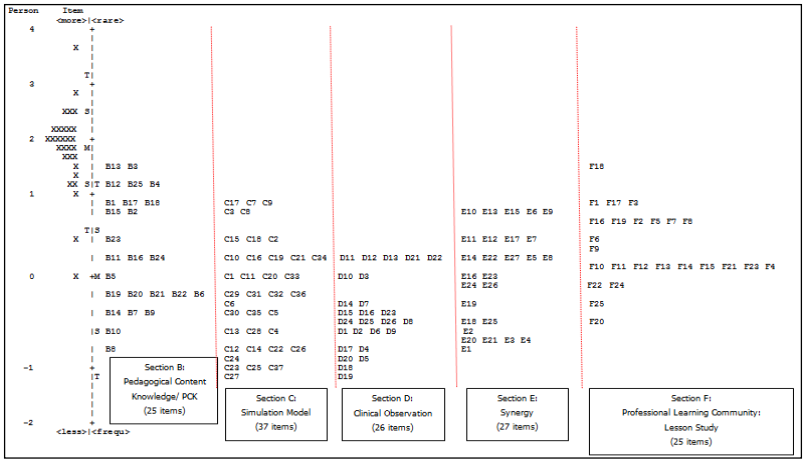


Figure 7. Item difficulty mapping

Figure 7 shows the empirical percentage for 2P2S instrument as defined by the results of the analysis item in Rasch model. Min for the measurement item was between -1.15 logit to 1.49. Researcher found that the distribution of items in this 2P2S instrument in the form of a symmetrical arc-shaped and has a positive skew. It means, the items in this 2P2S instrument were appropriated for measurement. From Figure 5 also showed the distribution of the respondents was in the top right. It means that all respondents can respond on their perception of the items 2P2S instrument and all items can be answered by the respondents. According to Fisher (2007) and Linacre (2011), the respondents were able to answer or at least agree on the difficulty items which had high ability while simple items were could be answered or agreed to by the respondents with high and low ability. Based on the mapping of item difficulty (item map), it was found that respondents with high ability did not have problems to agree on all the items in this 2P2S instrument.

CONCLUSION

The results of the analysis, enabled several iterations of the analysis included data generated by some combination of the Likert scale rating to drop items that do not fit (A1, A2, A3, A6, B1, B13, B18, B18, B2, B3, B4, B7 , C10, C12, C13, C14, C15, C17, C18, C20, C25, C26, C29, C30, C31, C5, C7, C9, D11, D12, D13, E12, E27, E7, F1, F3, and F6) including a person who provides an unexpected response (person-to-2, 5, 6.7, 14, 20, 22, 24 and 26). Finally, a purified instrument can be shaped to suit the parameters of good measurement.

REFERENCES

Abd. Ghafar Md. Din. (2003). *Prinsip dan Amalan Pengajaran*. Kuala Lumpur: Utusan Publications.

Baharudin Yaacob & Mohd. Yusof Abdullah. (2008). *Pencerapan pengajaran dan pembelajaran*. Pulau Pinang: Institut Perguruan Tuanku Bainun.

Chong, A. K. (1992). *Laporan pra instrumen Matematik guru pelatih ambilan 1992*. Kertas kerja Persidangan Kebangsaan Matematik/Institut Perguruan Malaysia. Melaka.

Chua, Y. P. (2006). Kaedah dan statistik penyelidikan: Kaedah Penyelidikan. Kuala Lumpur: Mc Graw Hill.

Cohen, L., Manion, L., & Morrison, K. (2007). *Research Methods in Education,* Sixth Edition. Oxon: Routledge.

Creswell, J. W. (2009). Research design: Qualitative, quantitative and mixed methods approachers. Los Angeles: SAGE Publications.

Creswell, J. W. & Plano Clark, V. L. (2011). *Designing and conducting mixed method research.* (2nd edition). Thousand Oaks, CA: Sage.

Creswell, J. W. (2012). Educational Research: Planning, Conducting and Evaluating Quantitative and Qualitative Research. (4th ed.). Boston, MA: Pearson Education

Daud Mohamad. (2000). *Pengetahuan Pedagogi Kandungan (PPK): Isu Bahasa Dalam Matematik*. Universiti Teknologi MARA: Seminar Pendidikan Sains dan Matematik.

Depdikna. (2005). Those who understand: Knowledge growth in teaching. Educational Researcher.

Fernandez, Clea & Yoshida, Makoto. (2004). *Lesson study*: A Japanese Approach to Improving Mathematics Teaching and Learning. Mahmah, New Jersey: Lawrence Erlbaum Associates, Publishers.Firestone, W.A., & Pennell, J. R. (1997). Designing state-sponsored teacher networks: A comparison of two cases. *American Educational Research Journal, 34*(2), pp. 237-266.

Fisher, W. P. Jr. (2007). Rating Scale Insturment Quality Criteria. *Rasch Measurement Transactions, 21*(1), p. 1095.

Goldhammer, R., Anderson, R.H., & Krajewski, R. J. (1981). *Clinical supervision*. New York: Holt, Rinehart & Winston, Inc.

Johnson, C. C. (2006). Effective professional development and change in practice: Barriers science teachers encounter and implications for reform. *School Science and Mathematics, 106*(3), pp. 150-161.

Joyce, C. R. B., McGee, H. M., & O'Boyle, C. A. (2002). Individual quality of life: review and outlook. In Joyce, C. R. B., O'Boyle, C. A. and McGee, H. (eds), *Individual Quality of Life: Approaches to Conceptualisation and Assessment.* Harwood Academic, Amsterdam, 215–24.

Kementerian Pelajaran Malaysia. (2009). *Standard Guru Malaysia.* Putrajaya: Bahagian Pendidikan Guru.

Lewis, C. (2009). What is the nature of knowledge development in lesson study? *Educational Action Research, 17*(1), pp. 95-110.

Linacre, J. M. (2011). 3PL, Rasch, Quality-Control And Science. *Rasch Measurement Transactions, 27*(4), pp. 1441-1444.

NKRA. (2009). Enam Bidang Keberhasilan Utama Negara. Putrajaya: KPM.

Noraini Idris. (2001). *Pedagogi dalam pendidikan Matematik*. Kuala Lumpur: Utusan Publications & Distributors Sdn. Bhd.

Noraini Idris. (2010). *Penyelidikan dalam Pendidikan*. Malaysia: Mc Graw Hill.

Noor Shah Saad. (2002). *Teori dan Perkaedahan Matematik Siri 1.* Petaling Jaya : Prentice Hall Pearson Malaysia Sdn Bhd.

Onwuegbuzie, A. J., Collins, K. M. T., & Leech, N. L. (2007a). *Mixed research: A step-by-step guide*. New York: Taylor & Francis.

Prahalad C. K., & Krishnan M. S. (2008). The New Age of Innovation: Driving Cocreated Value through Global Networks. US: McGraw Hill.

Roslee Talip et al. (2011). Penambahbaikan sekolah melalui Komuniti Pembelajaran Profesionalisme (KPP). *Retrieved on December 14, 2011, from http://eprints.ums.edu.my/456/*

Sato, M. (2003). Kyoshitachi no chosen (Challenge by teachers). Tokyo: Shogakkan.

Sato, M. (2006a). *Gakko no chosen (Challenge by schools)*. Tokyo: Shogakkan.

Sharifah Maimunah Syed Zin. (2011). *Pendekatan pengajaran dan pembelajaran Matematik KBSM*. Pusat Perkembangan Kurikulum, Kementerian Pendidikan Malaysia.

Shulman, L. S. (1986). Those who understand teach: Knowledge growth in teaching. *Educational Researcher, 15*(2), 4–14.

Shulman, L. S. (1987). Knowledge and teaching: Foundations of the new reform. *Harvard Educational Review, 57*(1), 1–22.

Shulman, L. S. (1991). *Pedagogical ways of knowing*. Keynote Address in 1990 ICET World Assmebly. Singapore, 27–31 Julai.

Sitti Haishah Abdul Rahman, Chiew Chee Mun & Norhaini Abdul Aziz. (2011). *Komuniti Pembelajaran Profesional (PLC)-Lesson study di Malaysia.* Perak: Bahagian Pendidikan Guru (BPG).

Solis, A. (2009). Pedagogical content knowledge: What matters most in the professional learning of content teachers in classrooms with diverse student populations. *Intercultural Development Research Association (IDRA) newsletter.* Retrieved on July 2012, from http://www.idra.org.

Teddlie, C., & Johnson, R. B. (2009). Methodological thought since the 20th century. In C. Teddlie & A. Tashakkori (Eds.), Foundations of mixed methods research: Integrating quantitative and qualitative techniques in the social and behavioral sciences. Thousand Oaks, CA: SAGE.

Wang-Iverson, Patsy & Yoshida, Makoto (Editors). (2005). Building Our Understanding of *Lesson study*. Philadelphia, PA: Research for Better Schools.