# *Thinking Aloud Protocol Based Self-Report Questionnaire to Measure Metacognitive Skills in Mathematical Problem Solving*

Uthpala Athukorala, Institute of Technology University of Moratuwa, Sri Lanka
Dileepa Fernando, Singapore University of Technology and Design, Singapore
Chanakya Wijeratne, University of Colombo, Sri Lanka

The Asian Conference on Education 2023
Official Conference Proceedings

## Abstract

Metacognitive skills play a major role in Mathematical problem solving. Metacognitive skills are required for monitoring and regulating the cognitive process of Mathematical problem solving. Different countries have declared that improving metacognitive skills is an essential component in Mathematics Education. Hence, having an instrument to effectively and efficiently measure metacognitive skills is important for both researchers and teachers. Think-aloud protocol is an endorsed method for assessing metacognitive skills in Mathematics. There, students verbalize their thoughts while working on the problem. However, this method has limited usability in large classroom settings due to the time consumed. Self-report questionnaires, on the other hand is an efficient metacognitive skill measurement instrument since it has ease of administration, suitable for larger classes and no need of special training on conducting. Though task general self-report questionnaires show low correlation with think-aloud which is an effective metacognition measurement tool, task specific questionnaires which were designed in line with think aloud show a significant correlation. To this date, there is no self-report questionnaire designed based on think-aloud for measuring metacognition in Mathematical problem solving. This study focuses on developing a task specific Likert type questionnaire for measuring metacognitive skills in Mathematical problem solving based on think aloud. The scale shows a high content validity (S-CVI/Ave=0.9), confirms the construct validity including both convergent and discriminant and higher internal consistency (ordinal alpha=0.89) assuring it as a successful measure for measuring metacognitive skills in Mathematical problem solving.

Keywords: Mathematical Problem Solving, Metacognition, Metacognitive Skills, Questionnaire, Thinking Aloud Protocol

## Introduction

Measuring metacognition is still debatable because it is a complex structure which is not visible from the outside. It is a process that happens inside the brain. There are currently various types of measurements used to assess metacognition. Questionnaires and interviews are offline measures while Think Aloud Protocol and systematical observations are online measures. This research specifically focuses on measuring the metacognition of students who learn in Digital Learning Environment (DLE). In DLEs, conducting Thinking Aloud or systematical observations may be impractical due to large number of students. These methods are individual-based and require more time for transcribing and analyzing students' language into a common coding system (Veenman & van Cleef, 2019). Different coding systems may create different results. Think Aloud Protocol has to be conducted by trained raters using well developed coding systems (Schellings et al., 2013). Quality of the online assessment depends on the adequacy of the coding system (Schellings et al., 2013). Further, there may be unrevealed thought processes even during think aloud tasks since students may not verbalize, all they think.

But many pieces of research confirm that online measures are the most effective method of assessing metacognition, especially in Mathematical problem solving (Veenman & van Cleef, 2019). Veenman and colleagues confirmed that self-report questionnaires exhibit a moderate relationship with the online Think Aloud method (Veenman & van Cleef, 2019). Veenman suggested that, for the assessment of metacognitive skills in Mathematics, online methods should be preferred over offline methods (Veenman & van Cleef, 2019). However, there are still numerous challenges in conducting the Thinking Aloud Protocol for larger classes. Schellings and colleagues concluded that the Thinking Aloud method is particularly suitable for research purposes rather than for practical aims due to its labor-intensive process (Schellings et al., 2013).

Hence, using the Think Aloud method for measuring metacognition in DLEs is not very practical due to large class sizes. However, there is still a possibility of creating a questionnaire based on the Thinking Aloud method. However, correlations between questionnaire data and think aloud measures are generally moderate to low (Schellings et al., 2013). When questionnaires are compared with Think Aloud, they present a varied picture. Researchers have found that general questionnaires exhibit a low correlation (0.22) with Think Aloud, whereas task specific questionnaires show moderately high (0.42) correlation (Schellings et al., 2013). Since questionnaires inquire about the activities that students performed, they rely on the long term memory. At times, questionnaires may not accurately represent the actual activities performed by students due to the limitations of memory (Schellings et al., 2013). Responses to the questionnaires may be influenced by varying reference points, such as comparing oneself with others, like the teacher or the best/worst student in the class. This also contributes to the low correspondence of the questionnaire. Even though students are using more strategic activities, they have to limit to the questionnaire given. Students report more activities to be effective not because they use them, but because they believe they are effective. Variations in rating the activities in questionnaires and Think Aloud create non-correspondence between the two. As an example, since Think Aloud uses a frequency scale, questionnaires measuring the usefulness of activities do not exhibit a high relationship (Schellings et al., 2013). Due to these reasons, self-report questionnaires and Thinking Aloud protocol suffer from low correlations.

Schellings and colleagues (2013) designed a questionnaire for measuring metacognition in reading activities based on taxonomy for coding Thinking Aloud protocols. Despite some validity issues with that scale, it showed a promising correlation (r=0.63) with Thinking Aloud. There is no such domain specific questionnaire built for Mathematical problem solving based on Thinking Aloud. In this present study, questionnaire was designed based on scoring system created by Veenman and colleagues (2000;2005) for systematical observations used in Thinking Aloud in Mathematical problem solving.

The objective of this research is to design a task-specific questionnaire aligned with thinking aloud in order to minimize administrative and time consuming issues while extracting maximum information comparable to an online measure. For Mathematical problem solving, no task specific questionnaire has been designed based on Thinking Aloud. If a task-specific questionnaire can be designed that closely measures metacognition skills, such as Thinking Aloud Protocol, this scale could be a successful alternative for Thinking Aloud Protocol.

Next sections will describe how the questionnaire was designed and how validity and reliability were tested for the designed questionnaire.

## Methodology

### Questionnaire Design

This questionnaire is aligned with systematical observations used in Thinking Aloud Protocol. Thinking Aloud Protocol is an online measure while questionnaire is an offline measure where students report what they do/have done. This self-report questionnaire contains the questions to measure the metacognitive skills in Mathematical problem solving.

Systematical observations during problem solving process were used to create the questions in the questionnaire (Veenman et al., 2005). That systematical observation process includes 15 activities which were used to evaluate students while they are thinking aloud. These 15 activities were designed and tested by Veenman (Jacobse & Harskamp, 2012; Veenman et al., 2000, 2005). Table no 1 shows the 16 questions designed in line with systematical observations in Thinking Aloud protocol.

From this self-report questionnaire, it is intended to measure how frequent, the student is applying metacognitive activities in Mathematical problem solving process. A three-point scale was used to score the items. A frequency scale was used, since Thinking Aloud is also measuring a frequency. The scale was same as the scale used by Schellings (Schellings et al., 2013). Respondents select an answer from three scales “almost never” (=1), “Sometimes” (=2) or “often” (=3). There are few reasons to select a three-point scale for the responses.

1. Due to the task-specific nature of the questionnaire's elements, it may be challenging for the respondents to distinguish between small distinctions between "often" and "very often" on a more complicated scale (Schellings et al., 2013).
2. Student’s perception on the task they performed, are represented in self-reports. When they select an answer, they may use some reference points (their own individual standard, view point of their teacher, standards related to an ideal student or poor student). Therefore students who use one reference point may have a consistent reference point (Schellings et al., 2013).

3. A three-point scale may reduce the variation among students' choices of a reference point even though it cannot be fully eliminated (Schellings et al., 2013). This will produce high reliability and stability in the questionnaire.

After the design process, next step was the questionnaire validation and finding the reliability to ensure that how well the data is representative of the subject under examination and how well it provides stable and consistent results (Taherdoost, 2016).

| | **Activities Recommended (Jacobse & Harskamp, 2012; Veenman et al., 2000, 2005)** | **Question Included to verify the activity** |
|---|---|---|
| 1 | entirely reading the problem statement (Planning) | 1. I read the question entirely, before I start the solving process. |
| 2 | selection of relevant data (Planning) | 2. I select/highlight all the relevant data from the question before starting the solving process. |
| 3 | paraphrasing of what was asked for(Planning) | 3. I paraphrase the question.<br>4. I make clear what I have to find before starting the solving process |
| 4 | making a drawing related to the problem (Planning) | 5. I usually draw a sketch related to the problem, before I start the solving process. |
| 5 | estimating a possible outcome (Planning) | 6. Before I solve the problem, I estimate/think about the nature of the possible solution that I would get. |
| 6 | designing an action plan before actually calculating (Planning) | 7. I usually design a plan (temporary) to solve the problem |
| 7 | systematically carrying out such plan (Monitoring) | 8.Every time I execute the designed plan systematically to reach the answer. |
| 8 | calculation correctness (Monitoring) | 9. I am always vigilant on the calculation process to verify that I am on the correct way to the solution. |
| 9 | avoiding negligent mistakes (Monitoring) | 10. I pay attention to avoid negligent mistakes during the solving process. |
| 10 | orderly note-taking of problem solving steps (Monitoring) | 11. I keep an eye on the problem solving steps which helps me to verify intermediate results. |
| 11 | monitoring the on-going process; (Monitoring) | 12. I always monitor the ongoing calculation process. |
| 12 | checking the answer (Evaluation) | 13. I check whether the final answer is acceptable and compatible with given data. |
| 13 | drawing a conclusion (Evaluation) | 14. I confirm that the final answer is correct. |

| | | |
|---|---|---|
| 14 | reflecting on the answer (Evaluation) | 15. I refer the final answer to the problem statement and verify that the answer is acceptable. |
| 15 | relating to earlier problems solved (Evaluation) | 16. I relate to similar problems solved earlier and reflect the accuracy of the answer. |

Table 1: Initial questions included in the questionnaire

### Questionnaire Validation and Reliability

To validate the questionnaire, content validation and construct validation were used. Internal consistency was calculated for reliability analysis.

### Sample Selection for Validation and Reliability

Students were classified according to their field of study(strata). To have 95% confidence interval with 5% margin of error, ideal sample size for 800 populations is 260. Table 2 represents how 260 students were selected proportionately from each discipline. Specific student from each discipline was selected randomly using a random number generator. In this sample 33% are females and 67% are males.

| Discipline (Strata) | Total Population | No of students from each strata |
|---|---|---|
| Chemical Engineering Technology | 50 | 16 |
| Civil Engineering Technology | 200 | 65 |
| Electronic Engineering Technology | 100 | 33 |
| Electrical Engineering Technology | 100 | 33 |
| Information Technology | 100 | 33 |
| Maritime Engineering Technology | 20 | 6 |
| Nautical Studies | 20 | 6 |
| Mechanical Engineering Technology | 100 | 33 |
| Polymer Engineering Technology | 50 | 16 |
| Textile Engineering Technology | 60 | 19 |
| **Total** | **800** | **260** |

Table 2: Sample Selection Details

### Content Validation

In general, content validity requires assessing a new survey instrument to make sure it has all the necessary items and omits any which are unimportant to a specific concept area (Taherdoost, 2016). In content validation, a survey is conducted to get the idea of the experts in the same field of research. Content validation questionnaire was distributed among one Advanced Level Mathematics Teacher, four Mathematics lecturers from higher education institutes (public and private) in Sri Lanka and two researchers from the field of Mathematics Education. According to the answers of those 7 experts, Content Validation Index (CVI) was calculated.

### Construct Validation

Construct validity refers to how well a concept, idea or behavior was operationalized into a working, operable reality (Taherdoost, 2016). To measure the construct validity, the questionnaire was distributed among selected 260 diploma students from 1st Semester who follow IS1104 Mathematics and Statistics in Institute of Technology University of Moratuwa. The questionnaire was distributed as soon as the students finished a Mathematical problem solving exercise. After removing incomplete responses, only 200(77%) responses were used for finding two forms of construct validity; convergent and discriminant.

### Convergent Validity

In convergent validity, it is studied that two measures of constructs that are theoretically related are, in fact related (Taherdoost, 2016). To validate the questionnaire, the Cognitive and Metacognitive Strategies section (30 questions) of the Motivated Strategies for Learning Questionnaire (MSLQ) was employed. Using Statistical Package for Social Sciences (SPSS) software, correlation was calculated between two questionnaires using Spearman's rho correlation analysis. If the newly designed questionnaire is convergent valid, it should exhibit a significant correlation with the MSLQ – Cognitive and Metacognitive Strategies questionnaire.

### Discriminant Validity

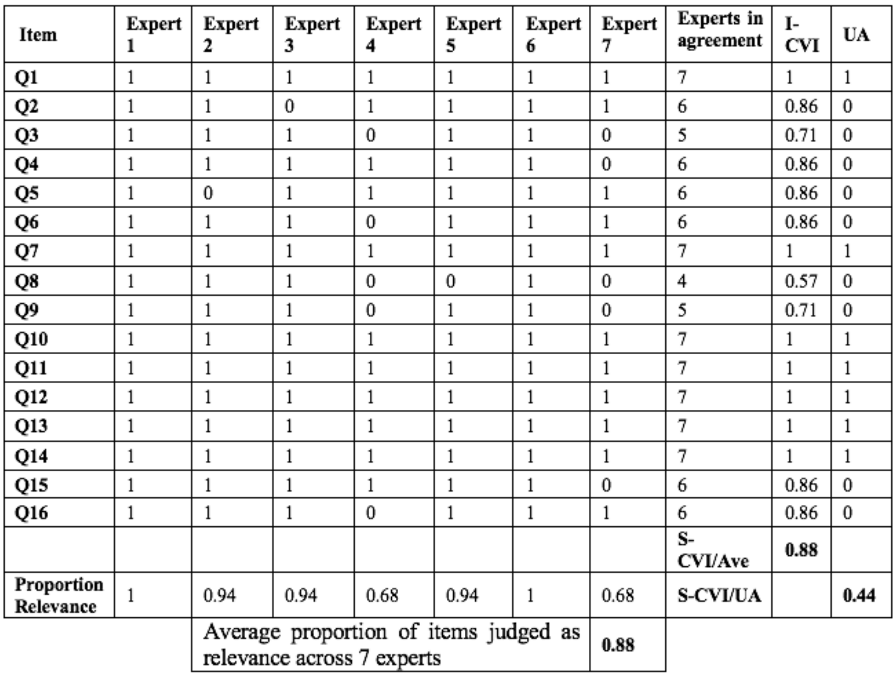
In discriminant validity, it tests whether constructs that are not related are in fact not related (Taherdoost, 2016). The questionnaire designed and the questions on Test Anxiety in the MSLQ, which are theoretically unrelated, were subjected to correlation testing using Spearman's rho correlation analysis. If the newly designed questionnaire has discriminant validity it should not show any significant relationship with responses from MSLQ-Test Anxiety questionnaire.

### Reliability Analysis

In the reliability analysis, internal consistency was assessed by calculating Cronbach's alpha from the data collected within the same sample.

## Results

### Content Validity



| Item | Expert 1 | Expert 2 | Expert 3 | Expert 4 | Expert 5 | Expert 6 | Expert 7 | Experts in agreement | I-CVI | UA |
|---|---|---|---|---|---|---|---|---|---|---|
| **Q1** | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 7 | 1 | 1 |
| **Q2** | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 6 | 0.86 | 0 |
| **Q3** | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 5 | 0.71 | 0 |
| **Q4** | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 6 | 0.86 | 0 |
| **Q5** | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 6 | 0.86 | 0 |
| **Q6** | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 6 | 0.86 | 0 |
| **Q7** | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 7 | 1 | 1 |
| **Q8** | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 4 | 0.57 | 0 |
| **Q9** | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 5 | 0.71 | 0 |
| **Q10** | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 7 | 1 | 1 |
| **Q11** | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 7 | 1 | 1 |
| **Q12** | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 7 | 1 | 1 |
| **Q13** | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 7 | 1 | 1 |
| **Q14** | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 7 | 1 | 1 |
| **Q15** | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 6 | 0.86 | 0 |
| **Q16** | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 6 | 0.86 | 0 |
| | | | | | | | | **S-CVI/Ave** | **0.88** | |
| **Proportion Relevance** | 1 | 0.94 | 0.94 | 0.68 | 0.94 | 1 | 0.68 | **S-CVI/UA** | | **0.44** |
| | | Average proportion of items judged as relevance across 7 experts | | | | | **0.88** | | | |

To measure the content validity of the questionnaire Content Validation Index (CVI) was calculated (Yusoff, 2019). Table 3 presents the calculation of CVI.

Table 3: Content Validation Index (CVI) Calculation

There three important CVI indices (Yusoff, 2019)

1. I-CVI (Item level content validity index)
2. S-CVI (scale-level content validity index based on the average method)
3. S-CVI/UA (scale-level content validity index based on the universal agreement method)

According to Lynn (Lynn, 1986; Yusoff, 2019), for 7 experts, minimum acceptable CVI value is 0.83. Except questions 3, 8 and 9 in designed questionnaire, all other questions are satisfying this minimum threshold value for I-CVI. S-CVI/Ave value is 0.88 and it also satisfy the minimum threshold value of 0.83. But S-CVI/UA is 0.44 and it does not satisfy the minimum requirement.

### Construct Validity

For this, Convergent validity and Discriminant validity was calculated using Spearman’s rho.

### Convergent Validity

Table 4 shows the results of correlation analysis for convergent validity.

**Correlations**

| | | | New Questionnaire | Metacognitive and Cognitive Component of MSLQ |
|---|---|---|---|---|
| Spearman's rho | New Questionnaire | Correlation Coefficient | 1.000 | .288$^{**}$ |
| | | Sig. (2-tailed) | . | .000 |
| | | N | 200 | 200 |
| | Metacognitive and Cognitive Component of MSLQ | Correlation Coefficient | .288$^{**}$ | 1.000 |
| | | Sig. (2-tailed) | .000 | . |
| | | N | 200 | 200 |

**. Correlation is significant at the 0.01 level (2-tailed).

Results show a significant correlation at the 0.01 level (2-tailed). Hence it can be concluded that new questionnaire shows a convergent validity with a same type of a questionnaire.

Table 4: Results for Convergent Validity

### Discriminant Validity

Table 5 shows the results of correlation analysis of discriminant validity. There is no any significant correlation between the new questionnaire and test anxiety component of MSLQ. Hence it doesn't show any relationship with Test Anxiety. These results ensure the discriminant validity.

**Correlations**

| | | | New Questionnaire | Test Anxiety component of MSLQ |
|---|---|---|---|---|
| Spearman's rho | New Questionnaire | Correlation Coefficient | 1.000 | .085 |
| | | Sig. (2-tailed) | . | .232 |
| | | N | 200 | 200 |
| | Test Anxiety component of MSLQ | Correlation Coefficient | .085 | 1.000 |
| | | Sig. (2-tailed) | .232 | . |
| | | N | 200 | 200 |

Table 5: Results of Discriminant Validity

⍰

### Reliability Analysis-Cronbach's Alpha

For evaluating the reliability, internal consistency was considered. For calculating internal consistency, Cronbach's alpha was calculated using SPSS (Table 6) (Schellings et al., 2013).

| Cronbach's Alpha | Cronbach's Alpha Based on Standardized Items | N of Items |
|---|---|---|
| 0.546 | 0.545 | 16 |

Table 6: Reliability Statistics

It seems that Cronbach's alpha (Table 6) is a low value (0.5 < α <0.6) and this will interpret low internal consistency in the newly designed questionnaire. The reasons for low Cronbach's alpha value should be investigated for improving the internal consistency and hence assuring that all items are measuring the same variable (metacognitive skills).

## How Questionnaire was Updated for High Validity and Reliability?

If the Cronbach's alpha is applied to a scale for calculating the reliability, three assumptions should be satisfied(McNeish, 2018). They are,

1. The scale items should be continuous and normally distributed.
2. The scale should adhere tau equivalence.
3. The errors of the items do not covary.

The observed covariance (or correlations) between items forms the foundation for a major part of Cronbach's alpha (McNeish, 2018). These item covariance is calculated using Pearson Correlation Analysis (PCA). It is well known that all variables in Pearson Correlation matrices are continuous in nature (McNeish, 2018). The scale described above is a Likert type questionnaire which contains discrete values. Hence, first assumption is violated.

By checking tau equivalence, it is assured that each item on the scale contributes equally to the total scale score (McNeish, 2018). To verify tau equivalence, exploratory factor analysis is run on the scale to verify that items have same relationship to underlying construct. For the scale developed above, exploratory factor analysis was run using SPSS.

## Factor Analysis

This questionnaire is designed to measure the metacognitive skills in Mathematical problem solving. Metacognitive skills are composed with three components; planning, monitoring and evaluation. Questions in the questionnaire are designed to measure these three components. The questions are organized as follows. Q1, Q2, Q3, Q4, Q5, Q6 and Q7are designed to measure planning skills, Q8, Q9, Q10, Q11 and Q12 are designed to measure monitoring skills and Q13, Q14, Q15 and Q16 are designed to measure evaluation skills. Table 7 presents the groups of questions that are initially assumed to be in three groups. Hence, it is assumed that factor analysis of this questionnaire should lie within three factors/groups.

| **Planning** |
|---|
| **PL1**.I read the question entirely, before I start the solving process. |
| **PL2**.I select/highlight all the relevant data from the question before starting the solving process. |
| **PL3**. I summarize the question and identify the main points. |
| **PL4**. I make clear what I have to find before starting the solving process |
| **PL5**. I usually draw a sketch related to the problem, before I start the solving process. |
| **PL6**. Before I solve the problem, I estimate/think about the nature of the possible solution that I would get. |
| **PL7**. I usually design a plan (temporary) to solve the problem |
| **Monitoring** |
| **MO1**.Every time I execute the designed plan systematically to reach the answer. |
| **MO2**. I am always vigilant on the calculation process to verify that I am on the correct way to the solution. |
| **MO3**. I pay attention to avoid negligent mistakes during the solving process. |
| **MO4**. I keep an eye on the problem solving steps which helps me to verify intermediate results. |
| **MO5**. I always monitor the ongoing calculation process. |
| **Evaluation** |
| **EV1**. I check whether the final answer is acceptable and compatible with given data. |
| **EV2**. I confirm that the final answer is correct. |
| **EV3**. I refer the final answer to the problem statement and verify that the answer is acceptable. |
| **EV4**. I relate to similar problems solved earlier and reflect the accuracy of the answer. |

Table 7: How questions are assumed to be in groups.

Factor analysis for above mentioned groups were conducted in SPSS. Exploratory Factor Analysis (EFA) was conducted using a principal component analysis and varimax rotation. Using EFA, items with high correlations are grouped together. Minimum factor loading was set to 0.5.

| | | |
|---|---|---|
| Kaiser-Meyer-Olkin Measure of Sampling Adequacy (KMO) | | .559 |
| Bartlett's Test of Sphericity | Approx. Chi-Square | 230.035 |
| | df | 120 |
| | Sig. | .000 |

Table 8: KMO and Bartlett's Test

Since KMO< 0.6, it indicates that it needs corrective actions (Shrestha, 2021). Bartlett's Test of Sphericity is highly significant at $p<0.001$ and it indicates that there are significant correlations among at least few variables in the matrix. It rejects the null hypothesis ($H_0$) that correlation matrix is an identity matrix. Since significant value is <0.05, variables are suitable for factor analysis (Shrestha, 2021).

Communality represents that the degree of variance a variable shares with all other variables being considered (Shrestha, 2021). For a sample size in between 100 and 200, communality value in between 0.5 and 0.6 is acceptable (Shrestha, 2021). Items which does not satisfy this requirement are removed. In Table 9, communality values of all 16 items in the questionnaire is represented. Question no 1 and 5 (MO1 and MO5) in Monitoring Skills group is not satisfying the criteria. Based on this, these two questions can be removed. But for further analysis, those two were kept as they are without removing in this very first step.

Table 10 demonstrates the eigenvalues and total variance. Extraction method used for factor analysis is principal component analysis. Before extraction, there were 16 factors. There are seven unique linear components in the data set with the eigenvalue > 1 after extraction and rotation. The portion of the total variance explained by a factor is indicated by its eigenvalue. The factors that have an eigenvalue greater than one are kept in factor analysis (Shrestha, 2021). The reasoning behind this rule makes sense. An eigenvalue larger than one is regarded as significant and denotes that the factor accounts for more of the common variance than the unique variance (Shrestha, 2021).

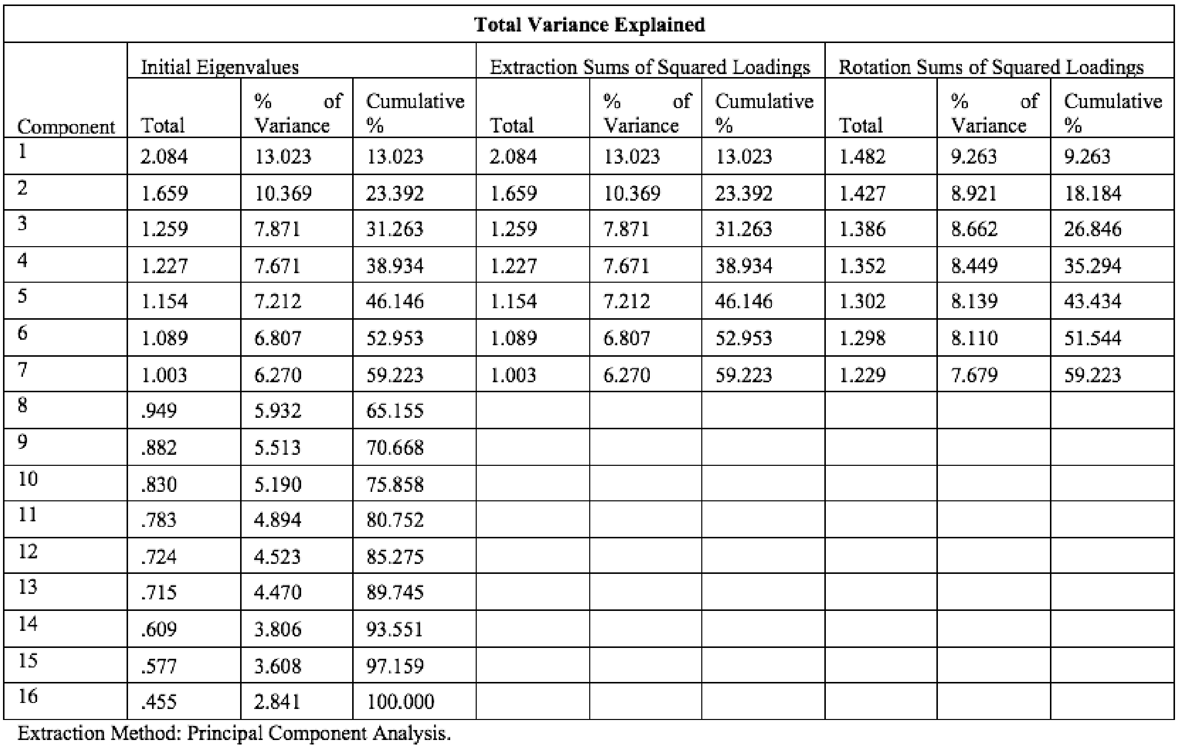
It is suggested that the retained components should account for at least 50% of the total variation. It reveals that 59.2% common variance shared by 16 variables is now sharing among seven variables. But this result is violating the initial decision of keeping all factors in three factor groups (planning, monitoring and evaluating). This is also indicated by initial KMO value (0.559) which indicated the need of corrective actions.

Table 11 represents the rotated component matrix with factor loading values for all seven factors. Factor loading values less than 0.5 are not displayed. Hence, items PL2, PL6, MO1 and MO5 does not include in any factor structure. Remaining items are scattered among seven factors and does not agree with three factor groups which was used for initial questionnaire design. Hence, factor analysis for this questionnaire is failed and it reveals the reason for low Cronbach’s alpha value which created low internal consistency in the questionnaire.

Variables in this questionnaire are ordinal (“Almost never”, “Sometimes”, “Often”) and discrete. It is recommended to use ordinal alpha for calculating internal consistency for ordinal type of data (Gadermann et al., 2012).

| **Communalities** | | |
|---|---|---|
| | Initial | Extraction |
| PL1 | 1.000 | .527 |
| PL2 | 1.000 | .646 |
| PL3 | 1.000 | .670 |
| PL4 | 1.000 | .598 |
| PL5 | 1.000 | .578 |
| PL6 | 1.000 | .600 |
| PL7 | 1.000 | .650 |
| MO1 | 1.000 | .498 |
| MO2 | 1.000 | .543 |
| MO3 | 1.000 | .612 |
| MO4 | 1.000 | .570 |
| MO5 | 1.000 | .371 |
| EV1 | 1.000 | .670 |
| EV2 | 1.000 | .679 |
| EV3 | 1.000 | .595 |
| EV4 | 1.000 | .669 |
| Extraction Method: Principal Component Analysis. | | |

Table 9: Communalities



**Total Variance Explained**

| Component | Initial Eigenvalues | | | Extraction Sums of Squared Loadings | | | Rotation Sums of Squared Loadings | | |
|---|---|---|---|---|---|---|---|---|---|
| | Total | % of Variance | Cumulative % | Total | % of Variance | Cumulative % | Total | % of Variance | Cumulative % |
| 1 | 2.084 | 13.023 | 13.023 | 2.084 | 13.023 | 13.023 | 1.482 | 9.263 | 9.263 |
| 2 | 1.659 | 10.369 | 23.392 | 1.659 | 10.369 | 23.392 | 1.427 | 8.921 | 18.184 |
| 3 | 1.259 | 7.871 | 31.263 | 1.259 | 7.871 | 31.263 | 1.386 | 8.662 | 26.846 |
| 4 | 1.227 | 7.671 | 38.934 | 1.227 | 7.671 | 38.934 | 1.352 | 8.449 | 35.294 |
| 5 | 1.154 | 7.212 | 46.146 | 1.154 | 7.212 | 46.146 | 1.302 | 8.139 | 43.434 |
| 6 | 1.089 | 6.807 | 52.953 | 1.089 | 6.807 | 52.953 | 1.298 | 8.110 | 51.544 |
| 7 | 1.003 | 6.270 | 59.223 | 1.003 | 6.270 | 59.223 | 1.229 | 7.679 | 59.223 |
| 8 | .949 | 5.932 | 65.155 | | | | | | |
| 9 | .882 | 5.513 | 70.668 | | | | | | |
| 10 | .830 | 5.190 | 75.858 | | | | | | |
| 11 | .783 | 4.894 | 80.752 | | | | | | |
| 12 | .724 | 4.523 | 85.275 | | | | | | |
| 13 | .715 | 4.470 | 89.745 | | | | | | |
| 14 | .609 | 3.806 | 93.551 | | | | | | |
| 15 | .577 | 3.608 | 97.159 | | | | | | |
| 16 | .455 | 2.841 | 100.000 | | | | | | |

Extraction Method: Principal Component Analysis.

Table 10: Eigenvalues and Total Variance Explained

According to the results obtained in factor analysis, questionnaire was updated by removing items PL2, PL6, MO1 and MO5. Then the questionnaire was distributed to the same sample again and calculated validity and reliability. This time ordinal alpha using R software was used to calculate internal consistency. The R code for calculating ordinal alpha is below.

```
>install. packages ("psych")              #polychoric correlation for ordinal data.
>require (psych)
>install. packages ("haven")              #loading SPSS files to R
>require (haven)
>data ← read_sav ("SPSS file path") #reading SPSS data file to R
>myfile ← polychoric (data)               #calculating polychoric correlation matrix
>alpha (myfile$rho)                       #calculating ordinal alpha
```

ExtractionMethod:PrincipalComponentAnalysis.
Rotation Method: Varimax with Kaiser Normalization.
a. Rotation converged in 9 iterations.

**Rotated Component Matrix**[a]

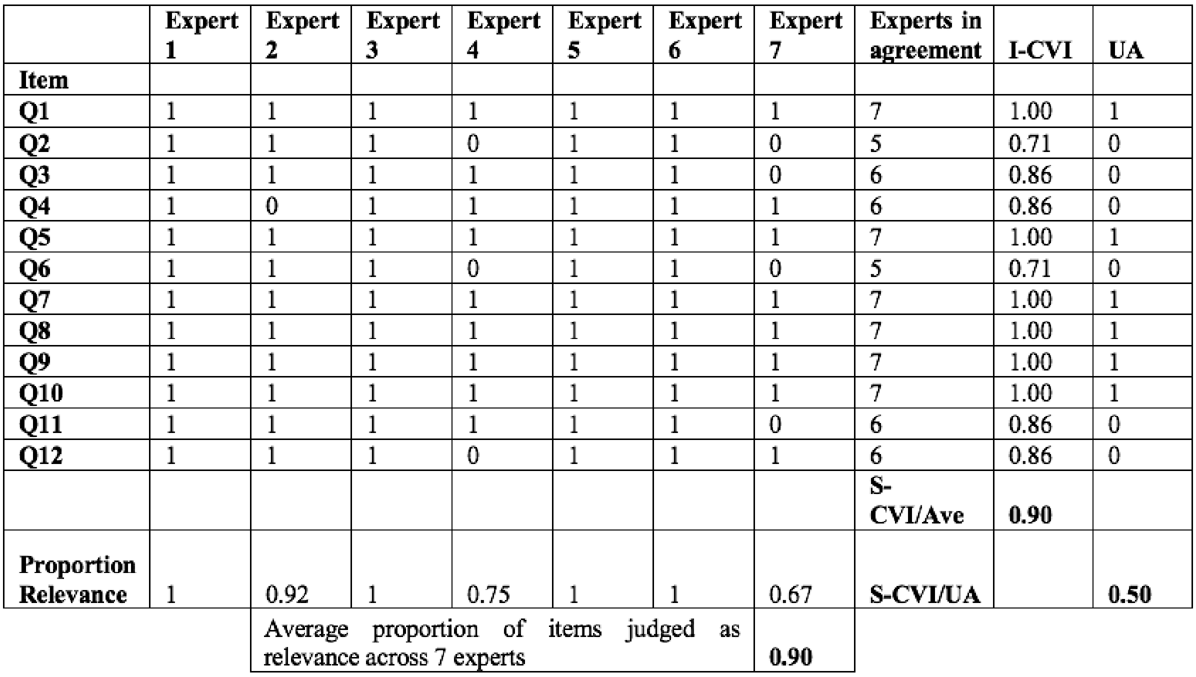
| | Component | | | | | | |
|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| PL1 | | | | .709 | | | |
| PL2 | | | | | | | |
| PL3 | | | | | .782 | | |
| PL4 | | | | .620 | | | |
| PL5 | | .727 | | | | | |
| PL6 | | | | | | | |
| PL7 | | .675 | | | | | |
| MO1 | | | | | | | |
| MO2 | | | .722 | | | | |
| MO3 | | | .524 | | | .565 | |
| MO4 | | | | | | .732 | |
| MO5 | | | | | | | |
| EV1 | .712 | | | | | | |
| EV2 | .794 | | | | | | |
| EV3 | | | | | | | .583 |
| EV4 | | | | | | | .780 |

Table 11: Rotated Component Matrix

Ordinal alpha calculated for the updated (12 questions) questionnaire was 0.89 (> 0.7). Content validity (S-CVI/Ave) was increased up to 0.9 and (S-CVI/UA) up to 0.50 by elevating the validity and the reliability of the questionnaire to an accepted level. Table 12 represents the updated content validity calculation.

## Discussion

Measuring metacognition is not an easy process as it involves a mental process that is not directly observable. The search for an optimum scale which measures metacognition successfully is still under discussion. Online methods, such as the Thinking Aloud Protocol (TAP), are preferred over offline methods to measure metacognition in Mathematics. (Veenman & van Cleef, 2019). But measuring metacognition using TAP is not feasible for large classes, especially in online settings. The self-report questionnaire, designed here based on Think Aloud, is suggested as a solution to practical issues in TAP. A teacher who wishes to measure metacognition in Mathematical problem-solving should share this questionnaire soon after the exercise to minimize memory loss.



| | Expert 1 | Expert 2 | Expert 3 | Expert 4 | Expert 5 | Expert 6 | Expert 7 | Experts in agreement | I-CVI | UA |
|---|---|---|---|---|---|---|---|---|---|---|
| **Item** | | | | | | | | | | |
| **Q1** | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 7 | 1.00 | 1 |
| **Q2** | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 5 | 0.71 | 0 |
| **Q3** | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 6 | 0.86 | 0 |
| **Q4** | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 6 | 0.86 | 0 |
| **Q5** | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 7 | 1.00 | 1 |
| **Q6** | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 5 | 0.71 | 0 |
| **Q7** | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 7 | 1.00 | 1 |
| **Q8** | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 7 | 1.00 | 1 |
| **Q9** | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 7 | 1.00 | 1 |
| **Q10** | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 7 | 1.00 | 1 |
| **Q11** | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 6 | 0.86 | 0 |
| **Q12** | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 6 | 0.86 | 0 |
| | | | | | | | | **S-CVI/Ave** | **0.90** | |
| **Proportion Relevance** | 1 | 0.92 | 1 | 0.75 | 1 | 1 | 0.67 | **S-CVI/UA** | | **0.50** |
| | | Average proportion of items judged as relevance across 7 experts | | | | | **0.90** | | | |

Table 12: Content Validity Calculation for Updated Questionnaire

This questionnaire initially contained 16 questions, corresponding to the 15 activities suggested in systematical observations by Veenman (2019) and colleagues. Despite confirming validity (content and construct), it exhibited very low internal consistency. First, Cronbach's alpha was used to measure internal consistency. However, it was later understood that Cronbach's alpha is not recommended for a scale with ordinal values, as it is associated with continuous values. (Gadermann et al., 2012). Questions with low factor loadings, which did not contribute at the same level to measuring metacognitive skills in Mathematical problem-solving, were identified after conducting an Exploratory Factor Analysis (EFA). After removing those questions, ordinal alpha was calculated to measure internal consistency, resulting in a value of 0.89, demonstrating high reliability. The removal of these questions also increased content validity to 0.9.

## Conclusion

The questionnaire designed in this paper provides a solution for measuring metacognitive skills in Mathematical problem-solving for large classes. This is a viable option for online classes with a large number of participating students, where applying TAP is not feasible due to the additional time and effort it requires. High validity and reliability of the newly designed questionnaire confirms that it is an effective and efficient scale for measuring metacognitive skills. As the next step, the intention is to study the relationship between the Think Aloud Protocol and the designed questionnaire to understand the extent to which it aligns with TAP.

## References

Gadermann, A. M., Guhn, M., & Zumbo, B. D. (2012). *Estimating ordinal reliability for Likert-type and ordinal item response data: A conceptual, empirical, and practical guide*. https://doi.org/10.7275/N560-J767

Jacobse, A. E., & Harskamp, E. G. (2012). Towards efficient measurement of metacognition in mathematical problem solving. *Metacognition and Learning*, *7*(2), 133–149. https://doi.org/10.1007/s11409-012-9088-x

Lynn, M. (1986). Determination and Quantification of Content Validity. *Nursing Research*, *35*(6), 382–386.

McNeish, D. (2018). Thanks coefficient alpha, we'll take it from here. *Psychological Methods*, *23*(3), 412–433. https://doi.org/10.1037/met0000144

Schellings, G. L. M., Veenman, M. V. J., Van Hout-Wolters, G., & Meijer J. (2013). Assessing metacognitive activities: The in-depth comparison of a task-specific questionnaire with think-aloud protocols. *European Journal of Psychology of Education*. https://doi.org/10.1007/s10212-012-0149-y

Shrestha, N. (2021). Factor Analysis as a Tool for Survey Analysis. *American Journal of Applied Mathematics and Statistics*, *9*(1), 4–11. https://doi.org/10.12691/ajams-9-1-2

Taherdoost, H. (2016). Validity and Reliability of the Research Instrument; How to Test the Validation of a Questionnaire/Survey in a Research. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3205040

Veenman, M. V. J., Kerseboom, L., & Imthorn, C. (2000). Test anxiety and metacognitive skillfulness: Availability versus production deficiencies. *Anxiety, Stress & Coping*, *13*(4), 391–412. https://doi.org/10.1080/10615800008248343

Veenman, M. V. J., Kok, R., & Blöte, A. W. (2005). The relation between intellectual and metacognitive skills in early adolescence. *Instructional Science*, *33*(3), 193–211. https://doi.org/10.1007/s11251-004-2274-8

Veenman, M. V. J., & van Cleef, D. (2019). Measuring metacognitive skills for mathematics: Students' self-reports versus on-line assessment methods. *ZDM*, *51*(4), 691–701. https://doi.org/10.1007/s11858-018-1006-5

Yusoff, M. S. B. (2019). ABC of Content Validation and Content Validity Index Calculation. *Education in Medicine Journal*, *11*(2), 49–54. https://doi.org/10.21315/eimj2019.11.2.6

**Contact email:** uthpalap@itum.mrt.ac.lk