Assessing the validity and reliability of questionnaires on the implementation of Indonesian curriculum K-13 in STEM education

R Oktavia, Irwandi, Rajibussalim, M Mentari, and I S Mulia
Syiah Kuala University, Jl. Teuku Nyak Arief Darussalam, Banda Aceh, 23111, Indonesia
E-mail: rini_oktavia@unsyiah.ac.id

Abstract. This study aims to investigate how educators of STEM-related subjects implemented K-13 in their classrooms. A forum group discussion (FGD) on the implementation of Indonesian curriculum known as Kurikulum 2013 (K-13) in STEM (Science, Technology, Engineering, and Mathematics) Education was held in March 2018 by STEM Research Center Unsyiah (STEM.id). Participants of the FGD consist of 60 educators of STEM-related subjects in Banda Aceh and Aceh Besar Districts. Since K-13 was implemented in July 2013, there is no survey study to evaluate its implementation in Aceh has been conducted. This study initiates to develop survey instruments to measure participants’ understanding and opinions regarding the implementation of K-13 in STEM Education and their satisfaction of the FGD. Eight items in participants’ satisfaction questionnaire are strongly valid with validity values range from 0.632 to 0.824. For the questionnaires on the implementation of K-13, nine items are valid with validity values range from 0.584 to 0.821. For internal consistency reliability, the Cronbach’s alpha of the FGD satisfaction responses is 0.838; meanwhile the Cronbach’s alpha of implementation of K-13 responses is 0.882. These results showed that the questionnaires have good internal consistency reliability and validity.

1. Introduction

STEM Research Center Unsyiah (STEM.id) is a center in Syiah Kuala University that initiate the activities to advocate the development of integrated instructions of Science, Technology, Engineering, and Mathematics (STEM) so that the learning processes took place in STEM-related subject classrooms become more interesting and enjoyable for students. The activities include conducting surveys and Focus Group Discussion (FGD) to gather information on how to integrate Investigative Science Learning Environment (ISLE)-Based STEM lessons in K-13, developing ISLE-Based STEM learning modules, conducting workshops for teachers and students on ISLE-Based STEM lessons, developing websites that provides resources to implement ISLE-Based STEM modules, summer camps for students etc. The slogan “learning is fun” becomes a goal in developing integrated instructional approaches for STEM-related subjects. As one of the programs in its current project titled “Integrated ISLE in Integrated Science Instructions to Improve Science Teachers’ Abilities in Science Education” funded by the United States of America (USAID-PEER) Awards, STEM.id held an FGD on the implementation of Kurikulum 2013 (K-13) in STEM Education on March 2018 that was participated by 60 STEM-related educators and educational policy makers from the city of Banda Aceh and Aceh Besar District.
The FGD aims to investigate how well educators of STEM-related subject understand and are able to implement the standards mandated by Indonesian national school curriculum known as Kurikulum 2013 or K-13. More importantly, it is intended to investigate the challenges that the secondary school teachers have in implementing K-13 in STEM related subjects such as Physics, Chemistry, Biology, and Mathematics. Since ISLE was developed for Physics Lessons by Prof. Dr. Eugenia Etkina [1], the USG-Partner of the USAID-PEER project, some of the information gathered will specifically relate to Physics courses.

Since it was implemented in July 2013, there is no survey study to evaluate the implementation of K-13 in Aceh has been conducted. Therefore, as one of the data collection techniques applied in this study, surveys on the implementation of K-13 in STEM education and on the satisfaction of participants of the FGD were administered at the end of the FGD. The survey instruments were developed to measure participants’ understanding and opinions regarding the implementation of K-13 in STEM Education and their satisfaction of the FGD. In this article, the results of classical test theory (CTT) analyses on the validity and reliability of the instruments will be reported.

2. Literature Review
The psychometric analyses reported in this article only involve the classical test theory (CTT) analyses. CTT is based on the underlying principle that the scores that respondents got on responding to an instrument (observed scores) is the sum of the true scores and some errors:

\[ X = T + E \]

where \( X \) is the observed score; \( T \) is the true score; and \( E \) is the error [2]. Two common measures of the quality of instrument in CTT are validity and reliability. A discussion regarding these two measures is described in the following paragraphs.

The importance of validity and reliability as criteria of the quality of an instrument has been addressed in many studies [3,4,5]. The term validity denotes the degree to which an instrument measures what it purports to measure. The validity of an instrument specifies how well the instrument met the standards that were set by certain criteria to judge the quality of an instrument. An instrument that has a poor validity indicates that there are items in the instrument that do not measure what the instrument intends to measure. In other words, the validity of an instrument represents how accurate the instrument measure what it is intended to measure. There are several types of validity of a measurement, including face validity, content validity, and construct validity. Descriptions of these types of validity are presented below.

2.1. Face validity
Face validity represents the estimation regarding the clarity of the purpose of an instrument. It means a questionnaire has a high face validity if its purpose or construct is clear by the respondents taking it. Researchers will let some experts to review whether the instrument has met the criteria of measuring the traits purportedly measured.

2.2. Content validity
Content validity represents the estimation regarding the propriety of content and how it is presented. In this study, content validity is ensured by establishing the connections between items and the implementations of K-13 in science lessons and the effectiveness of the FGD. Using Dillman’s Tailored Design Method [6], the questionnaire blueprint was created. The content validity can be estimated by asking experts in the content area to review the blueprint and the items in the instrument. In this study, we asked the experts of K-13 and STEM education to review the contents regarding the implementation of K-13 in STEM education. These experts were given the list of content areas specified in the test blueprint along with the test items. The experts were then asked to indicate whether they agree that each item is appropriately matched to the contents indicated in the blueprint and whether the ways the items were represented purportedly measure the intended contents.
2.3. Construct validity

Construct validity is the measure of the propriety of the constructs in the instrument. According to Westen & Rosenthal [7], despite its strongly important measure of the quality of an instrument, there is no simple formula to represents the extent that an instrument can be considered as construct valid. The most common calculation used by researchers to obtain the construct validity metrics is by presenting correlations between a measure of a construct and a number of other measures that are theoretically associated with the construct known as the convergent validity or vary independently of the construct known as the discriminant validity [7]. In this article, the construct validity of the instruments used in our survey is described by calculating the Pearson’s correlation coefficient \( r \) of the scores of respondents’ responses to an item with their total scores. Pearson’s correlation coefficient is also known as Pearson’s product-moment correlation coefficient (PPMCC) and calculated as below:

\[
r = \frac{n \sum xy - \sum x \sum y}{\sqrt[n]{\sum x^2 - (\sum x)^2}[\sqrt[n]{\sum y^2 - (\sum y)^2}]}
\]

where \( r \) is the Pearson’s correlation coefficient, \( n \) is the number of valid responses, \( x \) represents the score of an item, and \( y \) represents the total score of each respondent with valid responses; with an assumption that both variables (\( x \) and \( y \)) are normally distributed. The criteria of interpreting a validity coefficient of an item are presented in Table 1.

| Validity coefficient values | Interpretation               |
|-----------------------------|------------------------------|
| Above 0.35                  | Very beneficial (Strongly Valid) |
| 0.21-0.35                   | Likely to be useful          |
| 0.11-0.20                   | Depends on circumstances     |
| Below 0.11                  | Unlikely to be useful        |

If validity of a measurement represents the accuracy of the measurement, the reliability of an instrument indicates the consistency of the scores acquired from the instrument. The term reliability was first introduced by Spearman in 1904 who defined it as the ratio between the variances of the true scores with the variance of the observed scores [9]. By estimating the reliability of an instrument, several criteria of consistency of the scores can be explained. There are different ways to estimate the reliability of a questionnaire including: (1) Test-Retest reliability that is estimated by calculating the correlations between scores of two or more administrations of the questionnaire with the same participants; (2) Parallel-Forms reliability that is estimated by creating two or more forms of questionnaire that have the same content and then administering the forms to the same subjects at the same time; (3) Inter-Rater reliability that is estimated by comparing the scores of two or more independent judges on the responses of the participants to the instrument; the comparison is used to determine whether the raters’ estimates are consistent; and (4) Internal Consistency Reliability that is estimated to judge the consistency of results across items on the same test. There are different perspectives regarding the internal consistency reliability as explained by Tang, Cui and Babenko [9]. Without ignoring the other interpretations of the internal consistency reliability, in this CTT analyses the definition of internal consistency reliability by Haertel [10] will be used, that is the reliability estimate of an instrument based on a single administration.

According to Tang, Cui and Babenko [9] there are several measures of internal consistency reliability under CTT including Cronbach’s alpha, Revelle’s beta, McDonald’s \( \omega_h \), and Sijtsma’s ECV. Among those measures, the Cronbach’s alpha is the most common measure used. In our analysis, the Cronbach’s alpha coefficient will be used as the measure of the internal consistency reliability of the instrument. The mathematical expression of the coefficient alpha introduced by Cronbach’s [11] is presented below:
\[
\alpha = \left[ \frac{n}{n-1} \right] \left[ 1 - \frac{\sum_{i=1}^{n} \sigma_i^2}{\sigma_X^2} \right] \tag{3}
\]

where \( \alpha \) is a lower-bound estimate of the true reliability, \( n \) is the number of items in test \( X \), \( \sigma_X^2 \) is the observed score variance of test \( X \), and \( \sigma_i^2 \) is the variance of item \( i \). The criteria of interpreting an internal consistency reliability coefficient of an instrument are presented in Table 2.

| Internal consistency reliability coefficient value | Interpretation     |
|--------------------------------------------------|--------------------|
| Greater than or equal to 0.90                    | Excellent          |
| 0.80-0.90                                        | Good               |
| 0.70-0.79                                        | Adequate           |
| Below 0.7                                        | Less Applicable    |

3. Method
The questionnaire for the survey administered during the FGD was developed based on the blueprint created to determine the constructs and contents to be measured. There are two constructs measure in this survey: (1) the implementation of K-13 (K-13 construct), and (2) the effectiveness of the FGD to raise concerns regarding STEM education and ISLE (FGD construct). There are 10 items developed for each construct that measure contents including: (1) How well the participants understand and implement K-13 in their lessons; (2) Whether the support system is available to implement K-13 for their lessons; (3) How easy the implementation of K-13 to be executed; (4) How they understand STEM education; (4) How they integrate STEM in K-13 based lessons; (5) How they understand ISLE; (6) How they think about using ISLE in physics lessons; and (7) How effective the FGD enlighten their understanding about STEM and ISLE. To ensure the content validity and the face validity of the instrument, we requested two experts in STEM education and ISLE to evaluate and provide critiques and suggestions to improve the instrument. After being reviewed, the revised version of the instrument was evaluated by the experts. The final version is the one that has been declared to be construct and content valid.

The participants of this study were 32 teachers and STEM educators attending the FGD held by STEM.id in March 2018. The survey was administered during the FGD. The administration took place at the end of the discussions session of the FGD. Each item in the questionnaire is given five pre-coded responses (Likert’s scale): 1 for strongly disagree, 2 for disagree, 3 for neutral, 4 for agree, and 5 for strongly agree. Items that measured the same construct were arranged in the same group of items. Therefore there are two groups of items that later will be analyzed separately. Some examples of the items (translated into English from Bahasa Indonesia) are presented in Table 3 below.

| Item   | Statement                                                                 |
|--------|---------------------------------------------------------------------------|
| FGD_3  | Before attending this FGD, I have known about STEM                        |
| FGD_6  | I have known about ISLE learning approach before attending this FGD       |
| K13_   | I think the available laboratory equipment has supported me in conducting  |
|        | practicum activities                                                       |
| K13_9  | I have mastered the use of the available laboratory equipment             |

Before analyzing the internal consistency validity and reliability, a descriptive analysis was conducted, including the normality assessment of the variables. Since the sample size is smaller than 50, the Shapiro-Wilk Test is more appropriate [12]. The results of CTT analysis, especially the internal consistency reliability and validity of our instrument will be discussed below.
4. Results and Analysis

From the distribution of the scores of each item and the total scores, we found that the average total score is 36.41 with a standard deviation of 6.257 for FGD construct and the average total score is 35.56 with a standard deviation of 8.758 for K-13 construct. It was also found that a little bit more than half of the students (56.2 %) scored 38-45 (maximum of 50) for FGD construct; meanwhile 54.1% of the participants scored 39-47 (maximum of 50) for K-13 construct. In evaluating the tendency of respondents in the survey the average responses of each item ranges from 2.7813 to 4.1875. These results indicate that most participants tend to agree with the items. Only two items fall in the neutral category (item FGD_3 and FGD_6). To calculate the internal consistency validity of the items in the instrument, the Shapiro-Wilk test statistics is presented in Table 4 and Table 5 below.

| Item    | Stat | df | Sig |
|---------|------|----|-----|
| FGD_1   | .854 | 26 | .002|
| FGD_2   | .839 | 26 | .001|
| FGD_3   | .885 | 26 | .007|
| FGD_4   | .828 | 26 | .001|
| FGD_5   | .710 | 26 | .000|
| FGD_6   | .828 | 26 | .001|
| FGD_7   | .815 | 26 | .000|
| FGD_8   | .671 | 26 | .000|
| FGD_9   | .737 | 26 | .000|
| FGD_10  | .624 | 26 | .000|

| Item    | Stat | df | Sig |
|---------|------|----|-----|
| FGD_1   | .854 | 26 | .002|
| FGD_2   | .839 | 26 | .001|
| FGD_3   | .885 | 26 | .007|
| FGD_4   | .828 | 26 | .001|
| FGD_5   | .710 | 26 | .000|
| FGD_6   | .828 | 26 | .001|
| FGD_7   | .815 | 26 | .000|
| FGD_8   | .671 | 26 | .000|
| FGD_9   | .737 | 26 | .000|
| FGD_10  | .624 | 26 | .000|

The Saphiro-Wilk statistics of the items are all above 0.5 with p-values less than 0.05. The results indicate that all variables are normally distributed, therefore the use of Pearson’s correlation coefficients are appropriate to measure the internal consistency construct validity. The values of the Pearson’s correlation coefficients of each item with their total scores are presented in Table 6 and Table 8.

From Table 6 it can be seen that there are two items that fail to be valid items. Analyzing the items, we found that these two items have different nature than the other FGD items. When we exclude these two items from the analysis, the results still supported that the other items have strong validity values. Reviewing the statements of the items, we found that these two items measure respondents’ knowledge about STEM and ISLE, which are new terms for the respondents; meanwhile other items purely asked their opinion about the information that they obtain from the FGD. The Pearson’s correlation coefficients and the p-values showed that respondents, who tend to agree with the other statements on the questionnaire, disagree with the statements on these two items. To fix this problem, we have revised the items as showed in Table 7.
Table 6. Internal consistency validity of FGD items

| Item   | $r$  | Significance | Interpretation       |
|--------|------|--------------|----------------------|
| FGD_1  | .726 | .000         | Strongly Valid       |
| FGD_2  | .759 | .000         | Strongly Valid       |
| FGD_3  | .081 | .665         | Unlikely to be useful|
| FGD_4  | .824 | .000         | Strongly Valid       |
| FGD_5  | .781 | .000         | Strongly Valid       |
| FGD_6  | .264 | .144         | Likely to be useful  |
| FGD_7  | .632 | .000         | Strongly Valid       |
| FGD_8  | .859 | .000         | Strongly Valid       |
| FGD_9  | .750 | .000         | Strongly Valid       |
| FGD_10 | .788 | .000         | Strongly Valid       |

Table 7. Revised items in for FGD construct

| Revised Item | Statement                                                                 |
|--------------|-----------------------------------------------------------------------------|
| FGD_3        | After attending this event, I have a better understanding about STEM learning approach |
| FGD_6        | After attending this event, I have understood ISLE learning model            |

Table 8. Internal consistency validity of K-13 items

| Item   | $r$  | Significance | Interpretation       |
|--------|------|--------------|----------------------|
| K13_1  | .720 | .000         | Strongly Valid       |
| K13_2  | .652 | .000         | Strongly Valid       |
| K13_3  | .584 | .000         | Strongly Valid       |
| K13_4  | .794 | .000         | Strongly Valid       |
| K13_5  | .780 | .000         | Strongly Valid       |
| K13_6  | .727 | .000         | Strongly Valid       |
| K13_7  | .821 | .000         | Strongly Valid       |
| K13_8  | .790 | .000         | Strongly Valid       |
| K13_9  | .441 | .017         | Strongly Valid       |
| K13_10 | .719 | .000         | Strongly Valid       |

Since all items in K-13 construct are strongly valid following the rule of thumb in Table 1, then there are no revisions applied to the items in K-13 construct. For internal consistency reliability, it was found that the Cronbach’s alpha of the FGD instrument is 0.838; meanwhile for the K-13 instrument, the response data have Cronbach’s alpha of 0.882. If we combine all responses of the 20 items, we found that the Cronbach’s alpha coefficient of the instrument is 0.848. Based on the interpretation of internal consistency reliability value in Table 8, from these results, we can confirm that the internal consistency reliability of the instrument is good.

5. Conclusions
Based on the CTT analysis conducted in this study, there are two items that are not strongly valid and need to be revised. The internal consistency coefficients of validity and reliability of the questionnaire showed that the instrument has good quality as a measurement tool to measure participants’ knowledge and opinion regarding the implementation of K-13 in STEM education and the effectiveness of the FGD in enlightening participants about STEM and ISLE. The revised survey instrument will be administered in the future survey that will be conducted a few weeks after this report was written. The comparison among the initial and revised instruments will be reported in our next article.
Acknowledgment
This study is funded entirely by the NAS and USAID under the USAID-PEER Grant with USAID Prime Award Number AID-OAA-A-11-00012, and that any opinions, findings, conclusions, and recommendations stated in the article are from the authors only, and do not always reflect the views of USAID or NAS.

References:
[1] Etkina E 2015 Millikan award lecture: Students of physics—Listeners, observers, or collaborative participants in physics scientific practices? American Journal of Physics 83 669
[2] Kline T J 2005 Psychological testing: A Practical Approach to Design and Evaluation (Canada: Sage)
[3] Cohen L, Manion L and Morrison K 2011 Research Methods in Education (7th ed) (London: Routledge)
[4] Cook D A and Beckman T J 2006 Current concepts in validity and reliability for psychometric instruments: Theory and application The American Journal of Medicine 119 166
[5] Moskal B M and Leydens J A 2000 Scoring rubric development: Validity and reliability Practical Assessment, Research and Evaluation 7 71
[6] Dillman D A, Smyth J D and Christian L M 2014 Internet, Phone, Mail, and Mixed-mode Surveys: The Tailored Design Method (4th ed) (Hoboken, NJ: John Wiley and Sons)
[7] Westen D and Rosenthal R 2003 Quantifying construct validity: Two simple measures Journal of Personality and Social Psychology 84 608
[8] Saad S, Carter G W, Rothenberg M and Israelson E 1999 Testing and Assessment: An Employer’s Guide to Best Practices (Washington DC: Employment and Training Administration)
[9] Tang W, Cui Y and Babenko O 2014 Internal consistency: Do we really know what it is and how to assess it Journal of Psychology and Behavioral Science 2 205
[10] Haertel E H 2013 Reliability and Validity of Inferences about Teachers based on Student Test Scores (Princeton, NJ: Educational Testing Service)
[11] Cronbach L J 1951 Coefficient Alpha and The Internal Structure of Tests Psychometrika 16 297
[12] Hanusz Z, Tarasinzka J and Zielinski W 2012 Adaptation of Shapiro-Wilk test to the case of known mean Colloquium Biometricum 42 43