Jurnal Penelitian dan Evaluasi Pendidikan
Volume 24, No 1, June 2020 (76-87)
Online: http://journal.uny.ac.id/index.php/jpep

# OBSERVATION INSTRUMENT FOR STUDENT SOCIAL ATTITUDE IN PRIMARY SCHOOLS: VALIDITY AND RELIABILITY

**Ari Setiawan[1], Widowati Pusporini[1*], Hanandyo Dardjito[2]**
[1]Department of Educational Research and Evaluation, Universitas Sarjanawiyata Tamansiswa
Jl. Kusumanegara No. 157, Yogyakarta 55165, Indonesia
[2]Department of English Education, Universitas Sarjanawiyata Tamansiswa
Jl. Batikan UH-III/1043, Yogyakarta 55167, Indonesia
*Corresponding Author. E-mail: w.pusporini@ustjogja.ac.id

## ABSTRACT

This study aims to (1) identify the content validity of observations of students' social attitudes, (2) identify the construct validity of the observation instruments of students 'social attitudes, and (3) identify the reliability of observations of students' social attitudes. The subjects of this study were grades IV and V of elementary school students in Yogyakarta province selected using cluster random sampling. Observation guidelines were used to collect the data using a summative rating scale model. The content validity was analyzed by applying Aiken assisted by Microsoft Excel, the construct validity by using second-order Confirmatory Factor Analysis assisted by Lisrel, and the reliability by using the Omega reliability approach. The results indicate that all items are valid by which the content validity. The construct validity with the Confirmatory Factor Analysis is high. The reliability values of the observation instruments are reliable.

**Keywords:** *validity, reliability, observation instrument*

**How to cite:** Setiawan, A., Pusporini, W., & Dardjito, H. (2020). Observation instrument for student social attitude in primary schools: Validity and reliability. *Jurnal Penelitian dan Evaluasi Pendidikan, 24*(1), 76-87. doi:https://doi.org/10.21831/pep.v24i1.31868

## INTRODUCTION

Social attitude is a core education aiming to create generation attitude. The social attitude, unfortunately, lacks assessment as the teachers' limited time. Teachers are more likely to spend time teaching without seeing the importance of doing the right assessment. Stiggins (2005) suggests that teachers should spend as much as one-third to half of their available time to engage in assessment activities.

Currently, social attitude assessment instruments that are developed or used by teachers in elementary school in the province of Yogyakarta, are still partial. Additionally, observation is rarely used because it takes a long time and challenging. Moreover, the basic observation instruments are still very rare. Regardless of the challenges, the instruments for social attitudes with observation is essential to estimate the real conditions. Availability of standardized, valid, and reliable social attitude observation instruments for elementary school students are required to write the report of the learning outcomes.

Social attitude can be seen as something associated to the attitude which is related to social conditions (Setiawan & Suardiman, 2018). Student social attitude can be recognised by observing the student' consistent (repetitive) behaviour, for example, consistently being late for class. Based on this visible observation, it can be interpreted that the student is not disciplined to come to the class. It is very reasonable to recognised students' attitude toward some-

ISSN: 2685-7111 (print)
2338-6061 (online)

https://doi.org/10.21831/pep.v24i1.31868

thing from visible behaviour because behaviour is an indicator of individual attitudes. Observation on student behaviour, however, is not enough for measuring a student's attitude because certain behaviours/actions are sometimes intentionally raised to disguise the real attitude (Azwar, 2014). For instance, students appear to work on their assignments for their teacher supervises them. The students observed actions do not necessarily mean the students' attitude of responsibility toward the assignment given by the teacher. The observed actions may indicate the students' respect to or afraid of their teacher.

An observational assessment might be defined as a form of the assessment carried out continuously using the senses, both directly and indirectly, using an observation format containing some indicators of the observed behaviour. The observation is conducted either during the learning inside or outside a classroom. Observation technique is a data collection technique that performed through direct observation of objects. This technique allows for measuring or assessing the social attitudes in everyday life.

As an assessment technique, observation might apply in various situations. This statement is in line with the concept of authentic appraisal where it carries out in real conditions and does not wait for the completion of a process. The model assessment might be conducted continuously for a long time to obtain a more accurate assessment. It is necessary to make a rating scale observation guide aiming to ensure consistency. This observation technique has several advantages such as (1) this technique is conducted in real and accurate situations; (2) the response can be directly assessed; and, (3) it needs to understand that this observation reflects latent social attitudes.

Validity indicates that a test is essentially valid as long as it detects and measures what it alleges to measure and not something else (Thorndike & Thorndike-Christ, 2010). According to Anastasi and Urbina (2007), the validity of an instrument indicates how the measured thing is closely related to the instrument and how well the instrument can be used in measuring something from the designed measurement. Therefore, validity can be defined as the agreement between test scores or measurement and the quality it is believed to measure (Kaplan & Saccuzzo, 2017). In other words, validity has been defined as the extent to which a test measures what it was designed to measure (Aiken, 1980). Validity bears an aspect of precision in measurement (Azwar, 2014). Accuracy becomes important in measurement because it will produce accurate data. Based on the previous definitions, this study recognises validity interprets comprises "accuracy" and "precision," that is the extent to which an instrument can or is able to measure what it was designed to measure, or how far an instrument fit its measuring function.

Valid measuring instruments are not only able to provide data correctly but also must provide a careful overview of the data. Precision means that measurement can provide an overview of the smallest differences between subjects and other subjects (Sunyoto, 2012). For example, in measuring affective aspects aiming to know the honest attitude of the student, an instrument will be valid when the instrument can carefully measure the honest attitude. This is the only quality that the instrument is stated to be valid.

The instrument used to measure a particular aspect which might not provide a precise and accurate measurement will certainly lead to variance errors. The error might be overestimated or underestimated. A valid instrument has a small variance error because the error on measurement is low. Thus, the number can be trusted as an actual number or a number close to the actual state. As previously mentioned, the notion of validity is closely related to the problem of measurement objectivity. Therefore, no validity is generally accepted for all measurement purposes. A measuring instrument is usually only valid in measuring a specific purpose (Kartowagiran et al., 2019).

There are three types of validity; content validity, construct validity, and criterion relative validity (Kerlinger & Lee, 2000). First, the content validity of an instrument shows to what extent the instrument represents all aspects as a conceptual framework. The items in a test


ISSN (print) 2685-7111 | ISSN (online) 2338-6061

https://doi.org/10.21831/pep.v24i1.31868

must consider the representation of the relevant material, which means each item must assess in terms of its relevance to the measured characteristic.

Second, the construct validity of an instrument indicates the ability of the instrument to efficiently distinguish individuals in terms of ownership of a certain trait. The construct validity also determines the extent to which the test score can show the accuracy, adequacy of indicators according to the measured characteristics by the test (Graham & Naglieri, 2003). The instrument is valid when it can explain the construct of an instrument and provide the validation theory of the test. There are three aspects in construct validity; delivering the possible influence of the construct on the test results, making hypotheses based on theories involving on the construct, and testing the hypothesis empirically.

Criteria validity is applied research which basically refers to certain applied criteria, and not to its predictors. According to Isgiyanto (2009), validity is assessed by comparing the test scores with one or more external criterion/a that are known or believed as measuring attributes in the research. In this research; the measuring attributes are social attitudes and indicators. The validity prioritizes on the testability to make predictions.

Next, content validity is an expert agreement on a measured domain that will determine content validity. The designed assessment is believed to be able to measure the affective which is defined in the domain or construct of social attitudes - in this case, are Honesty, Responsibility, discipline, Politeness, care, and Self-Confidence. Content validity is needed in the development of measurement or attitude assessment (Munby, 1997). Content validity provides a positive input to the developed instrument in real terms on each item. Aiken proposed an index of items validity aiming to find out the content validity. Aiken's validity offers the analysis of the evaluators' (rater) perspective if an item is valid or not.

The validity of an item must be able to explain the measurement or assessment and what should be measured. Item validity has a range of validation areas, namely "appropriateness, meaningfulness, and usefulness." Kumaidi (2014) defines the aspect of appropriateness as "... the adequacy with which the content represents the content of the assessment domain about which interferences are to be made." Adequacy of the content (items) of the test represents the content of the assessment domain. The definition shows that the description of behaviour on the sample to be measured is evidence of feasibility. The instrument guide will determine, and it requires an agreement to the results of the expert's assessment in the field measured by the instrument i.e., the social attitude assessment.

The reliability of assessment instruments shows the level of stability, consistency, constancy, and reliability of the assessment model (i.e. social attitude instrument) (Nunnally, 1994). The assessment instrument is reliable when the instrument shows similar results in being used to assess the same subject in different time and conditions. Thus, reliability is the extent to which the results of a measurement can be trusted.

The results of the assessment are reliable when the assessment is conducted several times to the same subject group and obtains relatively similar results (Azwar, 2014) in a condition that the measured aspects of the subject do not change. The phrase "relatively similar results" means there is a tolerance for small differences between the results of several measurements. When the difference is very big over time to time, the measurement results are not reliable and cannot be trusted.

The notions of the reliability of measuring instruments and the reliability of measurement results usually considered as having the same definition. However, it needs to consider the use of notions. The concept of reliability in the reliability of measuring instrument is closely related to measurement error problems (Peterson et al., 2011). The measurement error itself shows the extent of inconsistencies of measurement if the measurements are repeated in the same subject group. One of the factors making the measurement error is the variation of the respondent's response (Viswanathan, 2005), for example, an extreme response where there are items responded massively by respondents. High reliability means having a low measurement


ISSN (print) 2685-7111 | ISSN (online) 2338-6061

https://doi.org/10.21831/pep.v24i1.31868

error and vice versa (Coaley, 2014). In line with this, errors in measurement need to consider aiming to obtain high instrument reliability.

Reliability is the correlation between item scale and all the answer on the instrument items (Robinson et al., 1991). Reliability estimation of the instrument for assessing the students' social attitudes is conducted using the Omega reliability (construct). Construct reliability coefficient is also recognised as Omega coefficient developed by McDonald (Zinbarg et al., 2005). This coefficient emphasizes how far the measuring indicator reflects the latent-compiled factor. It is a sense of the context of factors analysis translated from the classical assessment theory of reliability. The larger the indicator reflects the more the factor, the greater the value of reliability of the assessment. The Instrument reliability is achieved when the combined coefficient of items is > 0.70 (Mardapi, 2017; Nunnally, 1994; Schnabel & Asendorpf, 2013; Sunyoto, 2012), then the instrument is rated as reliable.

Noting the exposure related to the needs of observation instruments assessment of the students ' social attitudes (valid and reliable) then this research specifically aims to: (1) identify the content validity of observations of students' social attitudes, (2) identify the construct validity of the observation instruments of students 'social attitudes, and (3) identify the reliability of observations of students' social attitudes.

## RESEARCH METHOD

This is a research and development study by adopting a Mccoach model that focuses on the validity and reliability testing of the instruments proposed by McCoach et al. (2013). The research was managed on ten elementary schools in Yogyakarta Special Region. The study used an observation guide instrument with a summative rating scale model. The validity of observation guideline was done by the four experts with the Aiken method. Further, observation guidelines were shared on teachers in order to observe students in the limited tryout and expanded tryout. The limited tryout involved 180 students, while the expanded tryout involved 370 students.

The obtained instruments were put into limited trials and then revised based on the results. Data from the limited trial were analysed using the confirmatory factor analysis (CFA) to find out the validity and the Omega reliability used to find out the reliability. The steps aimed to find out the factors/components and valid items to produce valid and reliable instruments. The limited test aims to produce strong indicators to measure each component as an assessment of social attitudes.

The results of the revised limited trial continued to extended trials on a larger scale. The extended trial results were analysed using LISREL versions 8.80. The analysis results of the program produce the values of validity and reliability as well as the appropriate model. This stage produces instruments that are ready to be implemented to assess the social attitudes of elementary students.

This stage includes item validation and estimation of instrument reliability. The validity of an item must able to explain the measurement, assessment, and what should be measured. Validation of assessments was carried out through two stages. The scopes of the validation area are "appropriateness and meaningfulness. The first stage of the item validity is appropriateness, which, according to Popham's opinion in Kumaidi (2014), is defined as "... the adequacy with which the content of a test represents the content of the assessment domain about which inferences are to be made." This shows that the sample description of the desired attitude to be measured is evidence of the appropriateness. The instrument guide determines, and its evidence requires an agreement on the results by the expert review.

The expert agreement on social attitudes that has six components on measurement determines the content validity because the assessment is believed to measure the social attitudes including honesty, discipline, polite, responsibility, care and confidence.


ISSN (print) 2685-7111 | ISSN (online) 2338-6061

https://doi.org/10.21831/pep.v24i1.31868

Retnawati (2016b) states that content validity is determined using expert agreement. Expert agreement, also known as measured domain, determines the content validity stratification (content-related). The content validity of the expert agreement was calculated using the Aiken validity index (Kumaidi, 2014) formulated as in Equation (1), where V is the index of item validity; s is the score set by each evaluator and reduced by the lowest score in the category. (s = r-$I_0$), with r is evaluators' category score and $I_0$ is the lowest score in the scoring category); n is the number of evaluators, and c is the number of selected categories by evaluators.

$$V = \frac{s}{N(c-1)} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (1)$$

The value of V is in the range of 0-1, the higher the value of V; the more valid the instrument is. According to Retnawati (2016a), the results of the Aikens agreement index is categorized into three categories: (1) low validity if the score is less than 0.4; (2) medium validity if the score is 0.4-0.8; and (3) high validity if the score is more than 0.8.

To test the construct validity of the items in the survey instrument, both exploratory factor analysis and confirmatory factor analysis (CFA) were conducted (Shroff et al., 2019). The second stage of validity is the meaningfulness. This stage is identical to proving the construct validity by testing the constructs in the social attitude assessment instrument for elementary students who have six components in which each component has an indicator. The next step, it makes an item statement from the indicator for observational assessment (OA) instruments, it makes the statement item for the observation. After finding the items and indicators, then the construct is developed into three forms of assessment and continued to test to determine the correct construct of accurate measurement results.

This approach was chosen to prove the construct validity in phase I and phase II using confirmatory factor analysis (CFA), and to confirm that the factors/components were supported by data. The construct validity was estimated using the second-order confirmatory factor analysis to test the suitability of the social attitude assessment (Fuad, 2005; Jöreskog & Sörbom, 1996). If there is a correlation among indicators in one component, which suggest whether its need or not the indicator to be combined; the indicators need to be integrated or combined. The indicators that have a high correlation mean it has same components or measuring the same thing. The validity criteria are at a loading factor of at least 0.30 (Azwar, 2014, p. 143). The content factor is used as a reference for making decisions on valid items.

After constructing validity, reliability of social attitude assessment is estimated. Reliability shows that the extent of measurement results with social attitude instruments are consistent, steady, stable, and reliable so that the results can be trusted in the concept of reliability (Azwar, 2014, p. 7). The using of this formula is based on a summative rating scale adopted from Likert for forms observational assessment (OA). The reliability of instrument of social attitudes assessment for elementary school students was estimated under internal consistency approach by using the Omega reliability formula as presented in Equation (2).

$$\rho_{ii} = \frac{(\sum_{i=1}^{k} \lambda_i)^2}{(\sum_{i=1}^{k} \lambda_i)^2 + \sum_{i=1}^{k} (1 - \lambda^2)} \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (2)$$

The coefficient of Omega reliability of 0.70 or more is used on this assessing the instrument (Mardapi, 2017; Nunnally, 1994). The findings indicate that it is a high-reliability index (minimum 0.70) so they can use the instrument for assessment. In other words, the developed instrument of social attitudes assessment can be widely used.


ISSN (print) 2685-7111 | ISSN (online) 2338-6061

https://doi.org/10.21831/pep.v24i1.31868

## FINDINGS AND DISCUSSION

### Content Validity

The items of the instrument were validated by experts which is recognised as Delphi method and then were analysed by using Aiken. This validation aims to find out if the instrument's content designed by the researcher is valid. The experts assessed and gave suggestion on the instruments draft, assessed the guideline and consistency of the words used, the conformability between the guideline and the developed instrument items. The assessment results are presented in scores. The value criteria of V Aiken is less than 0.600 which is included in the less good category, between 0.600 - 0.88 is included in the good category, while greater V than 0.800 is included in the very good category (Suryani et al., 2017). These scores are then analyzed using Aiken approach to find out the level of validity as displayed in Table 1.

Table 1. Results of Aiken Index on the Instrument

| No | Component | Indicators | Items | Aiken Index | Criteria |
|---|---|---|---|---|---|
| 1 | Honesty | Complete the task independently | B1 | 1.000 | High |
| | | | B2 | 0.750 | Medium |
| | | | B3 | 1.000 | High |
| 2 | Discipline | Compliant/ obey with the rules | B1 | 1.000 | High |
| | | | B2 | 1.000 | High |
| | | | B3 | 1.000 | High |
| 3 | Responsibility | Return the borrowed goods | B1 | 0.750 | Medium |
| | | | B2 | 0.917 | High |
| | | | B3 | 0.833 | High |
| 4 | Politeness | Asking for permission when entering or leaving the room/ classroom | B1 | 1.000 | High |
| | | | B2 | 1.000 | High |
| | | | B3 | 1.000 | High |
| 5 | Care | Actively involved in maintaining the class or school cleanliness | B1 | 0.750 | Medium |
| | | | B2 | 1.000 | High |
| | | | B3 | 1.000 | High |
| 6 | Self-Confidence | Express/ State opinion in the class | B1 | 0.917 | High |
| | | | B2 | 0.833 | High |
| | | | B3 | 0.750 | Medium |

Based on the results of the Aiken index in Table 1, all items are valid and proved by the results of the Aiken index > 0.70. The majority of the developed items have medium and high levels of validity. The distribution of the items consists of 14 items in the high category and four items in the medium category. The valid items from the Aiken validity used for data collection on trial test.

### Construct Validity and Reliability

#### *Validity and Reliability on Limited Try Out*

Construct validity and reliability on limited try out were conducted in four elementary schools by involving 180 students. The data obtained were analyzed the Structural Equation Model (SEM) and the results provided construct validity. The analysis results were developed into a model and the valid items in the observational assessment (OA) model of social attitudes using the Lisrel 8.80 program are presented in Figure 1.

In Figure 1, there is also a correlation between errors. If observed, this correlation occurs on items with different indicators. This is more due to other factors such as gender. A small gender correlation is negligible so that the item still can be used.


ISSN (print) 2685-7111 | ISSN (online) 2338-6061

https://doi.org/10.21831/pep.v24i1.31868

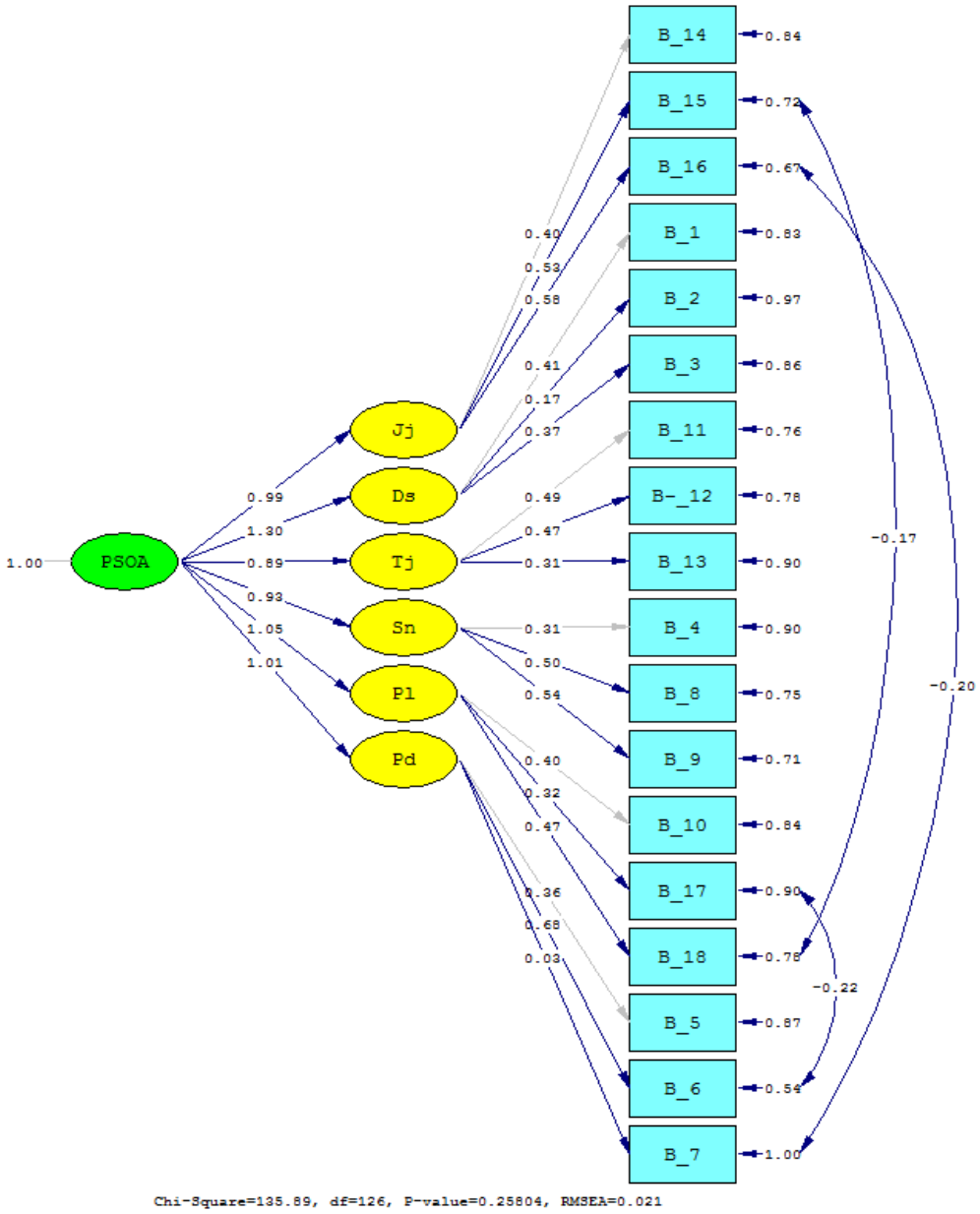


Figure 1. Analysis Result of CFA Second Order of Instrument OA on Limited Field Test

The results of variable analysis using CFA show Chi-Square ($x^2$) = 135.89 df = 126, P-value=0.25804, Root Mean Square Error of Measurement (RMSEA)=0.021 (RMSEA ≤ 0.08). The CFA calculation analysis results are statistically fulfilled. Thus, the social attitude assessment model with the OA instrument is appropriate for data in the field. Table 2 displays the loading factor scores. Based on the scores of loading factor, there are only two invalid items, which are item 7 with a score 0.03 of loading factor and item 2 with a score 0.17 of loading factor. Both the items (2 and 7) have a score of loading factor < 0.30, and the items state as invalid. Item 7 has a very low loading factor (0.03) under the valid item requirements. Item 7 is stated as invalid and revised for an extended trial. Meanwhile, item 2 has been revised. The instrument reliability in this study was estimated using the Omega reliability. The limited trial involving 180 students resulted Omega reliability by 0.7863. The reliability of 0.7863 indicated the high of the Omega reliability coefficient score > 0.70 (Mardapi, 2017). Thus, the social attitude instrument of Observasional Assessment meets the specified reliability requirements and can be used.


ISSN (print) 2685-7111 | ISSN (online) 2338-6061

https://doi.org/10.21831/pep.v24i1.31868

Table 2. The Score of Loading Factors for OA Instruments

| No | Component | Items | Loading Factor | Description |
|---|---|---|---|---|
| 1 | Honesty | Item 14 | 0.40 | Valid |
| | | Item 15 | 0.51 | Valid |
| | | Item 16 | 0.58 | Valid |
| 2 | Discipline | Item 1 | 0.41 | Valid |
| | | Item 2 | 0.17 | Not Valid |
| | | Item 3 | 0.37 | Valid |
| 3 | Responsibility | Item 11 | 0.49 | Valid |
| | | Item 12 | 0.47 | Valid |
| | | Item 13 | 0.31 | Valid |
| 4 | Politeness | Item 4 | 0.31 | Valid |
| | | Item 8 | 0.50 | Valid |
| | | Item 9 | 0.54 | Valid |
| 5 | Care | Item 10 | 0.40 | Valid |
| | | Item 17 | 0.32 | Valid |
| | | Item 18 | 0.47 | Valid |
| 6 | Self-Confidence | Item 5 | 0.36 | Valid |
| | | Item 6 | 0.68 | Valid |
| | | Item 7 | 0.03 | Not Valid |

### *Construct Validity and Reliability on Extended Trial Test*

The extended trial was carried out in six elementary schools in the Special Region of Yogyakarta Province. The six schools were SDN Pakel, SDN Gedongkiwo, SDN Sentolo 3, SDN Berbah, SDN Pakagung, and SD Sokowaten which totally involved 370 students. Before conducting this test, firstly, the instrument items should be maintained and improved because there were invalid items in the previous trial. This improvement was carried out by editing the invalid items.

The data obtained in the extended trial for social attitude assessment (observational assessment) instruments were analyzed by a second order CFA approach in order to determine the validity of instrument items. The Omega reliability estimation was used to test the instrument reliability. Furthermore, the construct validity of the social attitude in observational assessment instrument was estimated using the second order CFA. The analysis employed the Lisrel 8.80 program to determine the level of construct validity, seen from the determined indicators.

The analysis results of the estimation of construct validity on the instrument of social attitudes assessment in the form of observational assessment indicate a weak correlation (-0.12). The instrument OA of social attitudes consists of 18 items for an extended trial test among which the two items do not need to be combined. It means items 6 and 10 can used. The items are indicated as valid if the score of the loading factor is> 0.3. The analysis results using the Lisrel 8.80 programs are presented in Figure 2.

The analysis results of the variable/component of social attitudes using CFA (confirmatory analyzing factor) showed the score of *Chi-Square* ($\chi^2$)=139.64, df=128, P-value= 0.22723 (P-value>0.05), Root Mean Square Error of Measurement (RMSEA)=0.016 (RMSEA ≤ 0.08). The results of the CFA calculation analysis fit the provisions of the statistics suitability. It can be concluded that the OA instrument appropriate for data in the field. In addition, based on the loading factor, all items are valid. A summary of the loading factor scores are presented in Table 3.

The items validity can be seen from the loading factor for each item with the loading factor value 0.3. Therefore, the item is considered to be valid (Setiawan et al., 2019; Setiawan & Suardiman, 2018).


ISSN (print) 2685-7111 | ISSN (online) 2338-6061

https://doi.org/10.21831/pep.v24i1.31868

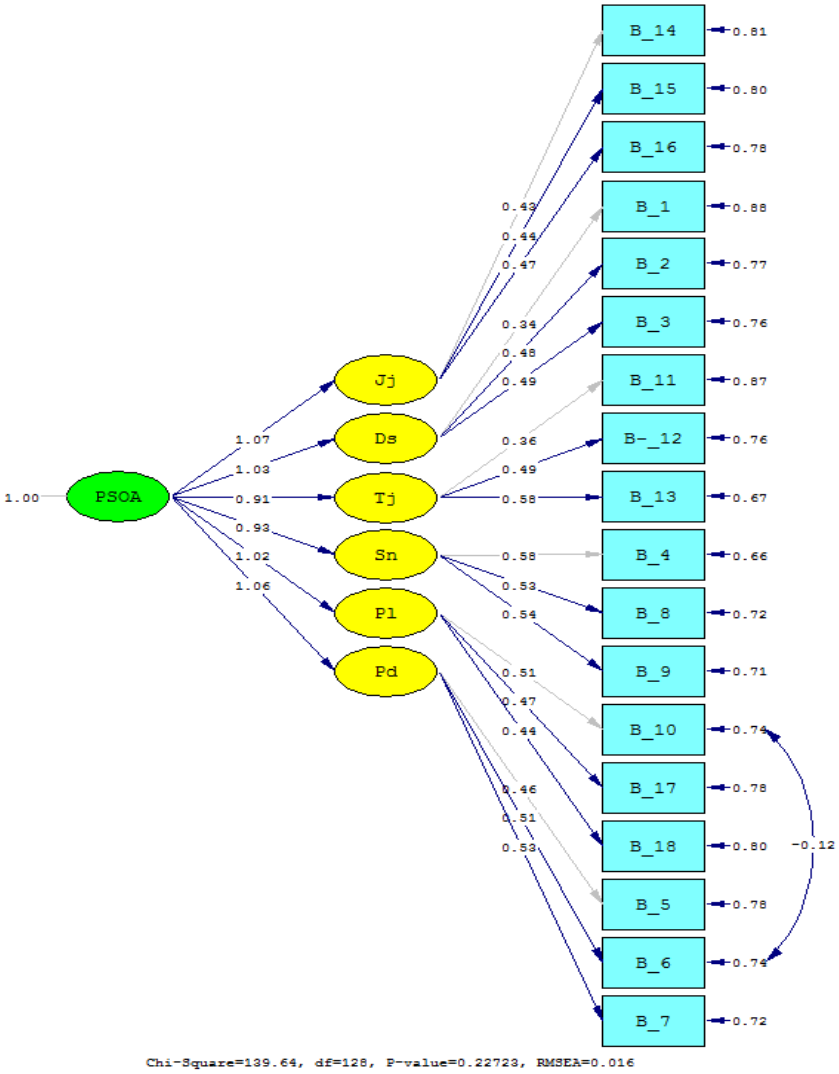


Figure 2. Analysis Result of CFA Second Order of OA Instrument in the Extended Trial

Table 3. The Score of Lading Factor of OA Instrument in the Extended Trial

| No | Component | Items | Loading Factor | Description |
|---|---|---|---|---|
| 1 | Honest | Item 14 | 0.43 | Valid |
| | | Item 15 | 0.42 | Valid |
| | | Item 16 | 0.47 | Valid |
| 2 | Discipline | Item 1 | 0.34 | Valid |
| | | Item 2 | 0.48 | Valid |
| | | Item 3 | 0.49 | Valid |
| 3 | Responsible | Item 11 | 0.36 | Valid |
| | | Item 12 | 0.48 | Valid |
| | | Item 13 | 0.58 | Valid |
| 4 | Polite | Item 4 | 0.58 | Valid |
| | | Item 8 | 0.53 | Valid |
| | | Item 9 | 0.54 | Valid |
| 5 | Care | Item 10 | 0.51 | Valid |
| | | Item 17 | 0.47 | Valid |
| | | Item 18 | 0.44 | Valid |
| 6 | Self-Confidence | Item 5 | 0.46 | Valid |
| | | Item 6 | 0.51 | Valid |
| | | Item 7 | 0.53 | Valid |


ISSN (print) 2685-7111 | ISSN (online) 2338-6061

https://doi.org/10.21831/pep.v24i1.31868

Table 3 shows 18 items have a loading factor score of more than 0.3, meaning that there are 18 valid items (all items) on the OA instrument. The lowest loading factor score is in item 1 with 0.34. In contrast, the highest score is 0.58 for points 4 and 13. The reliability of the assessment instruments in this study was estimated using the Omega reliability (construct) formula. The extended trial involving 370 students show the reliability scores by 0.8433. It is indicated by the high of coefficient score > 0.70 (Mardapi, 2017). A high construct reliability shows internal consistency, so all steps in the measurement consistently represent the same latent constructs (Pada et al., 2018, p. 126). These estimates were comparable to those found in studies using the same scales and, thus, were considered satisfactory (Trinidad et al., 2005). In sum, the social attitude instrument of OA can be used and meet the specified reliability requirements. Based on the limited and extended trials, the instruments are also reliable. The reliable OA instruments can be used to assess the social attitudes of elementary students.

## CONCLUSION

The developed observational assessment instruments have fulfilled the validity and reliability requirements. Validity through the content validity carried out by the experts' assessment and impact to Aiken Index value indicate the validity of the contents of each item resulted that all items met the valid requirements. The instrument also fulfilled the construct validity. Construct validity for limited trials used second-order CFA with invalid results for items 2 and 7. The trial extended on more samples and the results were analyzed by second-order CFA. Based on the results, all items are valid so that the construct validity is fulfilled. Whereas, reliability calculated by the Omega reliability (construct) has fulfilled the requirements for both limited trials and extended trials.

## REFERENCES

Aiken, L. R. (1980). Content validity and reliability of single items or questionnaires. *Educational and Psychological Measurement*, *40*(4), 955–959. https://doi.org/10.1177/001316448004000419

Anastasi, A., & Urbina, S. (2007). *Tes psikologi (psychological testing)*. PT Prehanllindo.

Azwar, S. (2014). *Validitas dan reliabilitas*. Pustaka Pelajar.

Coaley, K. (2014). *An introduction to psychological assessment and psychometrics*. SAGE Publications. https://doi.org/10.4135/9781446221556

Fuad, G. D. (2005). *Structural equation modeling: Teori, konsep, & aplikasi dengan program Lisrel 8.54*. UNDIP Press.

Graham, J. R., & Naglieri, J. A. (2003). *Handbook of psychology: Volume 10, assessment psychology*. John Wiley & Sons.

Isgiyanto, A. (2009). *Teknik pengambilan sampel pada penelitian non-eksperimental*. Mitra Cendikia.

Jöreskog, K. G., & Sörbom, D. (1996). *LISREL 8: User's reference guide*. Scientific Software International.

Kaplan, R. M., & Saccuzzo, D. P. (2017). *Psychological testing: Principles, applications, and issues*. Nelson Education.

Kartowagiran, B., Hadi, S., Wahyumiani, N., Alfarisa, F., & Pusporini, W. (2019). Effectiveness of the AA "4C" authentic assessment model: A single-case-research (SCR). *The New Educational Review*, *57*(3), 200–209. https://doi.org/10.15804/tner.2019.57.3.16


ISSN (print) 2685-7111 | ISSN (online) 2338-6061

Kerlinger, F. N., & Lee, H. B. (2000). *Foundations of behavioral research (PSY 200 (300) quantitative methods in psychology)* (4th ed.). Henry Holt.

Kumaidi, K. (2014). Validitas dan pemvalidasian instrumen penilaian karakter. *Prosiding Seminar Nasional Psikometri*.

Mardapi, D. (2017). *Pengukuran, penilaian, dan evaluasi pendidikan* (2nd ed.). Parama Publishing.

McCoach, D. B., Gable, R. K., & Madura, J. P. (2013). *Instrument development in the affective domain*. Springer. https://doi.org/10.1007/978-1-4614-7135-6

Munby, H. (1997). Issues of validity in science attitude measurement. *Journal of Research in Science Teaching*, *34*(4), 337–341. https://doi.org/10.1002/(SICI)1098-2736(199704)34:4<337::AID-TEA4>3.0.CO;2-S

Nunnally, J. C. (1994). *Psychometric theory 3E*. Tata McGraw-Hill Education.

Pada, A. U. T., Mustakim, S. S., & Subali, B. (2018). Construct validity of creative thinking skills instrument for biology student teachers in the subject of human physiology. *Jurnal Penelitian Dan Evaluasi Pendidikan*, *22*(2), 119–129. https://doi.org/10.21831/pep.v22i2.22369

Peterson, C. H., Schulz, E. M., & Engelhard Jr., G. (2011). Reliability and validity of bookmark-based methods for standard setting: Comparisons to Angoff-based methods in the National Assessment of Educational Progress. *Educational Measurement: Issues and Practice*, *30*(2), 3–14. https://doi.org/10.1111/j.1745-3992.2011.00200.x

Retnawati, H. (2016a). Proving content validity of self-regulated learning scale (The comparison of Aiken index and expanded Gregory index). *REiD (Research and Evaluation in Education)*, *2*(2), 155–164. https://doi.org/10.21831/reid.v2i2.11029

Retnawati, H. (2016b). *Validitas reliabilitas dan karakteristik butir (Panduan untuk peneliti, mahasiswa, dan psikometrian)*. Nuha Medika.

Robinson, J. P., Shaver, P. R., & Wrightsman, L. S. (1991). Criteria for scale selection and evaluation. In *Measures of personality and social psychological attitudes* (pp. 1–16). Elsevier. https://doi.org/10.1016/B978-0-12-590241-0.50005-8

Schnabel, K., & Asendorpf, J. B. (2013). Free associations as a measure of stable implicit attitudes. *European Journal of Personality*, *27*(1), 39–50. https://doi.org/10.1002/per.1890

Setiawan, A., Mardapi, D., Supriyoko, S., & Andrian, D. (2019). The development of instrument for assessing students' affective domain using self- and peer-assessment models. *International Journal of Instruction*, *12*(3), 425–438. https://doi.org/10.29333/iji.2019.12326a

Setiawan, A., & Suardiman, S. P. (2018). Assessment of the social attitude of primary school students. *REiD (Research and Evaluation in Education)*, *4*(1), 12–21. https://doi.org/10.21831/reid.v4i1.19284

Shroff, R. H., Ting, F. S. T., & Lam, W. H. (2019). Development and validation of an instrument to measure students' perceptions of technology-enabled active learning. *Australasian Journal of Educational Technology*, *35*(4). https://doi.org/10.14742/ajet.4472

Stiggins, R. J. (2005). High quality classroom assessment: What does it really mean? *Educational Measurement: Issues and Practice*, *11*(2), 35–39. https://doi.org/10.1111/j.1745-3992.1992.tb00241.x

Sunyoto, D. (2012). *Validitas dan reliabilitas*. Nuha Medika.


ISSN (print) 2685-7111 | ISSN (online) 2338-6061

Suryani, H., Kartowagiran, B., & Jailani, J. (2017). Development and validity of mathematical learning assessment instruments based on multiple intelligence. *Jurnal Penelitian Dan Evaluasi Pendidikan*, *21*(1), 93–103. https://doi.org/10.21831/pep.v21i1.15286

Thorndike, R. M., & Thorndike-Christ, T. M. (2010). *Measurement and evaluation in psychology and education*. Pearson.

Trinidad, S., Aldridge, J., & Fraser, B. (2005). Development, validation and use of the Online Learning Environment Survey. *Australasian Journal of Educational Technology*, *21*(1). https://doi.org/10.14742/ajet.1343

Viswanathan, M. (2005). *Measurement error and research design*. SAGE Publications.

Zinbarg, R. E., Revelle, W., Yovel, I., & Li, W. (2005). Cronbach's α, Revelle's β, and Mcdonald's ωH: their relations with each other and two alternative conceptualizations of reliability. *Psychometrika*, *70*(1), 123–133. https://doi.org/10.1007/s11336-003-0974-7


ISSN (print) 2685-7111 | ISSN (online) 2338-6061