The Reliability of the Indonesian Pedagogy of Science Teaching Test (POSTT)

Listiani[1,2,*] William W. Cobern[2], Brandy A-S Pleasants[2], Betty AJ Adams[2]

[1] *Department of Biology Education, Faculty of Teacher Training and Education, Borneo University Tarakan*
[2] *The Mallinson Institute for Science Education, Western Michigan University, USA*
[*]*Corresponding author. Email: nfn.listiani@wmich.edu*

ABSTRACT
The development of an instrument typically requires a reliability study. For formative assessment instruments, however, reliability is less of a concern unless a formative instrument gets used as a research tool or for summative purposes. When that happens, the formative instrument needs to be evaluated for reliability. The POSTT is one of those instruments that was developed for formative purposes and then subsequently used in research, but not much attention has been paid to reliability. Since only a few researchers have reported the reliability of the POSTT, it is important to investigate its reliability. This is a quantitative study with a secondary analysis of test/retest data. Seventeen preservice biology teachers in the fourth year of Biology department education were the respondents of the study. The data was collected using test/retest and analyzed using SPSS 26 software. The primary purpose of the study reported in this paper was to determine the reliability of the POSTT using a test and retest method. The results suggest test/retest coefficient do not support the Indo-POSTT as a reliable instrument. From eight items of Indo-POSTT, there is only one item that has significant high correlation coefficient. We conclude that the based on test/retest, the Indo-POSTT's reliability is not sufficient suggesting that this instrument is not stable. Therefore, a further investigation using other methods is needed.

Keywords: *Reliability, test-retest, summative assessment, Indo-POSTT.*

1. INTRODUCTION

Although the POSTT was developed for formative purposes, researchers have also put it to use. The original POSTT development work established the validity of the items but did not address reliability. However, if researchers use the POSTT for research purposes then both validity and reliability need to be established. Kimberlin and Winterstein [1] note that research instrument reliability is important for determining the instrument stability. Reporting Cronbach's alpha is one of methods that can be used to determine the reliability of an instrument. Taber [2], however, found that most reliability studies failed to interpret Cronbach's alpha correctly. Thus, this study examined the reliability of the eight-item, Indonesian POSTT (Indo-POSTT) using a test/retest method rather than relying on Cronbach's alpha.

Since its development, the POSTT has been widely used as a research instrument. Ramnarain and Schuster [3], Sondlo and Ramnarain [4], and Ramnarain, Nampota, and Schuster [5] studied the pedagogical orientations of South African Physical Science Teachers using the POSTT. Unfortunately, none of the research using the POSTT reported on the reliability of the POSTT, which is an oversight that needs to be addressed. There is a difference between formative and summative assessment in terms of validity and reliability. Validity is required of both, but Black and Wiliam [6] argue that the reliability of formative assessment is less important as long as teachers can detect irrelevant variations in a test during continuing interactions with the learner. Reliability of a formative assessment instrument can be low since students' knowledge is still evolving [7]. Thus, as a formative assessment, the reliability of POSTT instrument is less important. But, when a formative assessment instrument is used as a summative or research instrument, reliability becomes important.There are several ways for evaluating instrument reliability.

Ary et al. [8] describes three methods: test/retest, equivalent-forms, and internal-consistency. Test/retest reliability indicates consistency of scores over time.

Equivalent–forms reliability reflects variations of performance if the subject is given two different but equal forms. Internal–consistency is used to determine whether or not the items measure similar things. The test and retest can be conducted between 7 days [9] or two weeks [10]. A test/retest coefficient assumes that the characteristic being measured is stable over time [9-12]. The stability of the items can be affected by reactivity, memory, or instability of the underlying theoretical construct. Reactivity is caused by unfamiliarity with item content. Respondents may think about something unfamiliar and then when asked again they answer differently. This reactivity, thus, can affect the stability of the instrument [12]. If too little time passes between test and retest, respondents may remember items and hence respond as they did the first time. Instability refers to item ambiguity, which can lead to a respondent interpreting the item differently on the retest.

In case of reliability, many researchers report the reliability using Cronbach's alpha. Taber [2], having examined research published in high-status research journals published during 2015, found that researchers often reported Cronbach's alpha values. However, he also found that researchers used Cronbach's alpha inconsistently and rarely explained their use of this value. He pointed out that Cronbach's alpha is a measure of internal consistency and thus not necessarily of reliability. One important implication from his work is that researchers should be cautious about using Cronbach's alpha as an indicator of instrument reliability. Regarding the importance of reliability for research instruments, the main goal of this study is to examine the reliability of the eight Indo-POSTT items using test retest correlation as another method to examine the reliability of an instrument. Previously, the Indo-POSTT reliability has been examined using Cronbach’s Alpha method that seems problematic according to Taber [2].

2. METHOD

2.1. Data collection procedure

The data was collected using test-retest procedure and the participants were the fourth semester of preservice biology teachers in one of public universities in the North Borneo, Indonesia. The interval time between test and retest were 7 days [9, 10, 13] and the respondents were not given any treatments during this length of time.

2.2. Data Analysis

This study is secondary analysis. Previously, the reliability of Indo-POSTT was determined using Cronbach’s Alpha [14] suggesting that the instrument was considered as reliable. However, the use of Cronbach’s Alpha for measuring the reliability did not fit the characteristic of the POSTT. Therefore, the data from Listiani et al. [14] was subjected to correlational analyses, given that Cronbach’s alpha had been previously calculated for this data. From test to retest, respondents may change their responses, which could suggest unreliability. Furthermore, change could be directional which would also undermine reliability. Items can be considered stable (reliable) if the retest data highly correlates with the test data and there is minimal directional effect.

3. RESULTS AND DISCUSSIONS

3.1. Results

The test/retest correlation analysis was conducted to determine how well test scores predicted retest scores. The test and retest are considered as strongly correlated when there is very little change between these two administered tests.

Table 1 The Test/Retest Correlation Coefficient of Each Item

No. Item	Corr. Coef.	Sig.
Item_1	0.749**	0.001
Item_2	0.028	0.916
Item_3	0.158	0.544
Item_4	-0.073	0.781
Item_5	-0.009	0.973
Item_6	0.143	0.584
Item_7	0.271	0.292
Item_8	-0.361	0.154

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

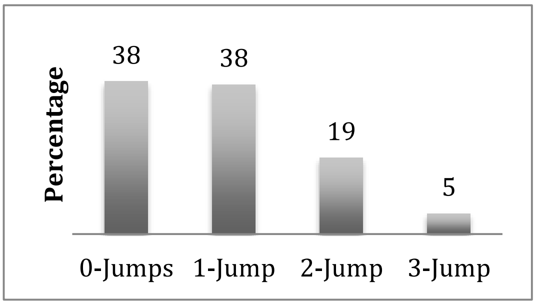

Figure 1 The overall percentage of jumps

Table 1 shows the result of test/retest correlation coefficient. Among eight items, only Item 1 has a significant test/retest correlation, indicating that only for Item 1 is there little change between test and retest. Figure 1 shows that across the eight items of the Indo-POSTT changes are dominated by 1-jump type, with fewer two-jumps and only a very few three-jumps. We use the term ‘jump’ to explain changing responses between test and retest. Zero jump means that the respondent did not change their responses (meaning that their responses on test and retest remain similar). One jump means that they changed their responses one step from previous test. In this study, there are four possible responses, they are didactic direct, active direct, guided inquiry, and open inquiry. Overall, about 62% of total responses changed between test and retest. Such variation indicates why the correlations in Table 1 are low. Thus, the correlational analysis does not support a reliability claim in contrast to the Cronbach's alpha result, which suggests that the items are reliable.

3.2. Discussions

In this study we examined existing data for the purpose of addressing the research goals: examining the reliability of the eight Indo-POSTT items using test-retest, investigating the strengths and weaknesses of this method, and providing recommendations on the use of POSTT items based on the reliability study.

The Indo-POSTT is the Indonesian version of eight POSTT items that have been translated and adapted into Indonesian language. As formative assessment, the Indo-POSTT has been validated during the transdaptation process. Therefore, the eight items of Indo-POSTT are considered as valid for formative purposes. However, as a research tool, the reliability of Indo-POSTT is important. Listiani et al. [14] reported the Cronbach's Alpha for the Indo-POSTT items as an indicator of reliability, which is a common practice in the literature. However, Cronbach's Alpha is specifically an indicator of item internal consistency while the POSTT items were not designed to be internally consistent. The main purpose for using POSTT items is to obtain profiles of teaching orientation, so it is expected that the responses should be spread, which can be a problematic for internal consistency [15].

Another method for measuring reliability is calculating correlation coefficients. In contrast to what one might expect knowing the Cronbach's alpha value for the Indo-POSTT items (~0.70) [14], our result showed that only one of eight items has a high test/retest correlation coefficient, while the rest of the items had low coefficients (Table 1). Furthermore, only the Item 1 correlation is statistically significant. It is not surprising that only one item would have a strong, statistically significant correlation coefficient given Figure 1, where the comparison between test and retest data shows that only 38% responses did not change. However, there is no obvious reason on why Item 1 had the fewest changes making it appear more stable than other items. Studying factors that contribute to item stability has potential for future research. According to Frankenburg et al. [9], the items are considered as stable if the correlation coefficients are high showing agreement between the tests and retest scores. In the case of Indo-POSTT, there is only one item achieving high agreement between test and retest. This means that the Indo-POSTT could be considered as not stable. However, it should be noted that the stability of an instrument is influenced by reactivity, memory, or instability of the underlying theoretical construct. In case of reactivity, this is related to the response of respondents who are not familiar with the instrument, such as the item content. Thus, the responses might change because of this unfamiliarity. Indeed, this reactivity, can affect the stability of the instrument [12]. The other factor that also causes reactivity or changing the responses between test and retest is the length interval between these two administrations of the test, which could be too short. However, in this study, the length of the interval between test and retest was sufficient [9, 10, 13].

Furthermore, regarding the general lack of statistical significance, we note that in this study, the N-size at 17 is relatively low and that N-size factors into significance calculations [16]. Increasing the N-size would likely lead to statistically significant correlations, however, these correlations would very likely still be low. As shown in Figure 1, about 62% of responses changed between test and retest, thus suggesting a low correlation.

Bearing in mind the lack of correlational significance, we still think that it is important to note that some items have positive correlations while others are negative. Table 3 shows that five items have positive correlations while three items have negative correlations. Although there is an item with a very high and positive correlation, overall the correlations are a mixed of positives and negatives. Having both positive and negative correlational coefficients indicates no directional effect. Having no directional effect is aligned with the nature of the development of POSTT, which is to get a spread of responses. In summary, the correlational analyses do not suggest that the items are reliably stable although there is no indication of a directional effect.

Aligned with Taber [2], the use of Cronbach's Alpha for examining the reliability of an instrument should be based on the characteristics of the instrument unless it will not depict the consistence of the instrument in measurement. There are also other various methods for examining the reliability of an instrument [8].

4. CONCLUSIONS

Reliability of a formative assessment instrument is not crucial. However, if the instrument is used for research purposes, it should be reliable. Given the likelihood that the eight items of the Indo-POSTT will get used in research, these items should be found reliable. Given Taber's (2018) critical analysis of how

Cronbach's alpha has been used in the science education literature for estimating reliability, we used test-retest approache for estimating reliability of the Indo-POSTT. Our findings showed that there is only one item out of eight Indo-POSTT items that has high correlation coefficient. On the other hand we also found another evidence showing the reason why these items tend to have low correlation coefficient. Therefore, further investigation of the reliability using other methods should be implemented.

AUTHORS' CONTRIBUTIONS

Listiani contributed to collect the data and data analysis. Listiani, William W. Cobern, Brandy A-S Pleasants, and Betty AJ Adams contributed to conceive and design the study, perform the analysis, and wrote the paper.

REFERENCES

[1] C. L. Kimberlin and A. G. Winterstein, "Validity and reliability of measurement instruments used in research," *American Journal of Health-System Pharmacy,* vol. 65, pp. 2276-2284, 2008.

[2] K. S. Taber, "The Use of Cronbach's Alpha When Developing and Reporting Research Instruments in Science Education," *Research in Science Education,* vol. 48, pp. 1273-1296, 2017.

[3] U. Ramnarain and D. Schuster, "The Pedagogical Orientations of South African Physical Sciences Teachers Towards Inquiry or Direct Instructional Approaches," *Research in Science Education,* vol. 44, pp. 627–650, 2014.

[4] A. Sondlo and U. Ramnarain, "Exploring the South African Physical Sciences Pre-Service Teachers Pedagogical Orientations," vol. 2, pp. 341-345, 2019.

[5] U. Ramnarain, D. Nampota, and D. Schuster, "The Spectrum of Pedagogical Orientations of Malawian and South African Physical Science Teachers towards Inquiry," *African Journal of Research in Mathematics, Science and Technology Education,* vol. 20, pp. 119-130, 2016.

[6] P. Black and D. Wiliam, "The Reliability of Assessment," in *Assessment and Learning,* J. Gardner, Ed., ed California: Sage Publications, 2006.

[7] M. Park, X. Liu, and N. Waight, "Development of the Connected Chemistry as Formative Assessment Pedagogy for High School Chemistry Teaching," *Journal of Chemical Education,* vol. 94, pp. 273-281, 2017.

[8] J. Ary, L. C. Jacobs, C. K. Sorensen, and D. Walker, *Introduction To Research In Science Education*. Boston: Cengage Learning, 2019.

[9] W. K. Frankenburg, B. W. Camp, P. A. V. Natta, J. A. Demersseman, and S. F. Voorhees, "Reliability and Stability of the Denver Developmental Screening Test," *Child Development,* vol. 42, pp. 1315-1325, 1971.

[10] S. M. Cruise, C. A. Lewis, and C. M. Guckin, "Internal Consistency, Reliability, and Temporal Stability of the Oxford Happiness Questionnaire Short-Form: Test-Retest Data over Two Weeks," *Social Behavior and Personality: an international journal,* vol. 34, pp. 123-126, 2006.

[11] B. Nevo, "Using Item Test-Retest Stability (ITRS) as a Criterion for Item Selection: An Empirical Study," *Educational and Psychological Measurement,* vol. 37, pp. 847-852, 1977.

[12] G. Torkzadeh and W. J. Doll, "Test-Retest Reliability of the End-User Computing Satisfaction Instrument," *Decision Sciences,* vol. 22, pp. 26-37, 1991.

[13] É. Dutil, C. Bottari, and C. Auger, "Test-Retest Reliability of a Measure of Independence in Everyday Activities: The ADL Profile," *Occupational Therapy International,* vol. 2017, pp. 1-7, 2017.

[14] Listiani, W. W. Cobern, and B. A. Pleasants, "An Indonesian Translation and Adaptation of the POSTT: A Science Teacher Pedagogical Orientation, Formative Assessment Device," *Journal of Research in Science Mathematics and Technology Education,* vol. 2, pp. 135-148, 2019.

[15] W. W. Cobern, D. Schuster, B. Adams, B. A. Skjold, E. Z. Muğaloğlu, A. Bentz, *et al.*, "Pedagogy of Science Teaching Tests: Formative Assessments of Science Teaching Orientations," *International Journal of Science Education,* vol. 36, pp. 2265-2288, 2014.

[16] L. D. Goodwin and N. L. Leech, "Understanding Correlation: Factors That Affect the Size of r," *The Journal of Experimental Education,* vol. 74, pp. 249-266, 2006.