ADDRESSING AN UNDERGRADUATE RESEARCH ISSUE ABOUT NORMALIZED CHANGE FOR CRITICAL THINKING TEST

S. P. Sriyansyah[1]*, D. Azhari[2]

[1]Department of Physics Education, School of Postgraduate Studies, Universitas Pendidikan Indonesia
[2]Department of Physics Education, Postgraduate Program, Univesitas Negeri Malang, Indonesia

ABSTRACT

Normalized change is a familiar expression used to measure student's improvement in physics education research, including critical thinking skill improvement. A widely used standardized critical thinking test is the Cornell Critical Thinking Test. The CCTT scoring method, rights minus one-half the number wrong, results from possible interval scores ranging from the negative minimum score to positive maximum score. The problem then arises in the use of the normalized change in CCTT scores, particularly in the situation when the post-test score is worse than the pre-test score. We reveal the used equation deficiencies and demonstrate the mistakes made by undergraduate researchers, as well as suggesting a modified equation that can be used under the normalized change rationale, i.e. the ratio of the gain or the loss of the maximum possible gain or loss. Some frequently asked questions about normalized change are also discussed.

© 2017 Science Education Study Program FMIPA UNNES Semarang

INTRODUCTION

The expression of normalized gain has been widely used in physics education research since it was first proposed by Hake (1998a). The normalized gain for treatment is defined as the ratio of the actual gain G (refers to the difference between the post-test score and the pre-test score) to the maximum possible gain G_{max}. The 'Hake gain' is used by instructors, who want to see how much conceptual learning their students achieve, and by researchers, who want to compare the difference in conceptual learning between groups exposed to different pedagogies (Lasry et al., 2014; Hoelwarth et al., 2005; Potter et al., 2014;). It is also considered as a rough measure of the effectiveness of a course in promoting conceptual understanding (Hake, 1998a; 2002; Meltzer and Manivannan, 2002; Von Korff et al., 2016; Berek et al., 2016), scientific literacy (Afriana et al., 2016; Khaeroningtyas et al., 2016) , and the effectiveness of a course curriculum (Colt et al., 2011), a developed learning material (Putra et al., 2016; Triyuni, 2016), online homework system (Cheng et al., 2004), and studio physics format (Cumming et al., 1999; Sorensen et al., 2006; Sorensen et al., 2011; Kohl & Kuo, 2012).

Along with the development of physics education research, some limitations were found in Hake's equation (see Marx and Cummings, 2007). In particular, when the post-test score is worse than the pre-test score, the 'Hake gain' permits ambiguous interpretation of the result. For example, when the average pre-test score $<S_{pre}>$ on the certain exam is 90%, and the average post-test score $<S_{post}>$ on the same exam is 75%, it would give an average normalized gain <g> of –1,5 by using Eq. (1).

*Alamat korespondensi:
E-mail: syaktiperdana@gmail.com

$$\langle g \rangle \equiv \frac{\%\langle G \rangle}{\%\langle G \rangle_{max}} = \frac{(\%\langle S_{post} \rangle - \%\langle S_{pre} \rangle)}{(100 - \%\langle S_{pre} \rangle)} \qquad (1)$$

This situation is often not understood well by undergraduate researchers. Regardless of whether or not they grasp the consideration of the difference in use the class-average normalized gain <g> and the average of the single-student gains g_{ave} by Hake (1998), undergraduate researchers tend to use the less appropriate equation, or they easily eliminate the anomaly data to simplify calculations. It certainly can lead to misinterpretation of the students' performance on the conceptual change after the instruction. Even an effective learning scenario by the experienced teachers might not improve student scores (Cumming et al., 1999) or can produce the post-test score is less than the pre-test score. The latter has likely happened if the students shift their correct answer of the certain item in the pre-test into the incorrect answer of the same item in the post-test, as reported by Lasry et al. (2014).

This deficiency leads Marx and Cumming (2007) proposed a formula called as the normalized change c to complete the 'Hake gain' equation. The normalized change measures all possible changes in students' performance after the instruction with Eq. (2), where the post and pre refer to the student score out of 100%, respectively. In the situation where the post-test score is less than the pre-test score defined as the ratio of the loss to the maximum possible loss, measured by Eq. (2d).

$$c = \begin{cases} \dfrac{post - pre}{100 - pre} & post > pre \quad (2a) \\ drop & post = pre = 100 = 0 \quad (2b) \\ 0 & post = pre \quad (2c) \\ \dfrac{post - pre}{pre} & post < pre \quad (2d) \end{cases}$$

However, Eq. (2d) needs more attention when applying directly to the raw score, in particular case ranged from negative to positive scores, such as a raw score of The Cornell Critical Thinking Test (CCTT) level X and level Z (Ennis & Millman, 2005). One of the recommended scorings of Ennis et al. (2005) for both level X and level Z test is a total score using the formula rights minus one-half the number wrong. Therefore, level X test, for example, which has 76 items (including 5 sample items), then all the possible total scores ranged from –35.5 to 71. Consider a worst case (although it is impossible, only showing an ambiguous interpretation of the result) in which

the pre-test raw score is 71, and the post-test score is –35.5 on the CCTT level X. An undergraduate researcher using Eq. (2d) Would assign a normalized change of –1.5. It can make undergraduate researchers who are not aware of this situation confuses or present a wrong data interpretation.

This paper highlights the issue above as well as addresses some issues about normalized change (we prefer to use the term 'normalized change c' than 'normalized gain g' in this paper) in undergraduate educational research. There are some adequate and important points to be discussed which often becomes frequently asked questions by an undergraduate educational researcher. This paper will help undergraduate researchers understand well the normalized change to avoid misinterpretation of the data.

Outline of Previous Research

Houveland et al. (1949) introduced a parameter independently for measurement of percentage change, who called g as the "effectiveness index." While Gery (1972) named g as the "gap-closing parameter." Hake (1998a,b) in his large student survey of mechanics test data for introductory physics courses, again familiarized g as the "normalized gain" expressed by Eq. (1) and Cohen et al. (1999) called g different than the others, namely "POMP − Percentage Of Maximum Possible".

In his study, Hake (1998) prefers to introduce the class-average normalized gain <g> than the average of the single-student gains gave to characterize a group's improvement (see Hake, 1998 for Hake's considerations). Eq. (1) employs the average class pre-test and post-test scores for obtaining <g>, while g_{ave} utilizes the single-student gain g which characterizes an individual's improvement, as shown in Eq. (3), where pre and post refer to the student score out of 100%, respectively. The comparison of these two methods in calculating the average g and how this comparison may be able making inference about how a group of students has changed as a result of instruction discussed carefully in Bao's paper (2006).

$$g_{ave} = \frac{1}{N} \sum_{k}^{N} g_k = \frac{1}{N} \sum_{k}^{N} \frac{post_k - pre_k}{100 - pre_k} \qquad (3)$$

The 'Hake gain' then became widely used equation by the instructors and the researchers, then Marx and Cummings (2007) found some limitations of using g and <g>, as well as proposing the calculation of the normalized change c in Eq. (2). The normalized change also encom-

passed all possible changes in student performances, including 'the normalized loss' situation for single-student loss, i.e. when the post-test score is worse than the pre-test score. Miller et al. (2010) stated that losses situation are fairly common in the classroom, but the reason behind these losses is still inconclusive whether it demonstrates actual conceptual losses or only lucky guess on the pre-test that became incorrect on the post-test. It may influence the conclusion that is drawn in a study, if or not these losses take into account in data analysis.

RESULTS AND DISCUSSION

An Issue of Using c for CCTT

This issue comes from a case of using c for the Cornell Critical Thinking Test, particularly when students show less performance in the post-test than in the pre-test. It's almost similar to the limitations of using g and $<g>$ previously addressed by Marx and Cummings (2007). In this case, undergraduate researchers usually employ CCTT raw scores in analyzing student pre-test and post-test score, as well as student individual's improvement.

There are two significant deficiencies to Eq. (2d) when directly used to the raw score of CCTT, for example, level X test, that ranges from −35.5 to 71. Firstly, it has a normalized loss bias. Student's positive pre-test scores exhibit an incorrect maximum normalized loss, but negative pre-test scores instead indicate a positive sign of maximum normalized loss. If a student with a pre - test score of −4, which has the number right (NR) of 21 and the number wrongs (NW) of 50, it can achieve a maximum normalized loss of +7.9. Otherwise, if a student with pre-test score of 6.5, which has NR of 28 and NW of 43, it can achieve a maximum normalized loss of −6.5 (coincidentally has the same magnitude with pre-test score). Noted that a positive value of c indicates the positive change or normalized gain and a negative value of c indicates the negative change or normalized loss. Hence, both examples obviously wrong because the maximum normalized loss in both cases should assign the same value of −1. The last, Eq. (2d) produces a non-symmetric range of scores which leads to misinterpretation of the results, as shown by dashed line at the isograms in Fig. (1a).

Responding the Issue

Since the main problem lies in the use of CCTT raw score ranged from the negative minimum of −35.5 to the positive maximum of 71,

we suggest two reasonable approaches address this issue. Firstly, we should scale raw scores into percentages to match Eq. (2). The student's raw score is converted by adding 35.5, then multiplied by 100 and divided by 106.5 (the range between −35.5 to 71). For example, a student whose raw score is −4 has a percentage score of 30% and whose raw score is 6.5 has a percentage score of 39%.

Finally, if we still set the raw score without converting into percentages, Eq. (2a, d) should be modified under a rationale the ratio of the gain or the loss to the maximum possible gain or loss, while Eq. (2b, c) remains unchanged. A general expression of normalized change in Eq. (2a, d) is modified as:

$$c = \begin{cases} \dfrac{post - pre}{max - pre} & post > pre \quad (4a) \\[2mm] \dfrac{post - pre}{pre - min} & post < pre \quad (4b) \end{cases}$$

where pre and post refer to the pre-test and post-test raw scores, as well as min and max, refer to the minimum and maximum raw scores. Eq. (4) can also be used to calculate the class average normalized change $<c>$ if pre and post are the averages of student pre-test and post-test respectively. In CCTT case, min and max score are respectively −35.5 and 71.

These approaches eliminate the normalized loss bias and shape in a symmetric range of scores as shown at the isograms for c in Fig. (1b, c). The normalized changes now range from −1 to +1. Consider a student whose the pre-test score is −4 (30%) or −6.5 (39%); it would have the same maximum possible normalized loss of −1. An undergraduate researcher is usually aware with Eq. (4a) for gain, but not with Eq. (4b) for loss.
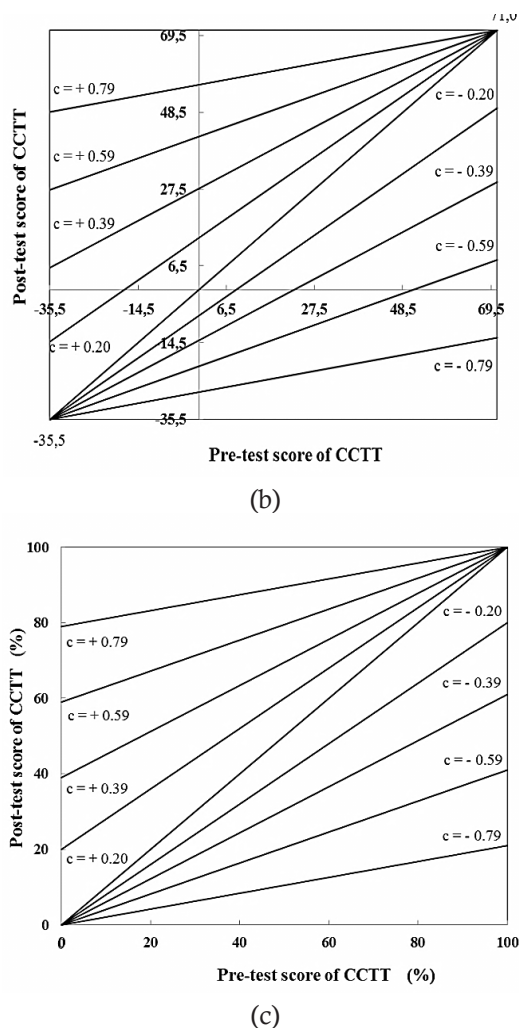

(a)

(b)

(c)
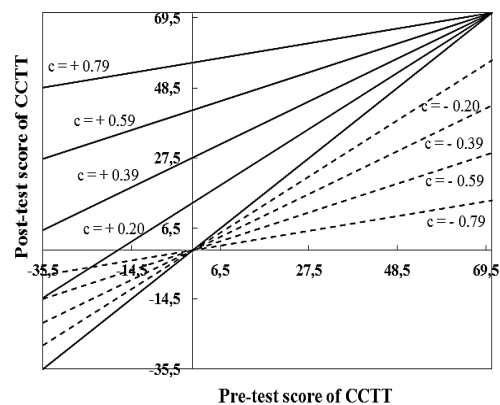
Figure 1. Lines of equal *c* for various CCTT pretest and post-test raw score combinations, calculated (a) using Eq. (2d) for the loss and Eq. (4a) for the gain, (b) using Eq. (4), as well as (c) percentage score using Eq. (2).

An Illustrative Example of the CCTT scores

We present an illustrative example from student's CCTT raw scores in Table 1 to demonstrate the problem and given approaches. The scores are randomly selected only to highlight the differences between the usual approach used by undergraduate researchers and the proposed approach.

Table 1 shows the differences between single-student normalized change scores, particularly when student post-test score is less than pre-test score. These differences usually are not aware of undergraduate researchers if *c*-scores which calculated using Eq. (2d) only within the interval 0 to −1, such as the student #2, #8 and #9. It is due to they understood that the maximum normalized loss should be −1. Undergraduate researchers

might then be aware their mistake when they encounter a case of student #6, #10 and #11, that instead indicates a positive normalized change greater than 1. For example, student #6 with the pre-test score is −4.0 (30%) and the post-test score is −19.0 (15%) if calculated using Eq. (2d), it yields results of +3.75 and yields −0.48 using Eq. (4b). It certainly would give significantly different average normalized change c_{-ave}, causing misinterpretation of the results.

Table 1 also demonstrates a fairly small difference between <c> and c_{-ave} calculated using Eq. (4). The difference between these two scores ranges at 5% for sample N≥20 (see footnote #46 in Hake, 1998). It indicates that the average single-student normalized change can also be an alternative to class-average normalized change. Marx and Cummings (2007) suggested the average single-student normalized change as a more effective way to characterize the whole class improvement and reveal how a group of students has changed after instruction (Bao, 2006). Finally, a suggested way to present a group's improvement is to use the average single-student normalized change associated with the standard error of the mean sem_c, calculated by the formula . Thus, the group's improvement of the illustrative example can be written as c_{ave} = 0.14 ± 0.16. The *c*-score can also be written in the form of 0-100 interval as 14 ± 16 (Cheng et al., 2004), which sometimes it is needed to represent data in a comparison graph effectively. This form often becomes one of the questions asked by the undergraduate students in the educational research methodology.

Frequently Asked Questions

The most frequent question arises in undergraduate research is about the divisions of *c*-score, comparison of the average *c*-score between two different groups, and another use of *c*-score.

Shifting the divisionsc-scores

There is a persistent belief in the undergraduate research that the division of 'Hake gain' (Hake, 1998) might not be shifted. Marx argues that depending on the test format, content area, assessment goals, and the specific sub-population, the divisions may shift[1]. Sometimes the divisions characterize definable shifts in understanding based on the test. Other times, though, the divisions may be more arbitrary and only used as a means to succinctly communicate broad findings. If we decide to shift these divisions, we should clearly define what these divisions represent and find reasons why we use other cut-off levels.

In the case of the Cornell Critical Thinking

Table 1. An illustrative example of fifteen pre-test and post-test scores and corresponding values of c calculated using Eq. (2) and Eq. (4) for each student, as well as the corresponding group average normalized change $<c>$, the average of individual c (c_{ave}), standard deviation and standard error of the mean for individual c's.

Student	Pre-test raw score	Post-test raw score	c calculated using	
			Eq. (4a, 2d)*	Eq. (4)
1	71.0	71.0		
2	71.0	60.5	–0.15	–0.10
3	50.0	60.5	0.50	0.50
4	50.0	71.0	1.00	1.00
5	12.5	12.5	0.00	0.00
6	–4.0	–19.0	3.75	–0.48
7	33.5	54.5	0.56	0.56
8	39.5	33.5	–0.15	–0.08
9	44.0	23.0	–0.48	–0.26
10	–10.0	–35.5	2.55	–1.00
11	–14.5	–25.0	0.72	–0.50
12	39.5	60.5	0.67	0.67
13	33.5	66.5	0.88	0.88
14	–14.5	39.5	0.63	0.63
15	–35.5	–35.5		
Averages	24.40	29.20	0.81	0.14
$<c>$ Eq. (4a)		0.10		
Standard deviation of the mean ()			1.16	0.61
Standard error of the mean ()			0.30	0.16

*Eq. (4a) for the gain and Eq. (2d) for the loss

Test, it doesn't have the divisions of c-scores to characterize the student's critical thinking improvement. Hence, it is free to make our own based on some external criteria. To set divisions, we need first to establish what the divisions mean and how we would measure them independent of the test. It is free to make the argument as to where the divisions are if we have evidence to affirm that our choices divide the test population in some meaningful way. One way to do that is base it on an analysis that compares test scores with some other measure, say interviews[1]. In other words, we try to ascertain that students who get a score above X have a reasonably good understanding of concept Y.

For example, the divisions of c-score by Hake (1998) demonstrates the improvement category of students' conceptual understanding as an impact of the active learning (in Hakes's case is Interactive Engagement) that promotes conceptual understanding and problem-solving. Therefore, it would be inappropriate to adopt Hake's

category into the increased critical thinking skills. Also, the divisions of c-score depend on the characteristic of the instrument and the populations where the instrument was measured.

Comparing two c-scores

Undergraduate researchers usually use mean difference test (t-test) to infer the difference between c_{ave} of two groups, but the difficulty then arises the distribution of the values of c not always approximate a normal distribution. One potential way to solve this issue and compare c-scores between groups exposed to different pedagogies is to look at the standard error of the mean (Marx and Cummings, 2007). A slightly more sophisticated approach would be to calculate the standard error for the scores above the mean and then repeat for the scores below the mean. It would help highlight the possible non-normal distribution of c-scores[1].

In details, when one calculates the standard error, there is a very basic formula that

incorporates the entire data set. So, out of that calculation, it gets one number, the standard error (SE) which is reported as MEAN±SE. The assumption being that numbers that are used to calculate the SE are normal about the mean. However, *c*-scores are typically not normal about the mean. As such, we used to calculate the SE only using the numbers above the mean (call it "SE+") and then do the second calculation with the numbers below the mean (call it "SE-"). Thus, we could report the mean as MEAN(+"SE+",-"SE-"). Once we have that mean with a range denoted by the upper and lower limits provided by MEAN+"SE+" and MEAN-"SE-," we could compare that range to the range from other classes. If the range of standard errors for two set of courses does not overlap, one could claim that as evidence that, subject to the limitation of the survey instrument, the two courses have different degrees of learning. For example, if two courses had *c*-scores 72±5 and 60±10, it would say there is no evidence that the instrument measured difference between those two populations.

However, one should be considered in the use of *c*-score is the issues of Performance Ceiling Effects (PCE) and a correlation between single-student *c* and the pre-test score (Hake, 1998). The researchers tend to accept *c*-score without worrying about this issues from because it is assumed they are not large effects[1]. We should always check to make sure that ceiling effects or strong correlations are not heavily influencing the analysis.

Another use of normalized loss

The normalized loss means negative normalized change. Regardless the use of *c* for the CCTT and based on the normalized change rationale, it can also be used to measure the effectiveness of an instruction to reduce misconception or a remedial instruction (Sriyansyah, 2015). Many previous types of research struggle to identify the students who hold misconceptions (Wijaya et al., 2016; Widarti et al., 2016), then reduce or remediate those misconceptions (Taufiq, 2012).The *c*-score formula, particularly Eq. (4b), has a great chance to be applied in such study to calculate the decreased number of students who hold misconceptions after instruction. This claim under rationale, i.e. the number of students who have been reduced or remediate divided by the total number of students that could have been reduced or remediate (the total number of students who hold a misconception at the beginning of the instruction).

CONCLUSION

We have revealed the mistakes made by undergraduate researchers when applying the normalized change in the CCTT raw scores, as well as suggesting a modified equation that eliminates normalized loss bias and shapes a symmetric range of *c*-scores. We have also discussed some frequently asked questions by the undergraduate educational researcher. We hope this paper help undergraduate researchers understand well the normalized change to avoid misinterpretation of the data, particularly when applying the normalized change in case of the Cornell Critical Thinking Test or the test which has interval score from the non-zero minimum score to the certain maximum score. It would also make a researcher or an educator to reconsider and involve the performance of the students who get 'losses' into data analysis to clearly infer the effectiveness of the different learning pedagogies.

REFERENCES

Afriana, J.,Permanasari, A., &Fitriani, A. (2016). Project based learning integrated to stem to enhance elementary school's students scientific literacy. *Jurnal Pendidikan IPA Indonesia*,5 (2), 261-267.

Bao, L. (2006). Theoretical comparisons of average normalized gain calculations. *American Journal of Physics*, *74*(10), 917-922.

Berek, F. X., Sutopo, S., & Munzil, M. (2016). Enhancement of junior high school students'concept comprehension in hydrostatic pressure and archimedes law concepts by predict-observe-explain strategy. *Jurnal Pendidikan IPA Indonesia*, *5*(2), 230-238.

Cheng, K. K., Thacker, B. A., Cardenas, R. L., & Crouch, C. (2004). Using an online homework system enhances students' learning of physics concepts in an introductory physics course. *American journal of physics*, *72*(11), 1447-1453.

Cohen, P., Cohen, J., Aiken, L. S., & West, S. G. (1999). The problem of units and the circumstance for POMP. *Multivariate behavioral research*, *34*(3), 315-346.

Colt, H. G., Davoudi, M., Murgu, S., & Rohani, N. Z. (2011). Measuring learning gain during a one-day introductory bronchoscopy course. *Surgical endoscopy*, *25*(1), 207-216.

Cummings, K., Marx, J., Thornton, R., & Kuhl, D.

(1999). Evaluating innovation in studio physics. *American journal of physics*, *67*(S1), S38-S44.

Ennis, R. H. & Millman, J. (2005). *Cornell critical thinking test level X fifth edition*. USA: The Critical Thinking. Co.

Ennis, R. H., Millman, J. &Tomko, T. N. (2005). *Administration manual: Cornell critical thinking tests*. USA: The Critical Thinking. Co.

Gery, F. W. (1972). Does mathematics matter. *Research papers in economic education. Joint Council on Economic Education, New York, New York, USA*, 142-157.

Hake, R. R. (1998). Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses. *American journal of Physics*, *66*(1), 64-74.

Hoellwarth, C., Moelter, M. J., & Knight, R. D. (2005). A direct comparison of conceptual learning and problem solving ability in traditional and studio style classrooms. *American Journal of Physics*, *73*(5), 459-462.

Khaeroningtyas, N., Permanasari, A., & Hamidah, I. (2016). Stem learning in material of temperature and its change to improve scientific literacy of junior high school. *Jurnal Pendidikan IPA Indonesia*, *5*(1), 94-100.

Kohl Kohl, P. B., & Vincent Kuo, H. (2012). Chronicling a successful secondary implementation of Studio Physics. *American Journal of Physics*, *80*(9), 832-839.

Lasry, N., Guillemette, J., & Mazur, E. (2014). Two steps forward, one step back. *Nature Physics*, *10*(6), 402-403.

Marx, J. D., & Cummings, K. (2007). Normalized change. *American Journal of Physics*, *75*(1), 87-91.

Meltzer, D. E., & Manivannan, K. (2002). Transforming the lecture-hall environment: The fully interactive physics lecture. *American Journal of Physics*, *70*(6), 639-654.

Miller, K., Lasry, N., Reshef, O., Dowd, J., Araujo, I., & Mazur, E. (2010, October). Losing it: the influence of losses on individuals' normalized gains. In C. Singh, M. Sabella, & S. Rebello (Eds.), *AIP Conference Proceedings* (Vol. 1289, No. 1, pp. 229-232). AIP.

Potter, W., Webb, D., Paul, C., West, E., Bowen, M., Weiss, B., ... & De Leone, C. (2014). Sixteen years of collaborative learning through active sense-making in physics (CLASP) at UC Davis. *American Journal of Physics*, *82*(2), 153-163.

Putra, M. I. S., Widodo, W., & Jatmiko, B. (2016). The Development of Guided Inquiry Science Learning Materials to Improve Science Literacy Skill of Prospective MI Teachers. *Jurnal Pendidikan IPA Indonesia*, *5*(1), 83-93.

Sorensen, C. M., Churukian, A. D., Maleki, S., & Zollman, D. A. (2006). The New Studio format for instruction of introductory physics. *American Journal of Physics*, *74*(12), 1077-1082.

Sorensen, C. M., McBride, D. L., & Rebello, N. S. (2011). Studio optics: Adapting interactive engagement pedagogy to upper-division physics. *American Journal of Physics*, *79*(3), 320-325.

Sriyansyah, S. P. (2015). *Penerapan pembelajaran konseptual interaktif dengan pendekatan multirepresentasi untuk meningkatkan konsistensi ilmiah dan menurunkan kuantitas mahasiswa yang miskonsepsi pada materi termodinamika* (Doctoral dissertation, Universitas Pendidikan Indonesia).

Taufiq, M. (2012). Remediasi miskonsepsi mahasiswa calon guru fisika pada konsep gaya melalui penerapan model siklus belajar (learning cycle) 5E. *Jurnal Pendidikan IPA Indonesia*, *1*(2), 198-203.

Triyuni, T. (2016). The influence of science learning set using scientific approach and problem solving model on learning outcomes of junior high school students in the subject of heat and temperature. *Jurnal Pendidikan IPA Indonesia*, *5*(2), 177-185.

Von Korff, J., Archibeque, B., Gomez, K. A., Heckendorf, T., McKagan, S. B., Sayre, E. C., ... & Sorell, L. (2016). Secondary analysis of teaching methods in introductory physics: A 50 k-student study. *American Journal of Physics*, *84*(12), 969-974.

Widarti, H. R., Permanasari, A., & Mulyani, S. (2016). Student misconception on redox titration (a challenge on the course implementation through cognitive dissonance based on the multiple representations). *Jurnal Pendidikan IPA Indonesia*, *5*(1), 56-62.

Wijaya, c. P., & muhardjito, m. (2016). The diagnosis of senior high school class x mia b students misconceptions about hydrostatic pressure concept using three-tier. *Jurnal Pendidikan IPA Indonesia*, *5*(1), 13-21.