Chen, W. et al. (Eds.) (2016). Workshop Proceedings of the 24th International Conference on Computers in Education. India: Asia-Pacific Society for Computers in Education

Reliability Investigation of Automatic Evaluation of Kit-Build Concept Map by Comparing with Handmade Evaluation Methods

Warunya WUNNASRI*, Jaruwat PAILAI, Yusuke HAYASHI & Tsukasa HIRASHIMA
Graduate school of Engineering, Hiroshima University, Japan
* warunya@lel.hiroshima-u.ac.jp

Abstract: This paper describes an investigation of the reliability of automatic evaluation of concept maps in the framework of Kit-Build concept map. A concept map is a graphical tool for organizing and representing knowledge. In this framework, a learner is provided parts of a concept map and requested to build the concept map by combining the parts. This framework provides a kit-building task of a concept map which is a promising task to enhance and assess learners' comprehension in a topic that they already learned. The framework is practically used in several kinds of school. However, the assessment reliability has not been investigated. In this research, we try to investigate the reliability of assessment learner's comprehension of Kit-build concept map by comparing with handmade evaluation methods of the concept map. Two well-known handmade concept map evaluation methods that include the structural and propositional scoring method are chosen for comparison. These handmade methods can evaluate concept map meaningfully. Moreover, it is flexible for scoring because the human can understand the meaning of each proposition in concept map even the words of proposition do not appear in a learning material. So, the handmade methods are claimed they have reliability for evaluating concept map. To confirm our hypothesis, we designed the preliminary experiment in two learning comprehension situations. Those are learners' reading and instructor's teaching situations. In a preliminary experiment, the correlation between Kit-Build concept map and two handmade concept map evaluation methods in teaching situation has the marginal medium correlation. Even though this is a preliminary result, it suggests that Kit-Build concept map evaluation is reliable for evaluating concept map in the teaching situation.

Keywords: concept map, Kit-Build concept map, concept map evaluation, reliability

1. Introduction

Concept maps are utilized for evaluating learners' understanding widely. Various evaluation methods are proposed in different focus points. Kit-Build concept map is a learning task of exercise for checking learner's comprehension of a topic that they already learned. We have already used Kit-Build concept map in classrooms practically. It confirmed that the framework and results of the diagnosis were useful to support teachers in science learning in elementary school (Sugihara et al., 2012; Yoshida et al., 2013), geography in junior high school (Nomura et al., 2014), and the learning English as the second language (Alkhateeb et al., 2015). These reasons prove Kit-Build is good for using in teaching situation that instructor gives the direction following instructor's interpretation. However, we have not examined the quality of the propositional level exact matching evaluation method that is used in our framework. So we try to investigate the reliability of this method by comparing with handmade concept map evaluation method. These are the motivation of this research.

In the first phase since the 1980s, concept maps were evaluated by investigating the structure of propositions which was the one of an important feature of a concept map. However, some evaluators thought that only structure was not sufficient for evaluating a concept map. So, they provided a more meaningful evaluation method for scoring concept map. The latter method pays attention on the meaning of propositions in a concept map. This propositional method is reasonable for evaluating, but it takes a long time for scoring each concept map. Therefore, the automatic concept map evaluation is invented for decreasing time cost and human workload. It is useful for evaluating

concept maps, but it does not get reliability for evaluating concept map as expected because it does not understand the relations which do not exist in the database. Although researchers try to apply synonym word matching which is very flexible for evaluating by using meaning of words, they have not yet accomplished enough level of accuracy. The propositional level exact matching is one familiar method for evaluating concept map by checking with the criteria map directly. It is very straightforward method that is used for assessing learner map in Kit-Build concept map. Nonetheless, it has been not investigated that it is reliable and proper for evaluating concept map. That is the reason why we want to confirm the reliability of the propositional level exact matching evaluation method by comparing with the reliable handmade evaluation methods.

Even though the reliability of the automatic concept map evaluation is still ambiguous, many educational researchers try to propose their method for using in a learning situation. The important features of concept map are differently focused on scoring. In this study, several automatic frameworks are investigated and compared with Kit-Build concept map about three important features that are the types of their criteria map, the level of evaluated meaning and the type of meaning matching. These features can be analyzed to identify a suitable situation to use Kit-Build concept map.

2. The Concept Map Evaluation Methods

In this study, many types of research about the concept map evaluation method are explained as the overview. First, we try to explore the handmade concept map evaluation method which is a typical one to compare with our framework reasonably. Moreover, detail of the automatic concept map evaluation method is described, and the automatic comparison concept map evaluating method is focused.

The Handmade Concept Map Evaluation Method

The evaluation methods in the handmade concept map evaluation method are used by a human who can understand the meaning of proposition well. In this study, we focus on the methods that pay attention to the structure of concept map and the meaning of proposition of concept map.

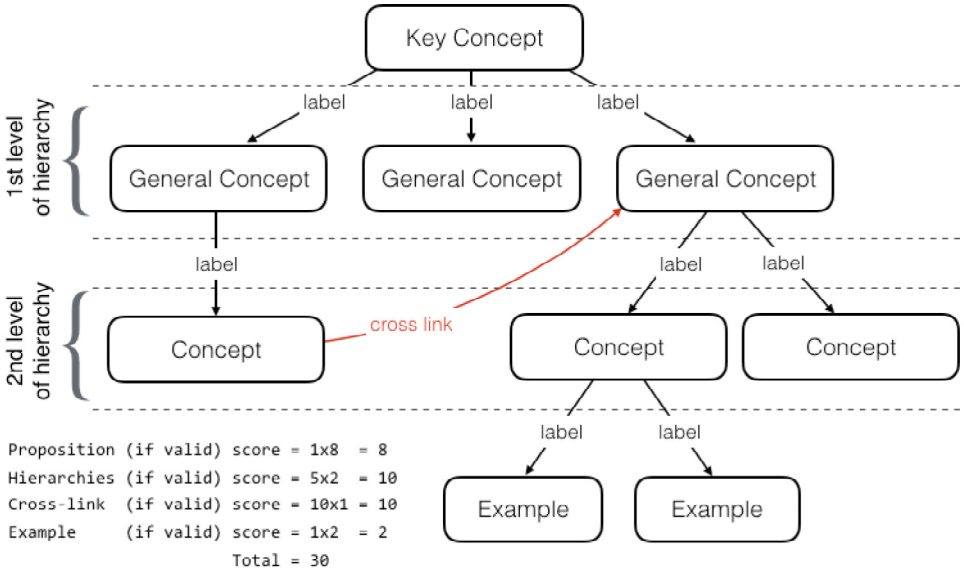

Figure 1. The example of Novak and Gowin structural scoring method

Several concept map evaluation methods evaluate concept map by investigating the structure of concept map such as the level of the hierarchy, the characteristic of the branch, crosslink and so on intentionally. These methods were developed from the 1980s. In this study, we introduce *the*

structural scoring method of Novak and Gowin that was proposed in 1984 and it is a typical structural method (Novak and Gowin, 1984). This method gives high scores for each correct level of the hierarchy and each valid crosslink because ordering the concepts into the hierarchy and connecting the crosslinks can facilitate constructor to have creative thinking. Nevertheless, it tends to the structure more than the meaning of the proposition. That is the reason why it gives only one score for each valid relationship of proposition and example. The example of the Novak and Gowin structural scoring method is illustrated in Figure 1. Furthermore, there are several methods that used concept map structure as the important criteria. The method that was developed by scoring from the connecting of the proposition is also grouped in the structural scoring method (Cronin et al., 1982).

The structural method tends to score the structure of concept map more than meaning may be the cause of important meaning leakage in a concept map. If concept map has much quantity of concepts in the concept map, it can get the point more than the concept map containing a few concepts. After the structural scoring method's phase, a lot of concept map evaluation methods were proposed to improve scoring by taking an interest in the meaning of the proposition.

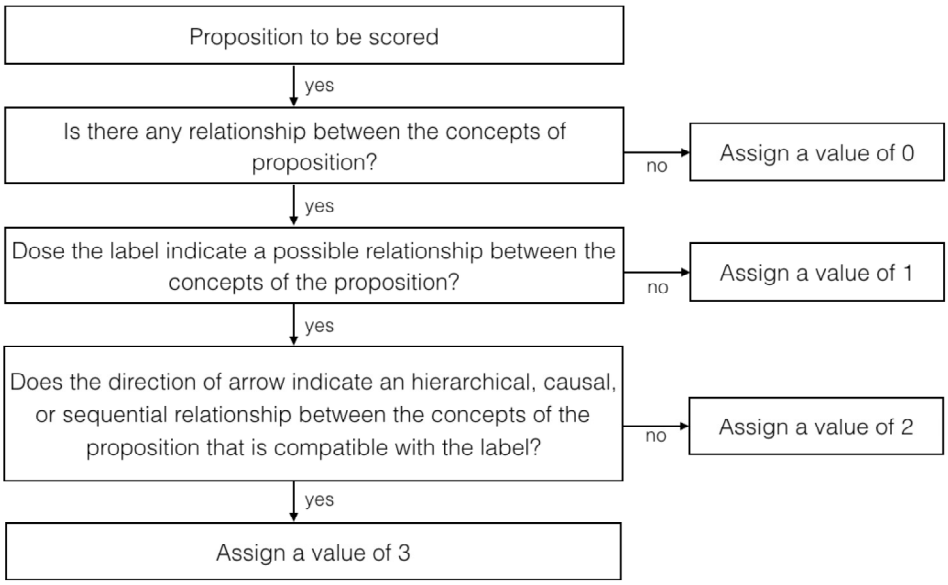

Figure 2. McClure and Bell relational scoring method

After the phase of the structural method, many handmade evaluation methods investigate the meaning of proposition for scoring concept map. They always consider on language and understanding representation, but they ignore the structure of concept map as the proposition precedence. In this study, we call the method of wide acceptance scoring which applies meaning criteria that *the propositional scoring method*. These meaningful methods always have a printed set of criteria as the rubric for assessing knowledge and giving feedbacks differently. However, we will focus on the relational scoring method from McClure and Bell in this study. It is one typical assessment method for concept map evaluation. In the method, the evaluator scores individual maps by evaluating each proposition on the concept map separately (McClure and Bell, 1990). So it will take very long time for scoring each concept map. Evaluator scores zero to three points for each proposition when considered by using a scoring protocol that occurs at Figure 2. Three points are assigned to a proposition which is correct label and representative a hierarchical, causal or sequential relationship between two concepts. Two points are given to proposition that can indicate a possible proposition's relationship. If it does not contain in two first conditions, it will be only checked about existing relation and get one point. Otherwise, the condition will be given zero point. For the reliability of this method, they claimed this method has the most reliability when using the master map by comparing with the holistic method and the structural method (Novak and Gowin structural scoring method). They confirmed the result by using g-coefficient value (McClure et al., 1999).

These handmade evaluation methods are obtained reliability more than the automatic evaluation certainly because the human consideration can understand the meaning of the relationship between concepts deeply, and it is more flexible than the automatic system. However, the evaluators

have to consume much time for evaluating each concept map, and it is necessary to require the evaluator who is an expert in the study area for evaluating following the criteria.

The Automatic Concept Map Evaluation Method

Because of the time consumption and workload for the handmade evaluation, many researchers try to propose the automatic concept map evaluations. Most of the automatic concept map evaluation methods utilize the criteria map as the target of learning. They compare the learner map with criteria map to evaluate learners' understanding. In this study, we call these methods as *the automatic comparison concept map evaluation method*. This comparison inherits the property from the handmade concept map evaluation methods which are the structure of concept map precedence and meaning of proposition precedence. If learner maps are similar to the criteria map, it is obvious that learners can understand in instructor's objective well, which includes the understanding of structure and meaning of the proposition.

In the detail of the automatic comparison evaluation method, it assesses concept map by checking the property of concept map with the correct answer (the criteria map) automatically. This method is desirably used in automatic assessment because the ease of using a matching function to compare learner map with the criteria map reasonably. There are two *types of concept maps* that we must choose for construct as the criteria map. The formal concept map is the first one that is constructed by using valid meaning in universe context and correct viewpoint of knowledge engineering. It also has more concise relations between concepts. That makes it is appropriate for the automatic evaluation but is hard for constructing the formal concept map even it was constructed by an instructor who understands in knowledge engineering fluently. That is different from the informal concept map. The informal concept map can be constructed freely by any words that the creators want to express. The informal concept map is easy to construct, but it is hard for evaluating the system because the system cannot guess the used words thoroughly. Besides the characteristic of the criteria map, we concern about the level of analysis and the type of matching method. For *the level of analysis*, some methods focus on the connectedness of selected concepts or topographical analysis methods to describe the overall geometric structure of concept map; we call the structure level analysis (Schwendiman, 2014). Nevertheless, some method chooses to investigate on the attribute of each proposition instead of the overall structure; we call the propositional level analysis. This level tries to find the valid proposition following its procedure and counts the number of a valid proposition as the evaluating score. One more interesting property is the *type of matching method* when the criteria map is compared with the learner map. The straightforward matching method that we call the exact matching is used widely. It will accept only the propositions that equal with the proposition of the criteria map. The others will be judged as incorrect proposition merely. While some researchers thought that the exact matching is too strict so they proposed the synonym matching to support more flexible comparison.

The Concept Mapping Tool (CMT) is the automatic concept map assessment that evaluated the learner map by using rules-based to check link label and link direction (Cline et al., 2010). The criteria map in this system must create as the formal concept map for supporting university science level major. So it does not facilitate to adjust for appropriate with the instructor's viewpoint as expected. However, they proposed nine rules such as synonym and antonym for more flexible synonym matching in proposition level. In the result part, the CMT reports feedback to learners via the table of proposition accuracy and path rules. Learners can check their mistakes and try to understand the instructor's viewpoint. The other synonym matching that is interesting is the CRESST Human Performance Knowledge Mapping Tool (HPKMT) (Chung et al., 2006). This lenient method also evaluates at the proposition level.

While our framework, Kit-Build concept map framework (Hirashima et al., 2011, 2015) is one of the automatic concept map evaluation method that uses the criteria map to compare with learner map by using the exact matching in propositional level. The task of Kit-Build concept map is separated into two subtasks. The first is the segmentation task that instructor has to prepare the criteria map, which is called the goal map in our framework. The example of the goal map is illustrated in Figure 3. For our system, the goal map is constructed as the informal concept map because the goal map should follow the instructor's objective that requires learners to understand not the overall universe context. The instructor is not necessary to know about knowledge engineering and should

feel free to construct his goal map. After submitting the goal map to the server, it is extracted to be the kit that contains a list of concepts and a list of relationships from the goal map. Moreover, this kit that is provided to learners can help learners to reduce their cognitive load more than the traditional concept map, which they must create all components by themselves. The second task is called the structuring task. Learners are given the learning task to reconstruct concept map by using the kit, which this map is called the learner map (Figure 4). After leaner maps are uploaded to our server, Kit-Build concept map will check leaner maps by matching each learner's proposition with goal map's proposition. That is the exact matching in propositional level. The system will generate a score of correction in a percentage format. However, the instructor can investigate learners' misunderstanding individually and can find the overview of all learners by overlaying the concept map as the group map (Figure 5) and the group-goal difference map (Figure 6) from Kit-Build concept map system immediately. In the group map, the link weight means the number of learners who link that relationship. From the example, the number of learners who connected "use to form" link between "Concept" node and "Propositions" node more than "used to form" link between "Linking Words" node and "Propositions" node. Furthermore, three types of error link are represented in the group-goal difference map. The lacking link is a link that exists in the goal map but does not exist in learner map. The excessive link is a link that occurs in learner map but does not occur in the goal map. Last, links that are not connected to any concepts in learner map are the leaving link. The instructor can use these links to find the holistic leaking understanding of all learners. Following Kit-Build concept map framework's ability, the instructor can use Kit-Build concept map to check understanding of individuals or group of learners and can use the diagnosis result to discuss with learners about the meaning of each error links. After error links analysis, the instructor can adjust the goal map or can teach learners about leaky content repeatedly.

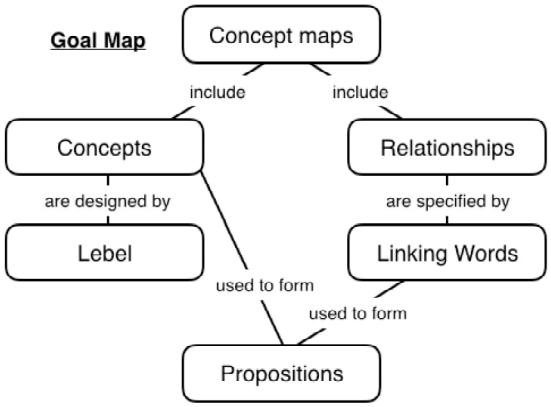

Figure 3. The example of goal map

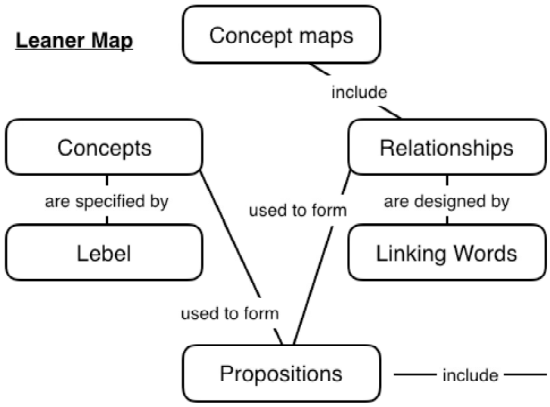

Figure 4. The example of learner map

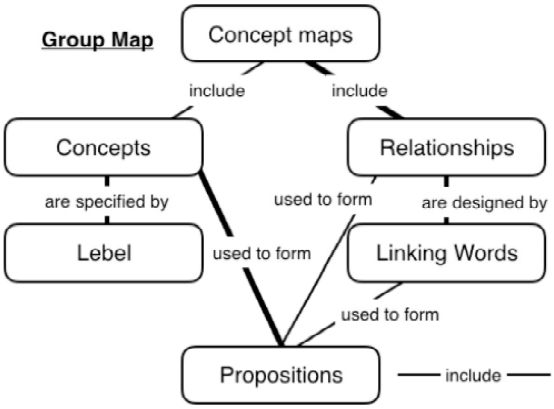

Figure 5. The example of group map

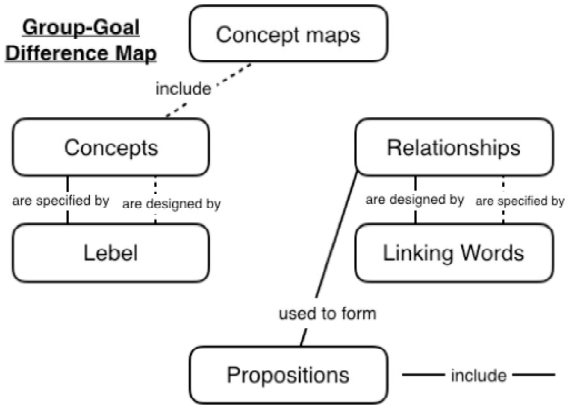

Figure 6. The example of group-goal difference map

3. Research Methodology

Because we have to investigate the reliability of Kit-Build concept map to confirm learners understanding, we tried to design the preliminary experiment for comparing Kit-Build concept map with the handmade concept map evaluation methods. The evaluators who use the handmade concept map evaluation methods can understand the meaning of each proposition even that relationships are expressed in a different viewpoint. So, the handmade concept map evaluation methods are claimed as the reliable evaluation methods. If the correlation between Kit-Build concept map and the handmade evaluation methods is a positive relationship, it means Kit-Build concept map is not much different from the handmade evaluation method and appropriate for using in concept map evaluation.

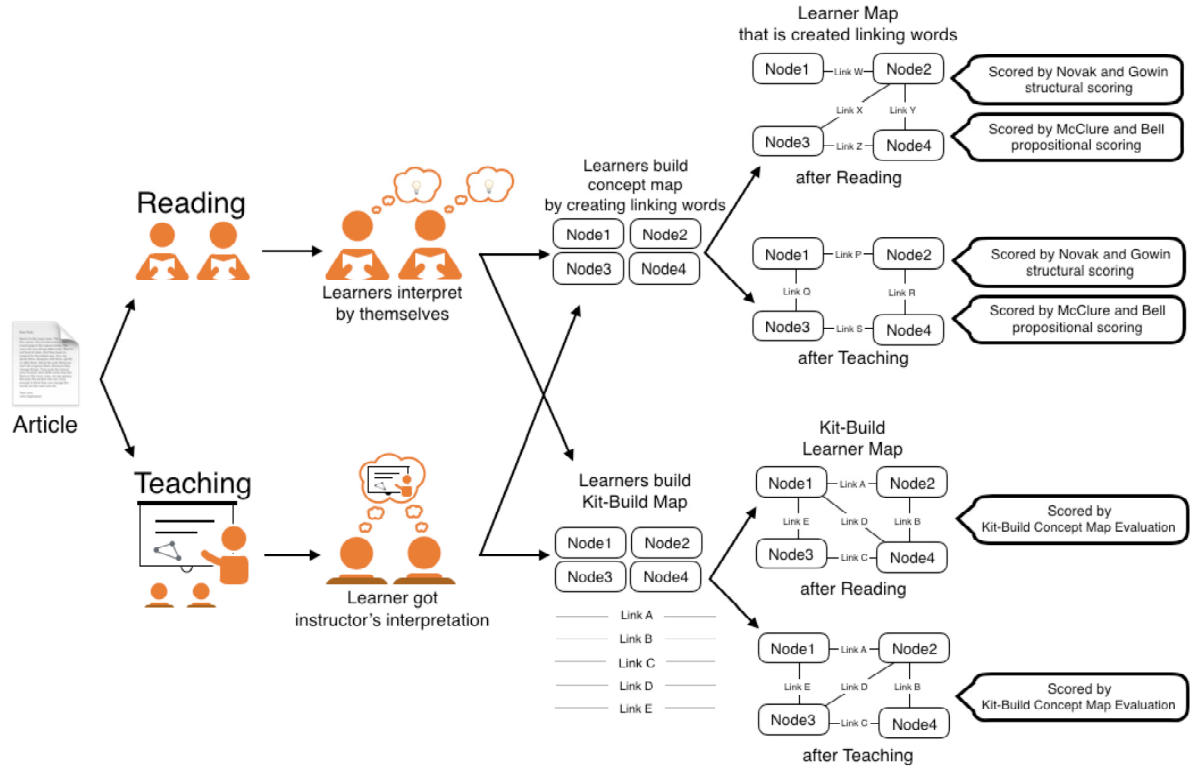

Figure 7. The preliminary experiment procedure

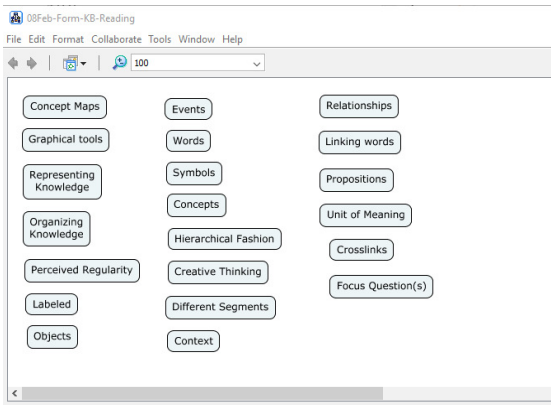

Figure 8. the experiment in CmapTools

Figure 9. the experiment in Kit-Build concept map

From the intentions that are expressed above, we designed the preliminary experiment for confirming our assumption displayed in Figure 7. In our preliminary experiment, ten university students were requested to read the article about "Introduction of concept map" (Novak and Cañas, 2008, p.1) and interpreted the article by themselves. After that, they constructed the concept map by using 21 provided concepts and had to create the linking words on their words in CMapTools application illustrated in Figure 8. These learner maps were evaluated by using the Novak and Gowin

structural scoring and the McClure and Bell relational scoring which is the handmade concept map evaluation. Then, they were requested to build concept map again by using Kit-Build concept map. Participants had to connect each proposition by using provided kit. In this preliminary experiment, the kit contained 21 concepts that are same as provided concepts in CmapTools application and additional provided 22 relationships. The initial screen of Kit-Build concept map in this preliminary experiment is represented in Figure 9. After they connected the propositions completely and uploaded their map to our server, these learner maps were evaluated by Kit-Build concept map evaluation method that is the exact matching in propositional level. When we finished in reading situation, all participants are taught about the same article following instructor's interpretation. Then they had to construct the leaner maps following the same procedure as reading situation, which contains the constructing learner maps by creating linking words by themselves and constructing learner maps by using Kit-Build concept map.

4. Preliminary Experiment Results and Discussion

In this preliminary experiment, the score from two handmade methods was normalized to the percentage score already. The scores from Novak and Gowin structural scoring were normalized by the perfect score of the goal map and the scores from McClure and Bell relational scoring were normalized by the perfect score of each learner map. The average score of each method in reading and teaching situation are shown in Table 1. In this section, we will discuss the result of reading and teaching situation.

Reading Situation

Reading situation means the learners are given the article. They have to read and interpret by themselves. However, their understanding may be different from the others and instructor's viewpoint because each learner has distinct existing knowledge before they read. So it is hard to make an agreement of understanding.

The score of Novak and Gowin's structural scoring in reading situation obtains the lowest percentage, and the score of Kit-Build concept map is better than the first one a bit. The highest score is from McClure and Bell relational scoring. The result between Novak and Gowin structural scoring and the McClure and Bell's relational scoring in our preliminary experiment is in the same way with the study of (McClure et al., 1999) that the latter get scores more than the former. The reason that the participants obtain a low score in reading situation is participants who read the material and interpret the article by themselves constructed learner map by using the different viewpoint from instructor's objective. Following their reading interpretation, they cannot understand the extracted part of goal map that Kit-Build concept map provides. In this situation, participants said they could read and understand the article well, but it is very difficult to create linking words properly for building concept map from provided concepts. The most of the reasons that we received are about the number of concepts is too much. In addition, the problem, when they use Kit-Build concept map, is they cannot understand the relationship between the provided nodes and links. Many components can make learners confuse and worry to connect.

Table 1: Map scores in the experiment (n=10).

	Structural Scoring (Novak and Gowin's)	Propositional scoring (McClure and Bell's)	Automatic comparison (Kit-Build map)
Reading	24.56	35.09	30.30
Teaching	28.99	53.41	50.00

Teaching Situation

Teaching situation means instructor explains about the essence of the material to learners in the same time and same presentation. Learners will be conducted in the same structure of knowledge from the instructor. So, it is an easy way to make the learners understand in the same way. It is properly with Kit-Build concept map, which requests the situation that shared time and presentation in the same structure. The advantages of Kit-Build concept map are it utilizes these shared components to arrange learners' viewpoint and check their understanding following instructor's objective. In contrast, learners have to interpret by themselves in reading situation. There is no any guidance from anyone, and it maybe makes some different understanding from the instructor's objective. The reading situation is good for representing their thinking, but it is very hard to evaluate when the instructor has specific purposes.

After teaching, participants can improve their score for both constructing concept map by creating linking words and Kit-Build concept map. Score from Novak and Gowin structural scoring slightly increases from the reading situation while the others exceedingly rise. Because learners do not change the structure of their leaner map significantly. They paid attention to adjust the words of relationship that they understand from teaching context more than changing structure of concept map, and they do not try to think deeply about creating more level of hierarchy and cross-link. So the structure is not much different from the previous one. However, the score we got from McClure's relational scoring remarkably increases because the participants tried to agree on the context from the instructor. They adjusted the relationship following instructor's viewpoint. Additionally, the adjusted more meaningful words impact to this scoring method that tends to investigate on the meaning of each proposition. In the same way with the score from Kit-Build concept map, it rises extremely but still less than the relational method's score. Because learners' viewpoint is fulfilled by instructor's viewpoint in a teaching situation which uses wording the same as provided linking words, so learners can connect the relationship between each concept more clearly. Nevertheless, the reason that Kit-Build concept map is less than the score from the McClure and Bell relational scoring because the latter is evaluated by a human, who tries to understand each proposition in the maps so the score can increase extremely. On the other hand, the automatic system like Kit-Build concept map does not have this flexible feature because of the limitation of development. It does not try to interpret learner's thought, but it just compares the learner map with the goal map intently. Even though the score of Kit-Build concept map is less than the meaningful human evaluation method in both situations, it is acceptable because this automatic system can check and confirm learners' understanding and get a resembling score with the human evaluation.

From the result of the experiment, we found the correlation of results that is represented in Table 2. The p-values show that we cannot discuss the correlation between both structural and relational method with Kit-Build concept map in reading situations, because, when the learners read the material, they interpret the information by themselves, and it is possible to be various ways. On the contrary, the result of teaching situation has a marginal medium correlation between both handmade evaluation method and Kit-Build concept map. Moreover, we assume the handmade concept map evaluation method is reliable. It refers the lecture from instructors can make an agreement on that material by teaching and conducting the learners' understanding to the same direction with the instructor. These results suggest that Kit-Build concept map is suitable for the teaching situation more than the reading situation.

Table 2. The correlations between the handmade evaluation method and Kit-Build concept map

	Kit-Build concept map In a reading situation	Kit-Build concept map in a teaching situation
Novak and Gowin's structural scoring method	0.1406 (p-value=0.6984)	0.6209 (p-value=0.0553)
McClure and Bell's relational method	0.2702 (p-value=0.4503)	0.5520 (p-value=0.0980)

The one interesting aspect of this preliminary experiment is the improvement of learners when their use Kit-Build concept map after teaching situation. The score increases from reading situation

obviously because participants cannot match their understanding with the provided nodes and links of Kit-Build in reading situation. However, teaching context can extend their viewpoint and makes them more acceptable the instructor's perspective. From the reason that the provided nodes and links will restrict the learners, it helps to keep their building following the instructor's perspective. In contrast, the teaching context can help participants to improve the concept map that is constructed by creating linking words slightly because learners try to represent their understanding including the instructor's viewpoint. However, it just expands their concept map not much and cannot go through the instructor's viewpoint properly.

Also, to think about the suitable situation for using Kit-Build concept map, three properties of Kit-Build concept map are discussed, because the goal map of Kit-Build is built in the form of an informal concept map. It suits for the situation that instructor need learners to understand in specific content and viewpoint. After the goal map is decomposed to be a kit and learners use it to reconstruct the learner maps, it is possible to use the exact matching for comparing between the goal map and leaner maps because of the restrictive set of words from the goal map's components. The semantic matching is not necessary when our framework provides the kit to learners. In addition, the words in the kit should be the common words that learners have to understand well. It is very important to discuss for making the common knowledge explanation with others. For the level of analysis, because our framework uses the goal map to compare with the learner maps, so the evaluating concept map by comparing in proposition level is very important for checking the meaning of the proposition. It is useful for generating informative diagnosis results. Our framework can show learners' misunderstanding or the part of the content that learners need more explain. While the reading situation learners must read the material by themselves and they may interpret in a different way from instructor's expectation. These reasons can support to confirm the exact matching in proposition level with the informal goal map of Kit-Build concept map is important and suitable in teaching context.

5. Conclusion

Our study tries to investigate the reliability of Kit-Build concept which is used in several kinds of school practically. We produce the preliminary experiment that compares Kit-Build concept map with the handmade concept map evaluation methods in two learning situation. The handmade concept map evaluation methods are categorized into two groups following the precedence. The structural scoring scores concept map by investigating composition straightforwardly. It is inconvenient for using in the classroom that instructor has to follow through the unit of instruction. On the other hand, the propositional scoring gives precedence to the meaning of propositions more than the structure. It is reasonable for evaluating understanding from concept map, but this method has to use an expert for checking and taking a long time for scoring. These methods are flexible and meaningful concept map evaluation, and the reliability of them is accepted widely. So we compare Kit-Build concept map with them for confirming our hypothesis. From our preliminary experiment, the results show the correlation between Kit-Build concept map and the handmade concept map evaluation methods in teaching situation has a marginal medium correlation. It means Kit-Build concept map is not much different from the handmade concept map evaluation methods. From the assumption that the handmade concept map evaluation method is reliable, so we can conclude that Kit-Build concept map is reliable and sufficient for evaluating concept map in a teaching situation.

For the future work, we desire to make more strong confirmation of Kit-Build reliability by using g-coefficient value in the full experiment. Moreover, the difficulty of evaluating by synonym matching of handmade evaluation may affect the reliability of evaluation. It depends on the evaluators who score concept map. While the exact matching of Kit-Build concept map always returns the score depend on the comparison between the goal map and learner map automatically. From this assumption, we try to confirm the reliability and stability of Kit-Build concept map is good enough for evaluating concept maps.

Acknowledgements

This work was partially supported by JSPS KAKENHI Grant Number 26280127 and 15H02931.

References

Alkhateeb, M., Hayashi, Y. & Hirashima, T. (2015). Comparison between Kit-Build and Scratch-Build Concept Mapping Methods in Supporting EFL Reading Comprehension. *The Journal of Information and Systems in Education*, *14*(1), 13-27.

Ausubel, D. P., Novak, J. D., & Hanesian, H. (1978). *Educational psychology: A cognitive view* (2nd ed.). New York: Holt, Rinehart and Winston.

Chung, G. K., Baker, E. L., Brill, D. G., Sinha, R., Saadat, F., & Bewley, W. L. (2006). *Automated Assessment of Domain Knowledge with Online Knowledge Mapping*. CSE Technical Report 692. National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Cline, B. E., Brewster, C. C. & Fell, R. D. (2010). A rule-based system for automatically evaluating student concept maps. *Expert systems with applications*, *37*(3), 2282-2291.

Cronin, P.J., Dekker, J., & Dunn, J.G. (1982). A procedure for using and evaluating concept maps. *Research in Science Education*, *12*(1), 17-24.

Hirashima, T., Yamasaki, K., Fukuda, H., & Funaoi, H. (2015). Framework of kit-build concept map for automatic diagnosis and its preliminary use. *Research and Practice in Technology Enhanced Learning*, *10*(1), 1-21.

Hirashima, T., Yamasaki, K., Fukuda, H., & Funaoi, H. (2011). *Kit-Build Concept Map for Automatic Diagnosis*. Proceedings of Artificial Intelligence in Education 2011 (pp.466-468). Auckland, New Zealand: Springer-Verlag Berlin Heidelberg.

Hu M. L. M., & Wu, M. H. (2012). The effect of concept mapping on students' cognitive load. *World Transactions on Engineering and Technology Education*, *10*(2), 134-137.

Liu, J. (2013). The Assessment Agent System: design, development, and evaluation. *Educational Technology Research and Development*, *61*(2), 197-215.

Luckie, D., Harrison, S. H., & Ebert-May, D. (January 01, 2011). Model-based reasoning: using visual tools to reveal student learning. *Advances in Physiology Education, 35*(1), 59-67.

McClure, J.R., Sonak, B. & Suen, H.K. (1999). Concept map assessment of classroom learning: Reliability, validity, and logistical practicality. *Journal of Research in Science Teaching*, *36*(4), 475–492.

McClure, J.R., & Bell, P.E. (1990). *Effects of an environmental education related STS approach instruction on cognitive structures of pre-service science teachers*. Pennsylvania: State University.

Mueller, J. (n.d.). *Concept map rubric*. Retrieved April 25, 2016, from http://jonathan.mueller.faculty.noctrl.edu /240/conceptmaprubric.htm

Nomura, T., Hayashi, Y., Suzuki, T. & Hirashima, T., (2014). Knowledge Propagation in Practical Use of Kit-Build Concept Map System in Classroom Group Work for Knowledge Sharing. *Proceeding of International Conference on Computers in Education Workshop* 2014 (pp.463-472). Nara, Japan: ICCE 2014 Organizing Committee.

Novak, J. D. (1972). The use of audio-tutorial methods in elementary school instruction. *The audio-tutorial approach to learning*, 110-120.

Novak, J. D., & Cañas, A.J. (2008). *Technical Report IHMC CmapTools*. Florida: Institute for Human and Machine Cognition.

Novak, J. D., & Gowin, D.B. (1984). Learning how to learn, New York: Cambridge University Press.

Schwendimann, B. A. (2014). Concept mapping. In R. Gunstone (Ed.), *Encyclopedia of Science Education* (pp. 1-5). Netherlands: Springer.

Sugihara, K., Osada, T., Nakata, S., Funaoi, H. & Hirashima, T. (2012). Experimental evaluation of kit-build concept map for science classes in an elementary school. *Proceedings of Computers in Education* 2012 (pp.17-24). Singapore: National Institute of Education.

Taricani, E. M. & Clariana, R. B. (2006). A technique for automatically scoring open-ended concept maps. *Educational Technology Research and Development*, *54*(1), 65-82.

Victorian Government. (2013). *Assessment Tools*. Retrieved April 25, 2016, from http://www.education.vic.gov.au /school/teachers/support/Pages/tools.aspx

Yoshida, K., Sugihara, K., Nino, Y., Shida, M., & Hirashima, T. (2013). *Practical Use of Kit-Build Concept Map System for Formative Assessment of Learners' Comprehension in a Lecture*. *Proceedings of Computers in Education 2013* (pp.906-915). Bali, Indonesia: Asia-Pacific Society for Computers in Education.