BE -01

CONTENT VALIDITY OF CREATIVE THINKING SKILLS ASSESSMENT

Andi Ulfa Tenri Pada* Badrun Kartowagiran, Bambang Subali****

**Universitas Syiah Kuala, Indonesia*

***Universitas Negeri Yogyakarta, Indonesia*

*Corresponding email address: andiulfa.usk@gmail.com

Abstract

Procedures for computing content validity coefficients and determining the statistical significance of these coefficients are described in this paper. The purpose of this study is to evaluate the content validity for creative thinking skills assessment that support conation aspect of prospective biology teacher. These procedures can be use in a variety of situations were judgements of the content validity of item or questionnaires are made on ordinal rating scales. A panel of six subject-matter experts (SMEs) rated the representation of content based on relevance, construction, and clarity. Computing data set using the Aiken's *V* formula for determining the statistical significanceof *V* for small samples of raters on the five point scale are given. The result, of 52 items, those with content validity coefficient greater than 0,79 (48 items) remained and the rest (four items) were discarded.

Keywords: Validity, content validity, Aiken's V formula.

INTRODUCTION

Creative thinking skills is one of the thinking skills dimension that need to be further developed and measured. Measurement of creative thinking abilities can be done by creating a divergent thinking tasks (Subali, 2011). Divergent thinking is part of the creative process ability. Divergent thinking is an ability to construct or produce a wide range of possible responses, ideas, alternative options to solve a problem (Isaksen, Dorval, &Treffinger, 1994). Thus, divergent thinking can be defined as the ability to elaborate a solutions to solve the problems with the procedures and the right reasons.

Creative thinking skills not only involve cognitive aspects, but also can not be separated from the conation aspect (Lubart, 2004; Poole & Van de Ven, 2004; Jo, 2009). Conation is a mental process that directs the behavior and actions (Huitt & Cainn, 2005). Various terms are used to represent the conation aspects including the intention or tendency to behave (Riyanti & Prabowo, 1998; BSNP, 2010). Conation components playing role in determining the readiness or willingness to act towards the object. Thus, although the students already have a good concept understanding, but their actions could be contradictive. According to Darmawan research (2013), although the students already have a fairly good concept understanding, he still uncertain the students apply their knowledge in the real world. When the students have learned about the circulation and respiration system, they should already know the bad effects of smoking for the heart and the lungs, but some student still smoking. On the other hand, students concept understanding also can produce constructive actions that can contribute to the character development. When they feel the benefits of their knowledge, they becoming more aware of the value contained in the subject materials. They will stop smoking, do not smoking, or remind their friend to stop smoking. So that, it is clear that the cognitive aspect which involved in the creative process can be supported or inhibited by the will or conation aspect.

Paying attention to the root of the problem, it is necessary to think how to solve it. Moreover, the application of the competency-based curriculum in higher education focused to train way of thinking and reasoning, developing creative activity, developing the ability to solve problems and communicate ideas. We offer a solution by developing assessment tool that can measure creative thinking skills that support the conation aspects of the students through a divergent task. This assessment is expected to improve the creative thinking abilities of the students. The students reasoning ability will be directed to produce arguments based on their concept understanding in the form of conation aspect ideas. As the result of the stimulus that have been given, educators will be able to see the students divergent production pattern in the form of rationality alternatives to explain the concepts which contradictive with the action.

After developing the assessment, we need to prove whether the assessment tool has been optimally constructed to evaluate the quality of the assessment. The most important consideration in evaluating the quality of the test as a measurement tools is validity. Messick defines validity as "an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions based on test scores or other modes of assessment" (Reynolds, Livingstone, &Willson, 2009). In line with Messick, the *Standards for Educational and Psychological Testing* defines validity as " the degree to which evidence and theory support the interpretations of test scores entailed by proposed use softests " (American Educational Research Association [AERA], the American Psychological Association, & National Council on Measurement in Education, 1999).

Content domain representation is of central importance in test validation (Kartowagiran, 2013). Validity evidence based on test content is one of the five “sources of evidence that might be used in evaluating a proposed interpretation of test scores for particular purposes” set out in the *Standards for Educational and Psychological Testing* (Miller, Linn & Groulund, 2009). Other evidence is response processes, internal structure, relations to other variables, and testing consequences. Content validity describes “a judgement of how adequately a test samples behavior representative of the universe of behavior that the test was designed to sample” or in the other words an instrument should cover the content that supposed to be measured (Cohen & Swerdlik, 2005). In line with this, Haynes et al. said that the meaning of content validity is “the degree to which elements of an assessment instrument are relevant to and representative of the targeted construct for a particular assessment purpose” (Haynes, Richard, & Kubany, 1995). Adequate representation of the content measured is a fundamental requirement of a psychological and educational tests instrument. According to Anwar (2012), the representation of the content can be estimated by testing the feasibility or relevance of the tests content through rational analysis by a competent panel that commonly called expert judgment. Expert judgment is a formal process for eliciting judgments from subject-matter experts (SMEs) about the value of a decision-relevant quantity (Hammit & Zhang, 2012).

This article discusses the evidence based of the validity based on the test content of the creative thinking skills assessment that support conation aspect of prospective biology teacher. Our goals are to describe the content validity evidence by using the SMEs, methods for gathering content validity data, and method to improve the measurement accuracy of the content validity. This type of validity can increase belief of the reader toward the assessment that have been developed. By reading this article, the reader can understand how to collectand analyze validity evidence based on the test content to evaluate the use of a test for a particular purpose.

RESEARCH METHOD

To examine content validity in the judgment stage, professional subjective ratings is required evaluate the relevant construct of the assessment. A panel consist of six SMEs were asked to rate the content representation based on construction, relevance,and clarity of the assessment. Assessment instrument were evaluated in this study is the "creative thinking skills assessment that support the conation aspect of prospective biology teacher through divergent task, consist of 52 items divergent tasks. This assessment consists of variety cases which is the application of human physiology courses that support the conation idea aspects. Assessment content serve as the basis for the research data analysis and gathering validity evidence based on test content with the help of SMEs.

Data collection wasper formed using Likert-type rating scales validation sheet to measure the aspects of each item of the content domain the SMEs are being asked to consider. Each SME was given a booklet containing all of the 52 test items to provided the data for this study based on the construction, relevance and clarity for each tests item on the five point scale (Table 1). To see the consistency between the validator, content validity index is calculated by using Aiken index [**V**] (Aiken, 1985). Aiken index calculation based on the result of SMEs ratings as "$\boldsymbol{n}$" people towards an item in terms of the extent to which the test measures the constructs it purports to measure. Aiken's V formula is defined as $\boldsymbol{V = \Sigma s/[n(c-1)]}$. The "$\boldsymbol{s}$" value obtained from the rating given by SMEs ($\boldsymbol{r}$) substract the integer assigned to the lowest validity category(**lo**). While "$\boldsymbol{c}$" is the integer assigned to the highest validity category.

Table1. Criteria for Measuring Content Validity

a. Construction
   1 = poor
   2 = fair
   3 = average
   4 = good
   5 = very good
b. Relevance
   1 = not relevant
   2 = item need revision
   3 = item need some revision
   4 = relevant but need minor revision
   5 = very relevant
c. Clarity
   1 = not clear
   2 = item need revision
   3 = item need some revision
   4 = clear but need minor revision
   5 = very clear

The criteria to evaluate the content validity of the assessment were analyzed separately through rating scale. Furthermore, the content validity index calculated by using Aiken's V formula with the ≥ 0.79 criterion. When Aiken index value ≥ 0.79 and statistically significant, there is agreement among the SMEs that the item is relevant to the specific content area. When Aiken index value < 0.79 and statistically significant, there is agreement among the SMEs that the item is not highly relevant to the specific content area. Moderate values of the Aiken index signify poor agreement among the SMEs about the relevance of the item to it sprescribed content area.

Table2.Example of SMEs rating task assessing item construction

Item	Goal	How well does the item measure its construction?				
		1 (poor)	2 (fair)	3 (average)	4 (good)	5 (very good)
1	Collecting information about the actions that student do if they meet people who has cell structure and functiond isorder in the context of science, environment, technology, and society (salingtemas)					
2	Formulating preventive actions that is useful to overcome the people with carbohydrate metabolism disorder (diabetes mellitus) in the context of salingtemas.					
3	Designing actions taken when facing people with impaired pyruvate metabolism in the context of salingtemas.					
4	Designing useful actions to overcome the people with fat metabolism disorder in the context of salingtemas.					
5	Formulating useful solutions to overcome the people with protein metabolism disorder in the context of salingtemas.					
....						
52	Designing actions taken to help people with pregnancy disorder mechanism in the context of salingtemas					
Directions: Please read each item and its associated benchmark. Rate how well the item construction using the rating scale provided. Be sure to give a check list one rating for each item						

RESULT AND DISCUSSION

Using the rating scale approach we can get an idea of how well specific items,and the group of items measuring a specific objective, adequately measure the intended objective. an example of how the data summarized according to the criteria illustrated in Table3. Aiken index ranges from zero to one and essentially indicates the proportion of SMEs who rate the item andit can also be evaluated for statistical significance (Sirecci, 1995; Sirecci & Bond, 2014). An item content relevance analysis with 6 judges should yield a V coefficient equal to or above 0,79 to be statistically significant (Aiken, 1985). This value was taken from a right-tailed binomial probability table provided by Aiken.

Table 3. Example summary of Item Relevance Results

Item	Goal	Mean	Median	V
1	Collecting information about the actions that student do if they meet people who has cell structure and function disorder in the context of science, environment, technology, and society (salingtemas)	4,3	4,0	0,83*
2	Formulating preventive actions that is useful to overcome the people with carbohydrate metabolism (diabetes mellitus) in the context of salingtemas.	4,5	4,5	0,88*
3	Designing actions taken when facing people with impaired pyruvate metabolism in the context of salingtemas.	3,7	4,0	0,67
4	Designing useful actions to overcome the people with fat metabolism disorder in the context of salingtemas.	4,3	4,0	0,83*
5	Formulating useful solutions to overcome the people with protein metabolism disorder in the context of salingtemas.	4,7	5,0	0,92*
...				
52	Designing actions taken to help people with pregnancy disorder mechanism in the context of salingtemas	4,8	5,0	0,96*
	Average for Item Relevance	**4,5**	**4,7**	**0,88**

Notes: Statistics based on 6 SMEs and rating scale where 1= not relevant, 5 = very relevant.
* p<0.05

The analysis showed that creative thinking skills assessment that support the conation aspect of prospective biology teacher through divergent tasks, has agood representation related to the extent to which the items are relevant to the domains. For item construction criteria, Aiken validity index identifies four items (i.e., 3, 14, 31, and 40) had a lower index than other items (<0,79). Of 52 items, those with content validity index equal or over 0,79 remained and the rest four items were discarded resulting to 48-item scale (Table 4).

To improve the content validity measurement accuracy, Penfield and Giacobbi (2004) established aconfidence interval for item content-relevance ratings applied to Aiken index (Table 4). Index values obtained from Aiken (V) may deviate substantially, and thus it is requre to construct a confidence interval for the value of the confidence interval V. A confidence interval that is not constrained by large numbers of raters or Likert-scale categories. The Score interval out performs the traditional Wald interval formula $\overline{X} \pm t_{df}\left(s/\sqrt{n}\right)$, especially when there are five or fewer categories and twenty or fewer raters (Penfield &Miller, 2004). Therefore, using Penfield score confidence interval is the solution of obtaining a more accurate estimate of content validity when the number of SMEs is small (i.e., less than 20) or the number of Likert-scale categories is small (i,e., less than 5).

Table 4. Outcomes Values of Aiken's V, and 95% Score Confidence Interval for 52 Items of Item Construction Results

		95% CI				95% CI	
Item	V	Lower Limit	Upper Limit	Item	V	Lower Limit	Upper Limit
1	0,83*	0,64	0,93	27	0,83*	0,64	0,93
2	0,83*	0,64	0,93	28	0,88*	0,69	0,96
3	0,58	0,39	0,76	29	0,83*	0,64	0,93
4	0,83*	0,64	0,93	30	0,83*	0,64	0,93
5	0,79*	0,6	0,91	31	0,67	0,47	0,82
6	0,92*	0,74	0,98	32	0,83*	0,64	0,93
7	0,92*	0,74	0,98	33	0,79*	0,6	0,91
8	0,79*	0,6	0,91	34	0,83*	0,64	0,93
9	0,83*	0,64	0,93	35	0,83*	0,64	0,93
10	0,88*	0,69	0,96	36	0,96*	0,8	0,99
11	0,88*	0,69	0,96	37	0,88*	0,69	0,96
12	0,92*	0,74	0,98	38	0,83*	0,64	0,93
13	0,83*	0,64	0,93	39	0,88*	0,69	0,96
14	0,67	0,47	0,82	40	0,67	0,47	0,82
15	0,83*	0,64	0,93	41	0,88*	0,69	0,96
16	0,83*	0,64	0,93	42	0,92*	0,74	0,98
17	0,88*	0,69	0,96	43	0,83*	0,64	0,93
18	0,83*	0,64	0,93	44	0,88*	0,69	0,96
19	0,88*	0,69	0,96	45	0,96*	0,8	0,99
20	0,83*	0,64	0,93	46	0,83*	0,64	0,93
21	0,83*	0,64	0,93	47	0,79*	0,6	0,91
22	0,83*	0,64	0,93	48	0,83*	0,64	0,93
23	0,92*	0,74	0,98	49	0,96*	0,8	0,99
24	0,83*	0,64	0,93	50	0,83*	0,64	0,93
25	0,83*	0,64	0,93	51	0,83*	0,64	0,93
26	0,83*	0,64	0,93	52	0,88*	0,69	0,96

Note. The critical value of *V* according to Aiken's (1985) table of critical values is 0,79 undera Type I error rate of 0,05.The items for which the null hypothesis is rejected according to Aiken's critical value are noted with *. CI = confidence interval.

Table 4 displays the Aiken index of 6 SMEs with a score of 95% confidence intervals for item construction criteria. The typical length of the 95% score confidence interval for the data presented in Table 4 is approximately ± 0,30, although this value varies across the items depending on the specific value of Vfor the item. Using the typical length of the intervalof precision of Vas an estimator, a researcher may make statements concerning the adequacy of the precision of V. For example, in this study a researcher may set a criterion level of typical length of a 95% confidence interval equal to 0,30 to ensure the accuracy of V. If the typical length of the Score confidence interval exceeds this (as is the case with the example item 4, 14, 31, and 40 provided in Table 4), then the researcher may opt to examine the content of the items for potential lack of content-relevance, or increase the number of expert judges providing ratings for the items of the scale. Increasing the number of expert judges will act to increase the precision of V, and thus decrease the length of the confidence interval.

Table 5. Outcomes Values of Aiken's V, and 95% Score Confidence Interval for 52 Items of Item Relevance Results

Item	V	95% CI		Item	V	95% CI	
		Lower Limit	Upper Limit			Lower Limit	Upper Limit
1	0,83*	0,64	0,93	27	0,92*	0,74	0,98
2	0,88*	0,69	0,96	28	0,83*	0,64	0,93
3	0,67	0,47	0,82	29	0,88*	0,69	0,96
4	0,83*	0,64	0,93	30	0,88*	0,69	0,96
5	0,92*	0,74	0,98	31	0,63	0,43	0,79
6	0,92*	0,74	0,98	32	0,92*	0,74	0,98
7	0,92*	0,74	0,98	33	0,88*	0,69	0,96
8	0,88*	0,69	0,96	34	0,83*	0,64	0,93
9	0,92*	0,74	0,98	35	0,88*	0,69	0,96
10	0,83*	0,64	0,93	36	0,79*	0,60	0,91
11	0,92*	0,74	0,98	37	0,88*	0,69	0,96
12	0,92*	0,74	0,98	38	0,92*	0,74	0,98
13	0,92*	0,74	0,98	39	0,96*	0,80	0,99
14	0,67	0,47	0,82	40	0,58	0,39	0,76
15	0,92*	0,74	0,98	41	0,92*	0,74	0,98
16	0,92*	0,74	0,98	42	0,96*	0,80	0,99
17	0,83*	0,64	0,93	43	0,88*	0,69	0,96
18	0,88*	0,69	0,96	44	0,92*	0,74	0,98
19	0,92*	0,74	0,98	45	0,96*	0,80	0,99
20	0,92*	0,74	0,98	46	0,92*	0,74	0,98
21	0,92*	0,74	0,98	47	0,96*	0,80	0,99
22	0,92*	0,74	0,98	48	0,92*	0,74	0,98
23	0,79*	0,60	0,91	49	0,92*	0,74	0,98
24	0,88*	0,69	0,96	50	0,96*	0,80	0,99
25	0,88*	0,69	0,96	51	0,92*	0,74	0,98
26	0,88*	0,69	0,96	52	0,96*	0,80	0,99

*** p<0.05**

The analysis shows most items have Aiken index value equal to or more than 0,79, but not on all items. There are four items that have Aiken index value below 0,79 (items 3, 14, 31, and 40). Not only in item construction category, but also on the relevance category (Table 5), and the item clarity category (Table 6). These four items were discarded, the number of items decreased from 52 to 48 items. A content validity coefficient (V value) of ≥ 0,79 indicates significant standard has been reached. Content validity assessment for each item show that the 48 items possess good content validity, indicating the items are effective measurement tool.

Table 6. Outcomes Values of Aiken's V, and 95% Score Confidence Interval for 52 Items of Item Clarity Results

Item	V	95% CI		Item	V	95% CI	
		Lower Limit	Upper Limit			Lower Limit	Upper Limit
1	0,83*	0,64	0,93	27	0,92*	0,74	0,98
2	0,83*	0,64	0,93	28	0,83*	0,64	0,93
3	0,58	0,39	0,76	29	0,83*	0,64	0,93
4	0,79*	0,60	0,91	30	0,83*	0,64	0,93
5	0,83*	0,64	0,93	31	0,71	0,51	0,85
6	0,83*	0,64	0,93	32	0,92*	0,74	0,98
7	0,96*	0,80	0,99	33	0,88*	0,69	0,96
8	0,92*	0,74	0,98	34	0,83*	0,64	0,93
9	0,88*	0,69	0,96	35	0,83*	0,64	0,93
10	0,83*	0,64	0,93	36	0,83*	0,64	0,93
11	0,83*	0,64	0,93	37	0,88*	0,69	0,96
12	0,83*	0,64	0,93	38	0,92*	0,74	0,98
13	0,92*	0,74	0,98	39	0,96*	0,80	0,99
14	0,63	0,43	0,79	40	0,67	0,47	0,82
15	0,92*	0,74	0,98	41	0,88*	0,69	0,96
16	0,92*	0,74	0,98	42	0,92*	0,74	0,98
17	0,83*	0,64	0,93	43	0,88*	0,69	0,96
18	0,88*	0,69	0,96	44	0,88*	0,69	0,96
19	0,83*	0,64	0,93	45	0,88*	0,69	0,96
20	0,92*	0,74	0,98	46	0,92*	0,74	0,98
21	0,88*	0,69	0,96	47	0,88*	0,69	0,96
22	0,83*	0,64	0,93	48	0,96*	0,80	0,99
23	0,88*	0,69	0,96	49	0,88*	0,69	0,96
24	0,88*	0,69	0,96	50	0,83*	0,64	0,93
25	0,83*	0,64	0,93	51	0,83*	0,64	0,93
26	0,88*	0,69	0,96	52	0,96*	0,80	0,99

*** p<0.05**

CONCLUSION AND SUGGESTION

Validity in educational testing refers to the measurement accuracy, it is a scale or measurement tool used to precisely measure characteristics for testing. Higher scale validity index indicates that the scale results can better present the actual characteristics of the tested subjects. Validity discussion whichis presented in this paper focuses on issue related to evaluating content validity. Content validity evaluation requires reputable SMEsto examine whether the test items assessing defined content (Lawshe, 1975). Content domain representation is critical for demonstrating the validity of inferences derived from test scores (Sireci, 1995). All inferences derived from test scores are valid only to the extent to which the test measures the constructs it purports to measure. The procedures studied in this paper will help test developers evaluate fundamental attributes of content representation.

Overall, of the 52 items, there are 48 items of the creative thinking skills assessment that support the conation aspect of biology prospective teachers through divergent tasks that has adequate content validity index value. Although documenting content validity of an instrument through SMEs may seem expensive in terms of time and human resources, but it is importance warrants greater attention when a valid assessment instrument is to be developed. Content validity is an important factor in identifying the concept of measuring, however, it is not a sufficient indication that the instrument actually measures what is that intended to measure. Finding from content validity could contribute to support the construct validity of an instrument. Based on this content validity study, the reader can understand the process of measuring content validity. By measuring content validity, the interpretations of results are precise.

REFERENCES

Aiken, L.R. (1980). Content validity and reliability of single items or questionnaires. *Educational and Psychological Measurement, 40*, 955-959.

Aiken, L.R. *(*1985*)*. Three coefficientsfor analyzing the reliability and validty of ratings. *Educational and Psychological Measurement*, 45, 131-142.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards for educational and psychological testing*. Washington, D.C.: American Educational Research Association.

Azwar, S. (2012). *Reliabilitas* dan *validitas*. Yogyakarta: Pustaka Pelajar

Cohen, R. J., Swerdlik, M. E., & Sturman, E. D. (2013). Psychological testing and assessment: An introduction to tests and measurement (6th ed.). New York: McGraw-Hill.

Darmawan, E. (2013). Pengaruh PBL terhadap sikap dan hasil belajar. *Jurnal Lentera Sains*, Vol 3, No. 2. p. 1-4.

Hammitt, J.K. & Zhang, Y. (2013). Combining experts' judgments: Comparison of algorithmic methods. *Risk Analysis* 33(1): 109-120.

Haynes, S. N., Richard, D. C. S., & Kubany, E. S. (1995). Content validity in psychological assessment: A functional approach to concepts and methods. *Psychological Assessment*. 7, 238-247

Isaksen, S. G., Dorval, K. B., & Treffinger, D. J. (1994). *Creative approaches to problem solving.* Dubuque, IA: Kendall/Hun. Google Book. http://books.google.co.id/books/about/Creative_Problem_Solving.html?id=qLZ3nMUPsyYC&redir_esc=y

Jo, S. M. (2009). *A study of korean students' creativity in science using structural equation modeling*. (Doctoral Thesis, Department of Special Education Rehabilitation, The University of Arizona, USA). [Electronic Version]

Kartowagiran, B. (2008). *Dimensional validity of mathematics test in the national exam for junior secondary schools (SMP) 2003-2006. Jurnal Penelitian dan Evaluasi Pendidikan*. Nomor 2, Tahun XII.

Lawshe, C. H. (1975). A Quantitative Approach to Content Validity. *Personnel Psychology*, (28), 563-575.

Lubart, T. (2004). Individual student differences and creativity for quality education. *Background paper prepared for the Education for All Global Monitoring Report 2005 The Quality Imperative*. UNESCO.

Miller, M. D., Linn, R. L., & Gronlund, N. E. (2009). *Measurement and assessment in Teaching*. 10th Edition. Upper Saddle River, NJ: Pearson Education, Inc.

Penfield, R. D., & Giacobbi, P. R. Jr. (2004) Applying a score confidence interval to aiken's item content-relevance index. *Measurement in Physical Education and Exercise Science*, 8:4, 213-225, DOI: 10.1207/s15327841mpee0804_3

Penfield, R. D., & Miller, J. M. (2004). Improving content validation studies using an asymmetric confidence interval for the mean of expert ratings. *Applied Measurement in Education,17*(4), 359–370.

Poole, M. S., & Van de Ven, A. H. (2004). *Alternative approaches for studying organizational change*. Presented at the First Organization Studies Summer Workshop on Theorizing Process in Organizational Research, Santorini, Greece, 12&13 June, 2005.

Riyanti, D. B. P. & Prabowo, H. (1998). *Seri diktat kuliah psikologi umum 2*. Depok: Universitas Gunadarma.

Reynolds, C. R., Livingston, R. B., Willson, V. (2009). *Measurement and assessment in education*. Pearson Education Inc. Upper Saddle River, New Jersey.

Sireci, S. G. (1995). *The central role of content representation in test validity*. Paper presented at the Annual Meeting of the National Council on Measurement in Education. San Francisco.

Sireci, S. & Bond, M. F. (2014). Validity evidence based on test content. *Psicothema*. Vol. 26, No. 1, 100-107.

Stelly, D. J., & Goldstein, H. W. (2007). *Application of content validation methods to broader constructs*. In S. M. McPhail (Ed.). *Alternative validation strategies: developing new and leveraging existing validity evidence*. San Francisco, CA: Jossey-Bass A Wiley Imprint

Subali, B. (2011). Pengukuran kreativitas keterampilan proses sains dalam konteks assessment. *Jurnal Cakrawala Pendidikan*, Februari 2011, Th. XXX, No. 1. p. 130-144. [Electronic Version]