Research and Evaluation in Education
Volume 2, Number 1, June 2016 (pages 1-12)
Available online at: http://journal.uny.ac.id/index.php/reid

A SEPARATION INDEX AND FIT ITEMS OF CREATIVE THINKING SKILLS ASSESSMENT

[1]*Andi Ulfa Tenri Pada;* [2]*Badrun Kartowagiran;* [3]*Bambang Subali*
[1]Syiah Kuala University; [2,3]Yogyakarta State University
[1]andi_ulfa@unsyiah.ac.id; [2]badrun_kartowagiran@uny.ac.id; [3]b_subali@yahoo.co.id

Abstract

This article discusses the evaluation results of the separation index and fit item of creative thinking skills assessment that supports the conation aspect of prospective biology teachers in Aceh. This assessment consists of 37 items of divergent tasks, which is the application of human physiology courses that support the conation aspects. The participants were selected from the Biology Education Program, Faculty of Teacher Training and Education, Syiah Kuala University. The data were analyzed using the Quest software including the separation index and fit item. The results indicate that the creative thinking skills assessment instrument that supports the conation aspect of prospective biology teachers has a good separation index and all the items fit PCM-1PL.

Keywords: *PCM-1PL, separation index, fit item, creative thinking skills, conation aspect*

How to cite item:
Pada, A., Kartowagiran, B., & Subali, B. (2016). Separation index and fit items of creative thinking skills assessment. *Research and Evaluation in Education, 2*(1), 1-12. doi: http://dx.doi.org/10.21831/reid.v2i1.8260

Introduction

Formal education in Indonesia todays generally shows fewer opportunities for the development of creativity. School prioritizes cognitive training only at the knowledge, memory and reasoning levels. This is proven by the teaching process in schools, where there is hardly any activity demanding creative thinking. Thus, students are stimulated to think, act, and behave creatively (Supardi, 2012, p.6). This statement is similar to Subali (2011, p.139), who insists that the creativity of science process skills is less developed by school teachers. The majority of biology teachers suggest that they are more concentrated in multiple-choice tests that are clearly oriented on the development of convergent thinking patterns, and are less oriented to the divergent patterns as the basis for creativity development.

The importance of creativity is stated in Article 3 of National Education System Act No. 20 of 2003, on the national education goals with the expectation that education can develop students' potentials in order to become pious, noble, skilled, creative, and independent human beings. Meanwhile, the goal of Indonesian national education emphasizes the importance of creativity. However, it is very much in contrast with the achievement of Indonesia in international creativity survey (The Global Creativity Index) in 2011. Indonesia currently ranks 81 of the 82 countries involved in the survey, far below the neighboring countries, Singapore, that ranks 9, and Malaysia that ranks 48 (Florida, Mellander, & Stolarick, 2011, p.41).

The results of the international surveys show that the performance of Indonesian people is included in 'low' category. The research which is conducted by Ramirez and Ganaden (2008, pp.22-33) reveals that the poor performance is due to the weakness in high-level thinking skills. The learning process in higher education today seems less than effective to improve the ability of thinking creatively. Despite this, creative thinking is the culmination of the cognitive dimension on the revision of Bloom's taxonomy by Anderson *et al.* (Krathwohl, 2002, p.215) and also New Bloom's taxonomy (Dettmer, 2006, p.73). DeHaan (2009, pp.172–181) indicates that the abilities to think creatively help the students find the value of evidence-based reasoning, increase high order cognitive skills (HOCS), and make them capable of solving problems.

Creative thinking is a process which is employed to yield ideas or brand new ideas (Runco, 2004, p.658). New ideas can result from combination (elaboration) of old ideas or newly emerging ideas. It may occur by combining the ideas of others to stimulate the rise of brand new ideas. Creativity is an ability to generate new ideas or new artifacts, which are surprising and valuable (Boden, 2001, p.95). The research results also show that creativity is an essential element of a problem solving (Mumford, Mobley, Uhlman, Reiter-Pamon, & Doares, 1991, pp. 91-122; Runco, 2004, pp.658-659). Thereby, it is normal if the creativity and intelligence are deemed the application of creativity among HOCS as described in Bloom's Taxonomy (Crowe, Dirks, & Wenderoth, 2008, pp. 368-381).

The creative thinking skill is one of the thinking skill dimensions that should be further developed and measured. In the opinion of Baer (2012, pp.1102-1119), the study on the creativity is complex. Nonetheless, it is not merely difficult to measure or perform (DeHaan, 2009, pp.172-181). The measurement of creative thinking skill ability of the students can be conducted by creating an assessment with divergent approach (Subali, 2011, pp.130-144). The divergent thinking process is a part of the creative capability. Divergent thinking is an ability to construct or generate sets of possible responses, ideas, options or alternatives to a problem (Isaksen, Dorval, & Treffinger, 1994, p.18). Therefore, the divergent thinking can be defined as the ability to deliver wide range of solutions to the problems with the proper procedures and reasons.

The characteristic of creativity is the uniqueness and originality that should be initiated with a search for various possible solutions. Afterwards, a person should know whether the solution is different from other solutions, and whether the solution has never existed before. In order to find various alternative solutions, one requires divergent thinking skills (Subali & Suyata, 2013, p.4). Creative thinking is the cognitive activities that can lead creative productions deemed useful and new to the groups or individuals (Isaksen, Dorval, & Treffinger, 1994, p. 31). Thereby, in this article creative thinking skill means the ability to construct an idea into a unique pattern or structure. It puts the priority on the element of originality in the idea formed, related to the problems identified.

Beside cognitive aspects, the creative thinking skill which developes in learners is inseparable from the conation aspect (Lubart, 2004, p.10; Poole & Van de Ven, 2004, p.41; Jo, 2009, p.86). Nowadays, the conation aspect is often ignored by most educators; not only at the level of primary and secondary education, but also at the level of university (Reeves, 2006, p.297). Hence, if the cognitive aspect is related to the idea, the connative aspect is associated with the concept of intrinsic motivation and willingness.

Conation as a mental process is to activate and/or guide the behavior and actions (Huitt & Cainn, 2005, pp.1-7). In the opinion of Pepper (1970, p.337), conation is illustrated as 'a drive-charged pattern of references positive or negative'. A variety of terms are used to represent aspects of conation, including the intention or tendency to behave (Riyanti & Prabowo, 1998, p.70; Board of National Education Standard, 2010, p.28). The connative performance is actions, willingness, or desire. Conation is a state where the mind has a purpose, and connative knowledge is to select or be willing to do an act in relation to a series of circumstances. It can be concluded that conation is a statement of desire which has a positive and negative direction.

According to Darmawan (2013, pp.1-4), the students who already have a fairly good concept of understanding do not necessarily apply their knowledge in the real world. By the time the students learned about the circulatory and the respiratory system, they should have already recognized the health impacts of smoking on heart and lungs, yet they are still on it. On the other hand, the concept understanding in the mind of the learners may also generate constructive actions that can contribute to character development. They are benefited by gaining more awareness on the value contained in these materials by quitting smoking or reminding their peers to stop smoking. Hence it is obvious that the cognitive factors getting involved in the creative process can be supported or inhibited by the willingness or conation factors.

By taking into account the problem's root, we need to think about the ways to overcome it. Moreover, the implementation of a competency-based curriculum at universities is focused on training how to think and reason, developing creative activities, developing the abilities to solve problems, and communicating ideas. One effort that has been done is to develop assessment tools to measure creative thinking skills supporting the conation aspect of the students through a divergent pattern in the course of human physiology. In order to prove whether the assessment has been constructed optimally, it is necessary to evaluate the quality of the assessment tools.

In order to obtain full information on the ability of creative thinking skills of students as the prospective biology teachers in the subject of human physiology, the information shall be collected at the end of learning. The expectation is that the results of the assessment does not only serve as the implications of the measurement result but also improve the thinking abilities of the students associated with the materials that they have learned, as well as provide information for classes and educators to improve the quality of teaching and learning process. This description illustrates the importance of the development of an assessment model to observe the attainment of thinking skills that support the conation aspect of prospective biology teachers in the subject of human physiology. Therefore, the assessment model used shall be able to support the attainment of the course objectives.

The selection of Human Physiology course is based on the consideration that with the course, the students' response to the conation idea response can be obtained more easily as the cases which are discussed are contextual. When studying human physiology course, students learn the normal function of organs, thereby in these instruments, the stimulus which is provided is in the form of disruption to the function of organs (disease) or the inverse of the normal body functions. What is expected is that, through such abnormal condition stimulus, the students can provide relevant solutions to various cases being presented.

The students were asked to provide a variety of responses in the form of a divergent production pattern to numerous cases presented through the concepts of human physiology materials that are considered essential. These responses are later capable of describing a certain tendency of behavior in accordance with the attitude of a person. In other words, they can describe a person's tendency to react against a stimulus in certain ways based on their understanding after studying human physiology.

The two aspects used in assessing the quality of good assessment tools are validity and reliability (Cohen, Swerdlik, & Sturman, 2013, p.98). In line with Cohen, Swerdlik, and Sturman's opinion, Reynolds, Livingston, and Willson (2009, p.4) mention the characteristics of tests include reliability and validity. The test users should seriously consider the use of the test results. The tests employed are only those which generate valid, reliable, and accurate evidence on the purposes they serve and for whom they are intended. Therefore, prior to using assessment instrument, the evaluation of the validity and reliability is necessary to conduct.

According to Reynolds, Livingston, and Willson (2009, p.4), the reliability of a test refers to the stability and consistency of the test scores, while validity refers to the accuracy of the interpretation of test scores. Wright and Stone (1999, pp.157-165) mention that reliability is a statement on the consistency and stability of scores of an instrument, while the validity is a statement of conformity of the test and its components, the truth of the test results and its interpretation. Based on several opinions explained before, it can be argued that good tests are those having reliable and valid condition or characteristics (Mardapi & Kartowagiran, 2011, p.332).

One technique that can be used to analyze the validity and reliability of test instruments is Item Response Theory (IRT). IRT is an alternative measurment method other than Classical Test Theory (CTT) (Gorin & Embretson 2006, pp.394-411). CTT is the psychometric technique which is allowing the presumption of test results, for example the item difficulties and individual talent (Alagumalai, Hungi, & Curtis, 2005, p.273). Meanwhile, IRT is a psychometric technique focusing on individual response towards specific test items influenced by the quality of the item.

IRT is a probabilistic model which is seeking to describe a person's response to an item (Hambleton, Swaminathan, & Rogers, 1991, p.9). In the simple form, IRT argues that the possibility of random people 'j' with the ability 'θj' to answer a random item 'i' with a degree of difficulty 'b', being conditioned on the ability of people and item difficulties. In other words, if a person has high ability in a specific field, he will probably answer the easy items correctly. In contrast, if a person has low ability and gets difficult items, he will perhaps answer the item wrongly.

IRT is made as an alternative model by psychometric experts to overcome the weaknesses of CTT. This model has the following properties: (1) The characteristics of the item are not dependent on the group of test participants subjected to the test item, (2) the scores which are stating the ability of test participants do not depend on the test, (3) the model is expressed in rank (level) of items, not in the level of the tests, (4) the level model does not require a parallel test to calculate the reliability coefficient, and (5) the model provides the proper measure for each ability score (Hambleton, Swaminathan, & Rogers, 1991, p.5).

There are two basic postulates of modern test theory (Hambleton, Swaminathan, & Rogers, 1991, p.7): (1) The performance of the test participants on an item can be predicted (described) by using a set of factors called properties, latent properties, or ability; (2) the relationship between the performance of test participants on a test item with the underlying characteristics can be described by a steadily increasing function, which is referred to as item characteristic function, or item characteristic curve. Such a function explains that if the ability level increases, the probability of a test to respond correctly to an item will also increase.

There are several assumptions in the item response theory model of Hambleton, Swaminathan and Rogers (1991, pp.9-12): (1) It is one-dimensional (unidimensional). This assumption is highly difficult to fulfill due to the factors affecting tests, such as cognitive, personality, and language factors. However, the most important point of this assumption is one component that is considered to be dominant in determining the abilities of the subject. According to Hutten (Hattie, 1985, p.146), the unidimensionality can be investigated through Eigen value in the factor analysis. The percentage of the total variance explained by the first component is commonly regarded as unidimensionality index. The higher the percentage of the main component total variance, the closer this test to unidimensional character. Reckase (1979, p.228) recommends that for a good calibration, the total percentage of variance explained by the first com-ponent, i.e. 20% or more is required by data to fulfill the unidimensional assumption. (2) It is locally independent. Such an assumption means that the test participants' response towards an item is not related to other items within the test.

The package program employed to perform item analysis in this study is QUEST. A central element of QUEST program is Rasch Model (RM). The program can use the response data scored in a politomus manner. The QUEST program is able to estimate the parameters, both for items and testee (case/ person) using unconditional (UCON) or joint maximum likelihood (Adam & Khoo, 1996, p.89).

In IRT, the instrument is declared valid when an item behaves consistently (fits) with what is expected by the model. The term 'valid' in IRT is used to assess the success of calibration in the effort to find out the data fitness with the model. An item is declared fit with the model when the calibration is 'valid' and when the testee (case/person) is declared fit with the model, thus the measurement shall be 'valid' (Wright & Stone, 1999, pp.169-171). The item and person fit resulted from the analysis of the Quest program is based on the average value of infit Mean Square (INFIT MNSQ) from 0.7 to 1.3 (Wright & Masters, 1982, p.100; Bond & Fox, 2001, pp.177-178)

The criteria for fit person through the analysis using the QUEST program is based on the average size of INFIT Mean of Square (INFIT MNSQ) of a person is equal to 1. Another criterion is that the expected mean value of INFIT t is equal to 0 with variance equal to 1. The determination of a fit item with the model is based on the value of INFIT MNSQ or the INFIT t value of the item. The expected value of INFIT MNSQ value is equal to 1 with a variance equal to 0, and the expected value of INFIT t is equal to 0 with a variance equal to 1 (Adam & Khoo, 1996, p.93).

In IRT, the precision test is conceptualized as something referred to as information, depending on the characteristic level being measured. The estimation of the internal consistency reliability of a test is based on the person separation reliability. Logit scale estimation is used for each testee to calculate the reliability (Bhakta, Tennant, Horton, Lawton, & Andrich, 2005, pp.1-13). A person separation reliability (R_p) can be calculated using the following formula:

$$R_p = \frac{SD_p^2 - MSE_p}{SD_p^2}$$

Where,

SD_p^2 : is observed variance of testee

MSE_p : is the mean squared error of measurement.

According to Wright and Masters (Mappiasse, 2006, p.584), using the Rasch model, item separation reliability and person separation reliability can be estimated as well. The interpretation of person separation reliability also encounters problems when an item fails to define a single variable leading to the use of alternative index which is called person separation index.

The person separation index is an estimation of how well each testee can be distinguished on the measured variables. It describes the placement repetition of a testee against other items, measuring the same construct (Mappiasse 2006, p.585; Curtis & Boman, 2007, p.251). The higher the person separation index (G_p), the more consistent each item is used to measure the respective testee. According to Wright and Stone (1999, p.163), the value $G_p = 2$ is equivalent to the R_p value of 0.80. The following formula is presented to calculate the person separation index:

$$G_p = \sqrt{\frac{R_p}{1 - R_p}}$$

Such a concept provides an estimation of sample standard deviation in standard error units. This index is useful to compare the use of different scales in an entire different classroom situation (Mappiasse, 2006, p.585). It is also applicable in the item separation reliability and item separation index. The consistency of a group of individuals in providing information on item difficulty forming the scale is reflected in the item separation index (Curtis & Boman, 2007, p.251). The higher the estimation of an item separation index, the more precise the whole items being analyzed according to the model used (Subali, 2010, p.38).

This article discusses the evidence of the validity and reliability of assessment instruments in creative thinking skills using the item response theory through the Partial Credit Model (PCM). In this analysis, there are two main things observed: Fit item for instrument validity testing and Pearson and Item Separation Index. The analysis results are later used to determine the quality of the test instrument.

Method

In order to evaluate Person Separation and item fit of assessment tools, empirical data are required. The data from the test product were analyzed using the Quest program. The employment of this program was based on the consideration that the logistic model chosen to estimate the item parameter and ability parameter of participants was Rasch model development or one parameter logistic model (1-PL), and for polytomous scoring technique was Partial Credit Model (PCM).

In this research, one parameter logistic model used PCM of the Quest program. Model of IRT 1PL or Rasch Model (RM) is a central element of the Quest program, using the joint maximum likelihood procedure to estimate items and case parameters (Adams & Khoo, 1996, p.89). PCM is developed from RM, where the RM is used on a dichotomous score data, whilst PCM is used in the polytomous score data (more than two categories) (Masters & Wright, 1997, p.100). In this model, it is assumed that the parameter of item difficulty level is the only item characteristics affecting the response characteristics of the test participants (Nering & Ostini, 2010, p.121).

The trial subjects were 218 students at the initial trial and 270 students at the main trial. The criteria were the students who had attended the teaching process of Human Physiology course. The test instruments were distributed to students at the end of the teaching process in two periods. The students were given two hours in each period to complete all test items. The test results were then employed as the data in this study.

The assessment instruments which were evaluated in this study were the assessment instrument of creative thinking skills which was supporting the conation aspect of prospective biology teachers through the divergent approach. The assessment instrument consisted of 37 items which were grouped into four components. These components consisted of: (1) the alternative solution component which was the the ability to generate a number of solutions to respond to an issue, which is consisting of 10 items; (2) the original solution components, i.e. the ability to generate a number of relevant solutions that are unique or unusual, that also consisted of eight items; (3) feasibility solution component, i.e. the ability to yield a number of effective solutions which are applicable for resolving the case given, that is consisting of 10 items; and (4) variation solution components, such as the ability to produce a number of categories of solutions, which is consisting of nine items. Items in this instrument consisted of a variety of cases which were an application of human physiology course which supported the conation aspect.

Responses were collected through four components of creative thinking skills, namely: (1) alternative solution or fluency which was produced in generating ideas, which could be observed through a number of relevant solutions resulted; (2) the original solution, such as the ability to generate a number of relevant solutions that are unique or unusual that can be observed through the frequency of testee's response. The score of the testee was calculated based on the response frequency given. The response which was less than 10% of the total testee was given a score of 4; lower than 25% was scored 3; lower than 50% was scored 2, and more than 50% was scored 1 (Diakidoy & Constantinou, 2010, p.405); (3) the feasibility solution, which was an effective solution to resolve the cases given, observable through a number of appropriate/proper responses; and (4) variation solution, such as the ability to produce a variety of categories with numerous solutions that could be observed from the number of relevant response categories with different types from the testee.

Findings and Discussion

Before analyzing IRT using PCM through the Quest program, the researcher tested the assumptions in advance. The first assumption is unidimensional. It can be proven using the factor analysis in order to view Eigen value of the inter-item covariance matrix (Hambleton & Rovinelli, 1986, pp.293-294). The second assumption is local independence. This assumption has been automatically proven after evidenced with unidimensionality of participants' data responses to a test (McDonald, 1981, p.101).

In the preliminary field testing, assumption testing is done at the data analysis stage using factor analysis which shows the largest Eigen value is 7.743; with the variation explained of 20.926% > 20%. This means that the assessment instrument developed is one-dimensional or unidimensional. With the proven unidimensional assumption, the local independence assumption is then automatically proven (Embretson & Reise, 2000 p.48). Thereby, the IRT analysis using PCM through the Quest program is feasible. The measurement data analysis results through the polytomous technique with five categories provide the results presented in Table 1.

Table 1. Summary of Pearson and Item/Cases Estimation using PCM

Criteria	Statistic Information	Estimation Results
Pearson Estimation/Case	Fit Statistics (Rerata Infit Mean Square)	1.00
	Standar Deviasi Infit MNSQ	0.21
	Separation Reliability	0.87
	Separation Index	2.58
	Zero Score	0
	Perfect Score	0
Item Estimation	Fit Statistics (RerataInfit Mean Square)	1.00
	Standar Deviasi Infit MNSQ	0.10
	Separation Reliability	0.72
	Separation Index	1.60
	Zero Score	0
	Perfect Score	0
Internal Consistency (CTT)		0.87

*$p < 0.05$, Item = 37 and *Cases* = 218

When an item fits in the sense that the item behaves consistently with what is expected by the IRT model, the instrument is declared valid (Wright & Stone, 1999, p.169-171). The term 'valid' in IRT is used to assess the success of calibration in an effort to find out that the data fit with the model. Table 1 shows that the entire items in the model are declared fit with the model for fulfilling statistics fit requirements that are obliged under the QUEST program. An item is declared fit to the model if it has an average infit Mean of Square (INFIT MNSQ) approaching 1 (Adams & Khoo, 1996, pp.24-25). Therefore, all the items analyzed are declared fit by the model with a standard deviation of 0.10.

In IRT, the estimation of internal consistency reliability of a test is based on the person separation reliability, where the estimation on a logit scale for each person is used to calculate reliability (Bhakta, Tennant, Horton, Lawton, & Andrich, 2005, pp.1-13). In other words, the value of test reliability is based on the error of measurement, presented in person/case; in this case it reaches 0.87. This means that the assessment instrument developed has a good reliability.

In addition to Person Separation Reliability (R_p), the reliability of a test can also be seen through Person Separation Index (G_p) which is an estimation on how well each testee can be distinguished on the measured variables. If the person separation reliability value ranges from 0 to 1, then it will be in contrast to the person separation index, which is not tied to a range of values from 0 to 1. The index quantifies reliability with a simple and direct manner, as well as having clear interpretation. The person separation index value (2.58) in Table 1 is classified as good. This is in line with the opinion of Wright and Stone (1999, p.163), that the value of G_p = 2 is equivalent to the value of R_p of 0.80.

The Quest program output also generates the item reliability analysis using the classical approach. In accordance with the reliability calculation using IRT, reliability calculation that is based on the internal consistency value of 0.87 shows that the test developed is qualified as a good test.

In the main field testing, the IRT analysis using PCM model through the Quest program is preceded by the unidimensional assumption and local independence tests. The unidimensional assumption test results of the instruments based on the factor analysis result can be seen through the Eigen value which is obtained at each factor. In this major field testing, the test result shows that Eigen value prior to rotation is 14.200, with the explainable variation of 38.378% > 20%, which means that the measuring instrument developed is unidimensional. It is proven with a unidimensional assumption showing a local independence assumption which is automatically proven. Therefore, the IRT analysis using PCM through the Quest program can be done.

Table 2. Summary of the comparison of main testing estimation by employing PCM

Criteria	Statistic Information	Estimation Results
Person Estimation/Case	Fit Statistics (rerata Infit Mean Square)	0.99
	Standar Deviasi Infit MNSQ	0.32
	Separation Reliability	0.94
	Separation Index	3.95
	Zero Score	0
	Perfect Score	0
Item Estimation	Fit Statistics (rerataInfit Mean Square)	1.00
	Standar Deviasi Infit MNSQ	0.38
	Separation Reliability	0.80
	Separation Index	2.00
	Zero Score	0
	Perfect Score	0
Konsistensi Internal (CTT)		0.94

*p < 0.05, *Item* = 37 and *Cases* = 270

The measurement data analysis result via five categories of polytomous scoring technique provides the results presented in Table 2. The analysis results of creative thinking skill tests that support the conation aspect using the PCM model are based on the value of infit Mean of Square (INFIT MNSQ) from 0.70 to 1.30 (Wright & Masters, 1982, p.100; Bond & Fox, 2001, p.230). The mean of INFIT MNSQ of 1 and a standard deviation of 0.38 indicates that the data fit with the model. Therefore, all items of the assessment instruments of creative thinking skills supporting the conation aspect are declared valid.

The estimation of separation reliability is based on the error of measurement presented in person/case. In the main field testing, person separation reliability value reaches 0.94, which means that the instrument assessment developed has good reliability. The separation reliability value also reports the data quality. The person separation reliability is used to classify people. The person separation value that is low (<2 person reliability <0.8) with the relevant people sample indicates the possibility that the instrument is not sensitive enough to distinguish the test participants with high abilities and low ability. Larger items may be needed. Item separation reliability is used to verify the hierarchy of items.

Table 2 shows a good person separation index value (G_p) of 3.95. The higher the value of person separation index, the more consistent each measuring item is used to measure the testee concerned (Mappiasse, 2006, p.585; Curtis & Boman, 2007, p.251). The low item separation value indicates that the person sample is not large enough to confirm the hierarchy of item difficulty level of the instruments (Linacre, 2015, p.656).

It also applies to the item separation reliability, and item separation index, i.e. 0.80 and 2.00. The consistency in a group of individuals in providing information on the item difficulty forming the scale is reflected in the item separation index (Curtis & Boman, 2007, p.251). The higher the index estimation, the more precise the entire item separation analyzes according to the model used (Subali, 2010, p.38).

The output of the Quest program for item reliability by using the classical approach is also presented here. In line with the person separation reliability, the reliability calculation based on the value of internal consistency of 0.94 in the main field testing suggests that the tests which are developed are qualified as good tests.

Conclusion and Recommendations

Conclusion

Based on the results and discussions, several conclusions can be drawn as follows. (1) All the items in the assessment instruments of creative thinking skills are declared fit with the model. (2) The estimation of person separation reliability shows a good reliability coefficient of 0.94. This coefficient can be used to calculate the person separation index of 3.95. (3) All items in the developed assessment instruments are qualified as creative thinking skill assessment instruments supporting the conation aspect of prospective biology teachers.

Suggestions

Based on the results, it is suggested that further research employ 2PL or 3PL data analysis for the polytomous type data. The findings in this article are able to contribute in favor of instruments' validity and reliability. Through this article, the readers can understand the estimation process on validity and reliability using the item response theory. Through validity and reliability coefficient tests, the measurement results can be interpreted more precisely.

References

Adams, R.J. & Khoo, S. (1996). *Acer quest (2.1).* Camberwell, Victoria: Australian Council for Educational Research.

Alagumalai, S., Hungi, N., & Curtis, D.D. (2005). *Applied Rasch measurement: A book of exemplars - papers in honour of John P. Keeves.* Dordrecht: Springer.

Baer, M. (2012). Putting creativity to work: The implementation of creative ideas in organizations. *Academy of Management Journal*, 55(5), 1102-1119.

Bhakta, B., Tennant, A., Horton, M., Lawton, G., & Andrich, D. (2005). Using item response theory to explore the psychometric properties of extended matching questions examination in undergraduate medical education. *BMC Medical Education*, 5(9). Doi: 10.1186/1472-6920-5-9.

Board of National Education Standard. (2010). *Panduan penulisan butir soal* [Handbook of writing questions item]. A material of technical guidance on school-based curriculum and 2010-standardized questions]. Jakarta: Badan Standar Nasional Pendidikan.

Boden, M.A. (2001). Creativity and knowledge. In A. Craft, B. Jeffrey, & M. Leibling (Eds.), *Creativity in education*. London: Continuum.

Bond, T.G., & Fox, C.M. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences*. Mahwah, NJ: Lawrence Erlbaum and Associates.

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

Crowe, A., Dirks, C., & Wenderoth, M.P. (2008). Biology in bloom: Implementing bloom taxonomy to enhance student learning in biology. *Journal of Life Science Education*, 7, 368-381.

Curtis, D.D & Boman, P. (2007). X-ray your data with Rasch. *International Education Journal*, 8(2), 249-259.

Darmawan, E. (2013). Pengaruh PBL terhadap sikap dan hasil belajar. [The impacts of PBL on learning attitude and outcomes]. *Jurnal Lentera Sains*, 3(2), 1-4.

DeHaan, R.L. (2009). Teaching creativity and inventive problem solving in science. *CBE—Life Sciences Education*, 8, 172–181.

Dettmer, P. (2006). New Blooms in established fields: Four domains of learning and doing. *Roeper Review*, 28(2), 70-78.

Diakidoy, I.A. & Constantinou, C.P. (2001) Creativity in physics: Response fluency and task specificity. *Creativity Research Journal*, 13, 3-4, 401-410, DOI: 10.1207/S15326934CRJ1334_17

Embretson, S.E. & Reise, S.P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.

Florida, R., Mellander, C., & Stolarick, K. (2011). *Creativity and prosperity: The global creativity index*. Toronto: Martin Prosperity Institute.

Gorin, J. & Embretson, S.E. (2006). Item difficulty modeling of paragraph comprehension items. *Applied Psychological Measurement, 30,* 394-411.

Hambleton, R.K. & Rovinelli, R.J. (1986). Assessing the dimensionality of a set of test items. *Applied Psychological Measurement,* 10(3), 287-302.

Hambleton, R.K., Swaminathan, H., & Rogers, H.J. (1991). *Fundamentals of item response theory*. London: Sage.

Hattie, J. (1985). Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement*, 9(2), 139-164.

Huitt, W. & Cain, S. (2005). An overview of the conative domain. *Educational Psychology Interactive*. Valdosta, GA: Valdosta State University. Retrieved from http:/www.edpsycinteractive.org/brilstar/chapters/conative.pdf

Isaksen, S.G., Dorval, K.B., & Treffinger, D.J. (1994). *Creative approaches to problem solving*. Dubuque, IA: Kendall/Hun. Retrieved on 22 January 2015 from https://books.google.co.id/books?id=dMGtBOux3LUC&pg=PA1&source=gbs_toc_r&cad=4#v=onepage&q&f=false

Jo, S.M. (2009). *A study of korean students' creativity in science using structural equation* *modeling* (Unpublished doctoral dissertation). University of Arizona, USA.

Krathwohl, D.R. (2002). A revision of Bloom's taxonomy: An overview. *Theory Into Practice*, 41(4), 212-264.

Linacre, J.M. (2015). A user's guide to WINSTEPS & MINISTEP Rasch-Model computer programs (Program Manual 3.90.0. winsteps.com). United States of America: Winstep Software Technologies.

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, Paris.

Mappiasse, S. (2006). Developing and validating instruments for measuring democratic climate of the civic education classroom and student engagement in North Sulawesi, Indonesia. *International Education Journal*, 7(4), 580-597.

Mardapi, D. & Kartowagiran, B. (2011). Pengembangan instrumen pengukur hasil belajar nirbias dan terskala baku [Developing unbiased and standardized instruments for student achievements in high schools]. *Jurnal Penelitian dan Evaluasi Pendidikan, 15*(2), 326-341. Retrieved on 20 January 2015 from http://journal.uny.ac.id/index.php/jpep/article/view/1100

Masters, G.N & Wright, B.D. (1997). The partial credit model. In W.J.V.D. Linden & R.K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 101-118). New York, NY: Springer-Verlag.

McDonald, R.P. (1981). The dimensionality of tests and items. *British Journal of Mathematical and Statistical Psychology*, 34(1), 100–117. Doi: 10.1111/j.2044-8317.1981.tb00621.x

Mumford, M.D., Mobley, M.I., Uhlman, C.E., Reiter-Pamon, R., & Doares, L. (1991). Process analytic models of creative capacities. *Creativity Research Journal,* 4(2), 91-122. Doi: 10.1080/10400419109534380

Nering, M.L., & Ostini, R. (2010). *Handbook of polytomous item response theory models*. New York, NY: Routledge.

Pepper, S.C. (1970). *The source of value*. Berkeley, CA: University of California Press.

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

Ramirez, R.P.B., & Ganaden, M.S. (2008). Creative activities and students' higher order thinking skills. *Education Quarterly*, 66(1), 22-33.

Reckase, M.D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics*, 4(3), 207-223.

Reeves, T. C. (2006). How do you know they are learning?: The importance of alignment in higher education. *International Journal of Learning Technology, 2*(4), 294–309.

Reynolds, C. R., Livingston, R. B., & Willson, V. (2009). *Measurement and assessment in education*. Upper Saddle River, NJ: Pearson Education.

Riyanti, D.B.P. & Prabowo, H. (1998). *Seri diktat kuliah psikologi umum 2* [Summary of general psychology lecturing vol. 2]. Depok: Universitas Gunadarma.

Runco, M.A. (2004). Creativity. *Annual Review Psychology,* 55, 657-687. Doi: 10.1146/annurev.psych.55.090902.141502

Subali, B. (2010). Bias item tes keterampilan proses sains pola divergen dan modifikasinya sebagai tes kreativitas [The bias of test item of divergent pattern science process skill and its modification as a creativity test]. *Jurnal Penelitian dan Evaluasi Pendidikan,* 14(2), 309-334. Retrieved on 15 January 2015 from http://journal.uny.ac.id/index.php/jpep/article/view/1084

Subali, B. (2011). Pengukuran kreativitas keterampilan proses sains dalam konteks assessment for learning [Creativity assessment of science process skills in the context of assessment for learning]. *Jurnal Cakrawala Pendidikan*. 30(1), 130-144. Retrieved on 7 January 2015 from http://journal.uny.ac.id/index.php/cp/article/view/4196

Subali, B., & Suyata, P. (2013). Standardisasi penilaian berbasis sekolah [The standardization of school-based assessment]. *Jurnal Penelitian dan Evaluasi Pendidikan, 17*(1), 1-18. Retrieved from http://journal.uny.ac.id/index.php/jpep/article/view/1358

Supardi, U.S. (2012). Peran berpikir kreatif dalam proses pembelajaran matematika [The role of creative thinking in mathematics instructional process]. *Jurnal Formatif*, 2(3), 248-262.

Wright, B.D. & Masters, G.N. (1982). *Rating scale analysis*. Chicago, IL: MESA Press.

Wright, B.D., & Stone, M. (1999). *Measurement essentials* (2nd ed.). Wilmington, DE: Wide Range.