RELIABILITY AND VALIDITY OF A GRAPHIC DESIGN ASSESSMENT RUBRIC

Bede Blaise Chukwunyere Onwuagboke[1]**, Termit Kaur Ranjit Singh**[2]
[1]Department of Curriculum and Instruction, Alvan Ikoku Federal College of Education, Owerri, Nigeria
[2]School of Educational Studies, Universiti Sains Malaysia, Pulau Penang, Malaysia
bbconwu@yahoo.com

Abstract— **Assessment in the visual arts is subjective in nature thereby raising a concern of how to effectively assess achievements in the subject. Quality assessment of artworks is highly dependent on accurate and reliable measurement. The objective of the paper was to construct and assess the reliability and validity of a scoring rubric for grading graphic design artefacts. The paper reports measure taken to ensure validity of the developed rubric. The reliability of the rubric was also investigated using Intraclass correlation coefficient (ICC). Findings indicate that the developed rubric show evidence of high reliability with an inter-rater correlation coefficient of 0.790 and intra-rater correlation coefficient of 0.828 which are within the very reliable range. It was concluded that the developed rubric was reliable. This conclusion has far reaching implications for the development and use of rubrics in assessment bearing in mind that subjectivity in the use of rubrics cannot be completely removed.**

Index terms- **Graphic Design; assessment rubric; validity; reliability.**

I. INTRODUCTION

The use of rubrics in instruction as assessment tool has been recognized by educationist. For example rubrics can be described as descriptive scoring instructional tool [1], [2] which form the foundation on which teachers make academic judgements about students' performances and measure students' achievements and progress [2], [3] as well as for ensuring the development of professional skills [4]. Rubrics generally give a connotation of a simple assessment tool that describes levels of performance on a particular task which is used to assess outcomes in a variety of performance-based contexts at all levels of education [5], [6]. They are adjudged beneficial in the learning situation in that they can enhance learning process by providing to both the teacher and the learners with a clear understanding of the objectives of a given (design) assignment as well as the criteria for assessment [7].

Fine and Applied Arts is a subject area which involves the production of artefacts as a measure of learners' accomplishment in the subject. This is applicable in almost all the branches of the subject whether it is drawing, painting, sculpture, ceramics, textiles or graphics except in art history and art education that are purely theoretical in nature. It is a subject that involves theory and practice in all the aspects that produces artefacts as measures of performance. While it may be easier to have an objective assessment of performance in the theoretical aspect, assessment of practical works produced by the students is usually subjective with instructors assessing the works according to personal appeal. The need for an assessment instrument capable of objectively assessing the artwork of students arises from the above nature of the subject. Every student who produces a design rarely finds any problem with his/her work thereby anticipating a very high grade. Often time's students have got reasons to grumble over some scores given to them by their lecturers stressing that such scores do not reflect the quality of work they had produced. To solve the problem of subjectivity in graphic design assessments the authors embarked on this study based on the need for objectivity and transparency in the assessment process.

Need for Assessment Rubrics

Rubrics are recommended in assessment methods where the students' responses to questions cannot be evaluated with complete objectivity, such as projects, artworks portfolios etc. as they can be employed for achieving reliable and valid professional judgement [8]. The use of rubrics is believed will lead to increased objectivity in the assessment of artefacts as criteria are explicitly defined [9]. To this end, different teachers or raters can make use of a common rubric across a subject to ensure that measurement of students' performance is consistent. They have been used by teachers in the classroom to communicate expectations for an assignment, providing focused feedback on works in progress, as well as in grading final students' products of learning activities [10], [11]. Rubrics are valuable instruments in the educational setting to both teachers and students alike. When the right type of rubric is used, it has the propensity to enhance the reliable scoring of performance assessments. They seem to have the potential of promoting learning and/or improve instruction. This is possible due to the fact that rubrics make expectations and criteria explicit, which also facilitates feedback and self-assessment on the part of the students [12].

Features and Types of Assessment Rubrics

There are three basic features which are considered essential in scoring rubrics which are evaluation criteria, quality definitions and scoring strategy [13]. The valuation criteria are the factors which the assessor of an artefact must put into consideration in order to determine the quality of such a work. In areas where it is difficult to define certain concepts, the criteria will highlight the indicators of what is to be measured to bring about clarity and understanding of requirements of performance. A well spelt out assessment criteria has the advantage of informing the students of the teacher/ assessors' expectation from a given assignment thereby giving them insight of what constitutes complete and effective response to the design problem [2]. When assessment is criterion-based, a room is created for assessment to play a leading role in the learning process [14], as it can motivate the learners to greater achievement, improve the instructional process and equally enhance assessment [15].

.

Quality definitions is a feature which entails stating a detailed explanation of what the learner have to do as a demonstration of the level of mastery or achievement of a skill, proficiency or criteria. They are statement which focuses on ascertaining a good response from a bad one. They are usually stated from the highest to the lowest level with other levels in between. It has not been widely accepted the ideal number of levels that a rubric should have though it is widely canvassed that the levels should be few. It has been suggested by some researchers that the minimum levels be three [13, 16] and a maximum of five levels [13]. Whatever the number of the levels is, it is important that the designer of the rubric should expressly state in clear and understandable language the expectation of the assessor from the student in the given task, so as to distinguish performances of students.

The next important feature of rubrics is the scoring strategy which entails the use of a measurement scale for judging and interpreting the product. Rubrics can be categorized into two namely holistic and analytical rubrics [17, 18, 5], specific or generic [19]. A holistic rubric is a scoring scale which assigns a level of performance by assessing performance across multiple criteria as a whole. There is no separation of levels of performance for each criterion. This type of rubric can be very useful in assessing students work in large classes where the assessor has to assess many works as it gives a broad picture of performance at a glance. Its shortcoming is actually that it does not provide detailed information. Advocates of using a holistic rubric maintain that it has the advantage of making it possible for works to be scored quickly thereby giving an overview of performance of students. The decision of which type of rubric to use lies with the designer and it is a function of how the result of the assessment will be put to use.

On the other hand, an analytical rubric is one which articulates levels of performance for each criterion so the teacher can assess student performance on each criterion. In this type of rubric, the scores for each criterion can be summed up to arrive at the final grade of the student. It may not be easy to provide one overall score of the student when a rubric with so many criteria is in use. Analytic rubrics would be preferable if the objective of its use is to provide a detailed diagnostic feedback of the strength and weaknesses of students' artwork and the effectiveness of instructional intervention [13]. Analytical rubric is highly time-consuming in scoring students work, however it provides more detailed feedback, with scoring being more consistent across students and raters [20].

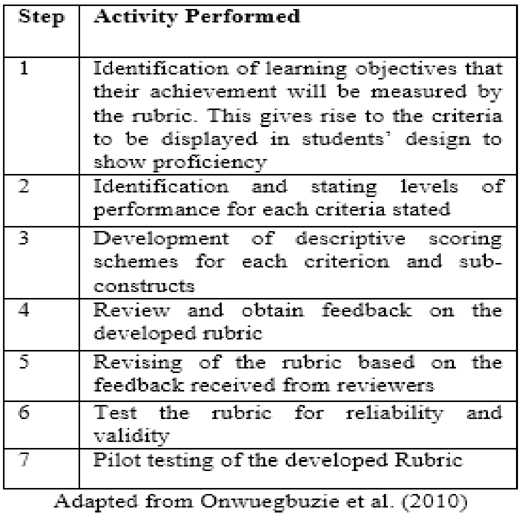
When a holistic or analytical rubric is designed to be used in assessing an individual assignment or task, it is said to be task specific whereas if it is designed so that it can be utilized in the assessment of a group of similar tasks, it is termed generic. Task specific rubrics lend themselves to thoroughness of detail which accounts for higher reliability and validity unlike generic rubrics [18]. The generic rubrics have the advantage of being used in assessment of wide range of similar courses and institutions because of its inherent flexibility. They have a wider scope and contain only the most essential ingredients of the learning outcome to be assessed across different tasks in the same assessment method [21].

For any measuring instrument to be considered effective in measuring students achievement in a given performance oriented task, there are two qualities the instrument must exhibit. Validity and reliability are essential qualities of good measuring tools in any setting. An assessment rubric whether it is in the holistic or analytical formant is expected to exhibit these qualities. Validity of a measuring instrument refers to the extent to which scores obtained using the instrument truly reflects the underlying variable of interest. On the other hand reliability of a measuring instrument is concerned with the consistency of scores across repeated measurements. Reliability and validity of assessment rubrics has often not been assessed probably because of the effort and time commitment required to undertake such assessment [22] despite the fact that some researchers have noted them as issues of concern in development of rubrics [23]. This paper designs and validates a rubric for scoring graphic design artworks produced in graphic design studios in schools and colleges.

II. Materials and Methods

The study reviewed literature on the design and construction of assessment rubrics and proposed and validated an assessment rubric for assessing graphic design projects/assignments. Following the steps for instrument development and validation methodology [24] the authors constructed the rubric. The step-by-step approach followed by the researchers in designing the assessment rubrics are as stated in table 1.

Table 1
Steps in developing the rubrics



Step	Activity Performed
1	Identification of learning objectives that their achievement will be measured by the rubric. This gives rise to the criteria to be displayed in students' design to show proficiency
2	Identification and stating levels of performance for each criteria stated
3	Development of descriptive scoring schemes for each criterion and sub-constructs
4	Review and obtain feedback on the developed rubric
5	Revising of the rubric based on the feedback received from reviewers
6	Test the rubric for reliability and validity
7	Pilot testing of the developed Rubric

Adapted from Onwuegbuzie et al. (2010)

Three experienced graphic design lectures not below the rank of senior lecturer from a College of Education participated in the rubric construction exercise. Two of the lecturers are males while one is a female.

(1) Designing the rubric (steps 1-3)

The first three steps for developing the rubric as recommended by scholars [24, 19] include 1. Conceptualize the construct of interest; 2. Identify and describe levels of behaviors that underlie the construct; 3. Develop the instrument by developing separate descriptive scoring schemes for each evaluation level and criteria. The experts were asked to list the criteria they normally consider in assessing designs produced by students in their classes. Four criteria of all that was listed were agreed upon after in-depth discussion on them. They are:

1. 1. General Appearance of the design
2. 2. Display of Creativity in the design
3. 3. Design layout
4. 4. Use of media and Technology.

Graphic design is seen as the activity that organizes visual communication in society with its primary concern being the efficiency of communication, the technology used for its implementation and the social impact it effects [25]. As the products of graphic design are visually appreciated and consumed, the general appearance of the products constitutes a major criterion to measure its effectiveness. Five items on the scale were constructed to address the general appearance of the final design. The graphic design process is a creative process which combines images and texts to convey information to a given audience, the layout and organization of the design elements on the working surface was adjudged by the team as a necessary criterion to be assessed. To this effect, five items were also developed to measure the design layout.

Art and design are ventures that have been seen as creative. In assessing any work of art and design, the team also maintained that creativity exhibited in the work should be the hallmark of such assessment. Though it was difficult for the team to arrive at a consensus of the attributes constitutes creativity they however pointed out originality, innovativeness, thoughtfulness, improved product compared to previous ones and acquisition of more skills as manifestations of a creative work of art amounting to another five items. Finally they also came to the agreement that the technical use of media and the technology involved in the creative process is worth assessing when evaluating a finished work of art hence five items were developed.

Each of the criteria so identified were subjected to further analysis to bring out detailed explanation of what the learner have to do as a demonstration of the level of mastery or achievement of a skill related to each of the set criterion. For each of the items, quality statements were constructed to state the levels of which a good response can be distinguished from a bad one. The statements were generally agreed on by the team to be at four levels as recommended [26]. Based on the above, a four point scale with labels as basic, acceptable, good and excellent with numerical values 1-4 were given to show advancement from the lowest performance to the highest anticipated performance. The initial rubric that resulted from the above exercise contained twenty items.

Review, Feedback and Revising the Rubric (steps 4-5)

The developed instrument was discussed with major stakeholders in graphic design in the College which comprises students and lecturers in graphic design. The purpose was to determine the usability and appropriateness of the language used in the measure as well as to ensure that the instrument covers the content area. Lecturers were asked to rate and comment on the rubrics with regards the following: clarity, completeness and applicability in measuring all aspects of a finished work in graphic design. The students on their part were asked to comment on the clarity of language and completeness of the rubrics. Statements which were not clear to the students were rephrased to ensure their clarity while very ambiguous ones were deleted.

The ratings and various comments by the lecturers and the students helped to reconstruction of some of the items to make them clearer as well as reduce the length of some of the descriptors. Some ambiguous items in the rubrics were deleted as well as items that seem to be repeated in the scale. In general, the rubric received high rating from the lecturers and the students found the rubric easy to interpret and apply in the context of their design assignments. The number of items in the rubric was reduced from twenty to fifteen items in line with the recommendations of the reviewers. Most items that appear to be repeated were deleted alongside items that do not seem measurable by merely observing and rating the finished artworks.

Reliability and Validity of the Developed Rubrics (step 5)

The essence of developing these rubrics is to overcome the shortcoming of subjective evaluation of students' products by teachers. There is the need for the assessments based on the use of the rubrics to be sound, unbiased and free from distortions [19]. Establishing validity and reliability of the rubrics is an exercise that was taken to ensure confidence in the use of the instrument. In an ideal situation, a sore a candidate receives in a test should be independent of the scorer and similar results obtained no matter when and where the assessment is carried out [12]. This is however rarely obtainable in practice. Multiple choice tests are prone to yielding similar results irrespective of the scorer, the time of scoring and the place. This makes it imperative that rubrics which measures complex performance assessment should be made to be reliable and consistent in measurement. The rubric is also expected to be valid by measuring what it is intended to measure. Simply put, the rubrics should accurately capture the intended outcome [26] for it to be said to be valid.

Validity

Validation of any measure is a process through which evidences that support the appropriateness of any conclusion drawn from students' work for specified assessment uses are accumulated [27]. Test validity has been seen by researchers as the most important factor to be taken into consideration in test evaluation [28]. Validity of any measuring instrument can be looked at from varying perspectives, through which the needed evidences are gathered content, construct and criterion validity. The content validity dimension focuses on evidence that students' responses/performance to the given instrument is a reflection of the students' knowledge of the relevant content area. It is also concerned with the extent which the assessment tool adequately samples the content domain and also ensure that measurement does not go beyond the scope of the content being measured. Construct related evidence deal with gathering of evidence that responses provided by individuals are as a result of their internal thought processes rather than by chance. Criterion validity deals with evidence that performance of a student in a given task maybe generalized to other relevant activities. It accumulates evidence of transfer of learning to solving real world problems.

Furthermore, [26] posit that validity is often harder to establish than reliability, it is preferable for assessments to contain multiple forms of validity as described above. Content validity of the rubric was ensured by the involvement of experts in the field of graphic design in the college to determine the criteria for assessment. It is most common to use experts to establish content validity to ensure that the rubric covers the full range of the concept's meaning [29, 30, 31]. Researchers have also stressed the importance of appropriateness of language in the understanding and use of rubrics by both teachers and students [21], [32]. Stating the criteria and descriptors of the levels of performance expected of the student in clear and understandable language enables the students to work to specifications as well as make clear to the rater what is required of the product thereby making sure that no extraneous content was inherent in the rubrics. This in no small measure increases the validity of the rubric.

To this end therefore not only were the experts involved in the determination of the criteria for assessment, the students were also involved in the review process. They read the rubrics and offered their feedback on their understanding of what the teacher requires of them in the given task being evaluated. The essence of the review was to eliminate ambiguity from the instrument as much as possible as it does not make room for proper interpretation [33]. All the above measures were taken, in order to ensure as well as enhance the validity of the rubric.

Reliability

Literatures on reliability of rubrics identify two ways of estimating the reliability of such a measure which are inter-rater reliability and intra-rater reliability [27, 19, 22, 33].

Inter-rater reliability

Inter-rater reliability is concerned with the likelihood of variance in students' score based on the subjective judgement of different raters. In effect the same piece of artwork produce by a student may receive as many different score values as there are scorers. Should there be a scoring rubric acting as a guide formalizing the criteria used by the assessors, the similarity in the score values given to such a work by different assessors can be highly enhanced.

To ensure that the rubric developed can yield similar results when used by different raters, the poster designs produced by fifteen pre-service art teachers were rated by three lectures in the department using the rubric. The inter-rater reliability of the rubric was computed using Intraclass correlation coefficient (ICC) which is a measure of correlation, consistency or conformity for a data set with multiple groups. A two-way random effects model of ICC was utilized in this study. To determine the sample size needed, [34] concluded that the optimal sample size for two ratings varied from 15 for ICC=.9 to 378 for ICC = .1; for three ratings, it varied from 13 to 159; five ratings, 10 to 64; and 10 ratings, 8 to 29.

In other words, fewer ratings and the smaller the ICC level, the larger the needed sample size. Fifteen designs were randomly selected from the products of the pre-service teachers. Three of the lecturers who participated in the development and review of the rubric were selected to act as raters for the designs. All the raters were made to rate the fifteen artworksindependently giving three scores for each poster. Following the recommendation of [35] all the individual scores awarded by the raters constitutes the unit of analysis rather than an average measure. This analysis was carried out using SPSS version 22 and a 95% confidence interval was defined.

Intra-rater reliability

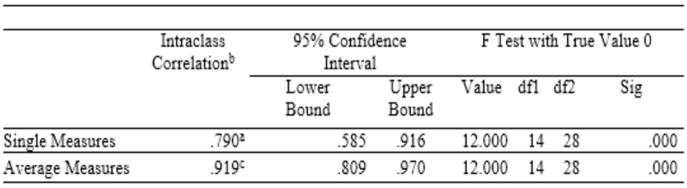
Intra-rater reliability is concerned with the probability of a rater obtaining similar scores from the same set of sampled products at measurements taken at two different points in time (test-retest reliability). It is a metric for rater's self-consistency in the scoring of subjects [36]. Experience has shown that certain psychological conditions around the rater at two different times most often affects the persons' performance of a task. When this happens using the same measure, as a measure's reliability is suspect. A reliable measure will yield a very high level of similarity in result when used by the same individual at different times. To ensure that this type of reliability is inherent in the rubric, the instrument was subjected to intra-rater reliability investigation. One of the raters was asked to rate the samples using the rubric and the ratings recorded, after an interval of two weeks he was asked to rate the same samples using the rubric. The ICC was computed for the two ratings of the rater taken at two different times.

III. Results and Discussion

Either Consistency Agreement (CA-ICC) or Absolute Agreement (AA-ICC) Intraclass correlation coefficient can serve as a useful measure of agreement depending on whether rater variability is relevant for determining the degree of agreement. According to [37] CA-ICC is useful when comparative judgments are made about objects of measurement thus representing correlation when the rater is fixed. Based on that, the researchers report the absolute agreement ICC. The result of inter-rater investigation carried out shows that the absolute agreement ICC is 0.790 with 95% confidence interval

(0.585-0.916) for the rubric on a single measure as shown in table 2.

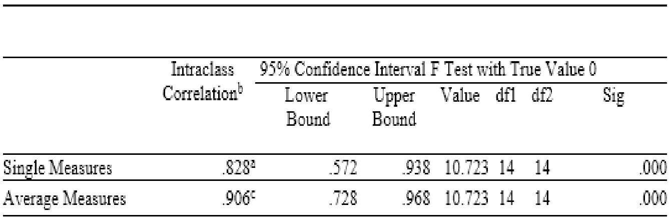
Table 2
Inter-rater reliability of assessment rubric using Intraclass correlation



	Intraclass Correlation^b	95% Confidence Interval		F Test with True Value 0			
		Lower Bound	Upper Bound	Value	df1	df2	Sig
Single Measures	.790^a	.585	.916	12.000	14	28	.000
Average Measures	.919^c	.809	.970	12.000	14	28	.000

The intra-rater correlation coefficient computed gave an output of 0.828 with 95% confidence interval (0.572-0.938) for one rater who repeated his rating of the sampled designs after two weeks of the first rating (repeated measures). Consistency measures of 0.70 or above are usually considered acceptable in literature [12, 38]. Table 3 shows the intra-rater reliability achieved using the rubrics to assess the designs of fifteen students by a single rater. The single measure ICC is reported instead of the average measure.

Table 3
Intra-rater reliability of assessment rubric using Intraclass Correlation Coefficient



	Intraclass Correlation^b	95% Confidence Interval		F Test with True Value 0			
		Lower Bound	Upper Bound	Value	df1	df2	Sig
Single Measures	.828^a	.572	.938	10.723	14	14	.000
Average Measures	.906^c	.728	.968	10.723	14	14	.000

The use of single measure ICC is recommended if further research will use the ratings of a single rater [39]. The ICC (.828) is within acceptable range according to [12, 38], < 0.40 is poor, 0.40to <0.74 is adequate and acceptable while > 0.75 is regarded as excellent correlation. Thus the rubrics yielded high inter-rater and intra-rater consistency agreements among the raters unlike newly developed rubrics [22]. The authors attributed the high correlation coefficient established to the fact that the raters used in the scoring of the samples were actually part of the team that developed the rubrics as they may have become familiar with the rubrics. In other words the development process may as well have served as enough training in the use of rubrics in scoring [22]. Secondly their wealth of experience in rating of artworks over the years in their field as all are within the rank of principal lecturer and above may have made them to be conversant with what constitutes a good work of art. The practice of moderating scores given to students' work by an external judge which is inherent in the fine art evaluation praxis may have likely influenced their pattern of evaluation and rating of artworks towards achieving consensus scores.

IV. CONCLUSION

The need to give an objective assessment to the graphic design artefacts produced by pre-service teachers so as to give them orientation on how to objectively assess students' products necessitated the development of the graphic design assessment rubric (GDAR). For the developed rubrics to be effectively utilized in the instructional process, there is need to investigate its' validity and reliability. The Graphic Design assessment rubric is thus validated following the laid down procedures. Similarly, the rubric is a reliable measure for assessing Graphic Design artefacts based on the established ICC.

The relationship between reliability and validity of instruments is such that the establishment of reliability is necessary condition for establishing validity. It does not however imply that a reliable assessment is by extension valid even though a valid assessment is a reliable assessment [27]. Reliability and validity of a rubric is not a function of the type of rubric whether it is holistic or analytical, task specific or generic in nature but rather dependent on the pains taken and carefulness in the design process.

Assessment rubrics are very useful assessment tools for teachers at all levels of education. Their use in literature is not limited to only subjects that produce artefacts as evidence of achievement in a learning setting. We therefore recommend that teachers in subject areas that require the use of rubrics in assessment to follow development procedures that will ensure very high reliability as well as validity of their rubric so as to give accurate measurement of students' performances in their subject at all times.

REFERENCES

Oakleaf, M. (2009). Using rubrics to assess information literacy: An examination of methodology and interrater reliability, *Journal of the American Society for Information Science and Technology, 60*(5), 969-983

[2] Egodawatte, G. (2010). A rubric to self-assess and peer-assess mathematical problem solving tasks of college students, *Acta Didactica Napocensia 3*(1), 78.

[3] Reynolds-Keefer, L. (2010). Rubric-referenced assessment in teacher preparation: An opportunity to learn by using, *Practical Assessment, Research & Evaluation* 15(8). Retrieved December 10, 2013 from http://pareonline.net/getvn.asp%3Fv%3D15%26n%3D8.

.

[4] Mertler, C. A. (2001). Designing scoring rubrics for your classroom. Practical Assessment, Research & Evaluation, 7(25), 1-10.

[5] Mueller, J. (2014). *Rubrics (authentic assessment toolbox)* Retrieved March 5, 2014 from http://jfmueller.faculty.noctrl.edu/toolbox/rubrics.htm

[6] Hanfer, J.C. & Hanfer, P. M. (2003). Quantitative analysis of the rubric as an assessment tool; An empirical study of student peer-group ratings. *International journal of Science Education 25*, 1509-1528

[7] Andrade, H. G. (2005). Teaching with rubrics: The good, the bad and the ugly. *College Teaching, 53*, 27-30.

[8] Pellegrino, J. W., Baxter, G. P., & Glaser, R. (1999). Addressing the" two disciplines" problem: Linking theories of cognition and learning with assessment and instructional practice. Review of research in education, 307-353.

[9] Peat, B. (2006). Integrating writing and research skills: Developing and testing of a rubric to measure student outcomes. *Journal of Public Affairs Education*, *12*, 295-311.

[10] Andrade, H. (2000). Using rubrics to promote thinking and learning. *Educational Leadership*, *57*(5), 13-18

[11] Moskal, B. M. (2003). Recommendations for developing classroom performance assessments and scoring rubrics. *Practical Assessment, Research & Evaluation, 8*(14). Available: http://pareonline.net/getvn.asp?V=8&n=14

[12] Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review*, *2*(2), 130-144.

[13] Popham, W.J. (1997), "What's wrong and what's right with rubrics", *Educational Leadership*, *55*(2), 72-75.

[14] Eshun, E. F. (2011). Report on the action research project on adopting innovative assessment for learning in communication design in higher education. *Paper presented at the Conference on Design, Development & Research, 26-27 September, 2011*, Cape Town, 382-395

[15] Gasaymeh, A-H. (2011). The implications of constructivism for rubric design and use, *Higher Education International Conference (HEIC 2011).* Retrieved 23rd February 2014 from http://heic.info/assets/templates/heic2011/papers/05-Al-Mothana_Gasaymeh.pdf

[16] Stevens, D. D., & Levi, A. (2005). Leveling the field: Using Rubrics to achieve greater equity in teaching and grading.

[17] Brookhart, S. M. (1999). *The art and science of classroom assessment: The missing part of pedagogy*. ASHE-ERIC Higher Education Report (Vol. 27, No.1). Washington, DC: The George Washington University, Graduate School of Education and Human Development.

[18] Moskal, B. M. (2000). "Scoring rubrics: What, when and how?" *Practical Assessment, Research & Evaluation, 7*(3). Retrieved from http://pareonline.net/getvn.asp?v=7&n=3

[19] Reddy, M. Y. (2011). Design and development of rubrics to improve assessment outcomes: A pilot study in a Master's level business program in India. *Quality Assurance in Education*, *19*(1), 84-104Callison, D. (2000). Rubrics. *School Library Media Activities Monthly*, *17*(2),

[20] Zimmaro, D. M. (2004). *Developing grading rubrics*, Retrieved March 5th 2014 from http://www.utexas.edu/academic/mec/research/pdf/rubricshandout.pdf

[21] Tierney, R., & Simon, M. (2004). What's still wrong with rubrics: focusing on the consistency of performance criteria across scale levels. Practical Assessment, Research & Evaluation, 9(2), 1-10.

[22] Stellmack, M. A., Konheim-Kalkstein, Y. L., Manor, J. E., Massey, A. R., & Schmitz, J. A. P. (2009). An assessment of reliability and validity of a rubric for grading APA-style introductions. *Teaching of Psychology*, *36*(2), 102-107.

[23] Thaler, N., Kazemi, E., & Huscher, C. (2009). Developing a rubric to assess student learning outcomes using a class assignment. *Teaching of Psychology, 36*(2), 113-116.

[24] Onwuegbuzie, A. J., Bustamante, R. M. & Nelson, J. A. (2010). Mixed research as a tool for developing quantitative instruments. *Journal of Mixed Methods Research*, *4*(1), 56-78.

[25] Frascara, J. (2006). Graphic design: Fine art or social science. Design Studies: Theory and Research in Graphic Design. Audrey Bennett (ed.), 26-35.

[26] Finley, A. (2012). Making progress?: What we know about the achievement of liberal education outcomes. Association of American Colleges and Universities.

[27] Moskal, B. M., & Leydens, J. A. (2000). Scoring rubric development: Validity and reliability. *Practical Assessment, Research & Evaluation*, 7(10), 1-11.

[28] Downing, S. M., & Haladyna, T. M. (1997). Test item development: Validity evidence from quality assurance procedures. *Applied Measurement in Education*, *10*(1), 61-82.

[29] Chambliss, D. F., & Schutt. R. K. (2009). *Making sense of a social world: Methods of investigation.* Thousand Oaks, CA: Pine Forge Press.

[30] Huck, S. W. (2008). *Reading statistics and research*. Boston, MA: Pearson Education.

[31] Reynolds, C. R., Livingston, R. B., & Wilson, V. (2006). Measurement and assessment in education. In A. E. Burvidovs (Ed.), *Validity for teachers* (pp. 117–140). Boston, MA: Pearson Education.

[32] Moni, R.W., Beswick, E. & Moni, K. B. (2005). Using student feedback to construct an assessment rubric for a concept map in physiology. *Advances in Physiology Education, 29*, 197–203.

[33] Payne, D. A. 2003. Applied educational assessment. 2nd ed. Belmont, CA: Wadsworth/Thomson Learning.

[34] Watts, F., Marin-Garcia, J. A., García Carbonell, A., & Aznar-Mas, L. E. (2012). Validation of a rubric to assess innovation competence. *WPOM-Working Papers on Operations Management*, *3*(1), 61-70.

[34] Bonett, D. G. (2002). Sample size requirements for estimating Intraclass correlations with desired precision. *Statistics in Medicine 21*, 1331-1335.

[35] Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin*, *86*(2), 420.

[36] Gwet, K. L. 2014. Intrarater Reliability. Wiley Stats Ref: Statistics Reference Online.

[37] McGraw, K. O. & Wong. S. P. (1996). Forming inferences about some Intraclass correlation coefficients. *Psychological Methods*, *1*(1), 30–46.

[38] Stemler, S. E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability *Practical Assessment, Research, and Evaluation*, *9* (4) Retrieved from http://PARAonline.net/getvn.asp?v9&n=4

[39] David Garson, G. (2009). *Reliability analysis*. Retrieved 11th March, 2014 from http://tx.liberal.ntu.edu.tw/~purplewoo/Literature/!DataAnalysis/Reliability%20Analysis.htm