Educational Research Journal《教育研究學報》, *Vol. 22, No. 2, Winter 2007*

Which Rubric is More Suitable for NSS Liberal Studies? Analytic or Holistic?

Sze-wing A. Kuo
The Chinese University of Hong Kong

This paper tries to gather evidence from literature review, to answer the question asked by the Hong Kong Examinations and Assessment Authority (HKEAA): whether analytic rubric or holistic rubric should be chosen in scoring candidates' written responses of New Senior Secondary (NSS) Liberal Studies public examination. Criteria for comparison include: impact, discriminative power, inter-rater reliability, bias-free, and time-saving. Analysis of the research findings suggests that analytic rubric appeared to be more appropriate for the assessment. Further studies are suggested to be done in implementing a large scale research project for facilitating learning from performance-based assessment, accommodation of students with special learning difficulty, combination of holistic and analytic rubric and comprehensive raters' training.

Key words: liberal studies, rubric, performance-based assessment

Correspondence concerning this article should be addressed to Sze-wing A. Kuo, e-mail: szewingkuo@gmail.com

Background

In September 2000, the reform proposal for the education system in Hong Kong was submitted to the Government by the Education Commission (EMB, 2000), that signified the beginning of a major educational reform. One of the prominent changes is the introduction of Liberal Studies, which will be a core subject, in addition to Chinese, English and Mathematics, in the new senior secondary school curriculum (CDC & HKEAA, 2007). Matching with the new curriculum, a new public examination, Hong Kong Diploma of Secondary Education (HKDSE) will be held in 2012 (CDC & HKEAA, 2006).

As part of the development process of the public examination, HKEAA and Education Bureau launched five consultation-seminars for teachers and principals about the progress of assessment of New Senior Secondary Liberal Studies (LS) during September to November in 2006. During the meetings, a question about the assessment of LS was thrown to the audience for comments: Which rubric, analytic or holistic, should be used for scoring students' essays in the public examination of LS?

In this paper, I attempt to answer the question by reviewing related research studies. First, concepts of LS, performance based assessment and rubrics are explained. Second, criteria of a suitable rubric for LS will be discussed. Third, based on the criteria, analytic and holistic rubrics are compared with each other by reviewing related empirical studies. Lastly, suggestions and future directions of research in this area will be explored.

LS, Performance-Based Assessment, and Rubrics

What is LS?

According to the Liberal Studies Curriculum and Assessment Guide (CDC & HKEAA, 2007), LS is neither textbook-bound nor necessarily tied to any one particular ideology, e.g., humanism or postmodernism. Daily news or any phenomenon happening around the students could be the learning

materials. It aims at helping students to understand themselves and their surrounding environment like contemporary issues in Hong Kong and global context. "It is built on the foundation of the eight key learning areas in basic education and extends this into new areas of knowledge" (CDC & HKEAA, 2007, p. 3). Based on student-oriented approach and issue-enquiry approach, students are expected to learn how to learn. For example, the recent financial storm originated in the United States could be a good issue for students to explore the phenomenon of globalization and might have some insights about their own personal financial planning.

Unlike history or mathematics, there is no definite domain-specific knowledge of this subject. The assessment objectives of LS could be found in the Appendix. Traditional assessment items like multiple choice or knowledge-specific and context-free questions, e.g., "what is globalization?" could not assess students' learning in LS. Instead, authentic questions which are more related to their daily life and their real life experiences, is more aligned to the aims of LS. This type of authentic question is used in performance-based assessment.

Performance-based Assessment

In performance-based assessment, students are provided with authentic questions in real life context (Harmon et al., 1997). This assessment consists of two parts: a performance task and a set of scoring criteria (Perlman, 2002). For example, Miss Chan and Miss Lee are two art teachers of primary 2 in the same school. Mother's day is coming, and the students are asked to draw a mother's day card to cheer up their mother. The card is a performance task and the mother's day is the real life context. Their scoring criteria are message written and creativity.

Performance-based assessment has been gained attention in large-scale assessment (Baker, 1994), as in the case of LS. Niemi, Baker, and Sylvester (2007) reported a 7-year performance assessment development and validation project done by the Los Angeles Unified School District and UCLA's Center for Research on Evaluation, Standards, and Student Testing. Assessment

has been seen as an effective strategy to change educational practice, and policy directives. More than 3,000 teachers participated in the development of performance assessment.

In addition to measure student achievement (Baker, 1994; Goldberg, & Roswell, 2000), advantages of using performance-based assessment include: assist student learning by providing information to guide changes in day to day practice (Jonsson & Svingby, 2007; Niemi, et al., 2007), promote self and peer assessment, enhance teachers' reflective practice (Jonsson & Svingby). This shows a paradigm changed, from solely "teaching the examination" to "assessment for learning" (Baker, 2007; Goldberg & Roswell, 2000). Assessment for learning refers to any assessment for which the first priority in its design and practice is to serve the purpose of promoting students' learning (Black, Harrison, Lee, Marshall, & Wiliam, 2004).

Moreover, developers of LS public examination expect that the performance-based assessment could provide useful information to students for what they could do, to teachers and schools on the quality of teaching provided, and also to communicate with parents, tertiary institutions, employers and the public about what students could do after studying LS. Information has to be exchanged between assessment and other settings. The impact of performance-based assessment on students' learning is expected to be very high.

To make performance-based assessment work well, fairness is very important for measuring student achievement in large scale examination. Being fair means same response will get same scores when they are rated by different scorers. Therefore, inter-rater reliability should be high and bias-free. Take the previous example of Miss Chan and Miss Lee. If both of them give same score to same mother's cards, it is said to have a good or high inter-rater reliability. Also, Miss Chan will not give a green mother's card lower score because she does not like the color green. The score is biased because of the teachers' own preference which is unrelated to students' performance. From Ni's review (1997) of performance-based assessment researches, score variability to same response due to different raters was

found to be one of the major problems for this kind of assessment. Therefore, monitoring and building a high inter-rater reliability is very important.

In addition, the performance task should also be able to discriminate good responses from bad ones. This is called discriminative power. For example, in the mother's day cards exercise, if every kid get the same score, then the discriminative power of the scoring exercise is low. However, kids who draw better get high scores while kids who draw worse get low score, then the discriminative power is high.

In sum, for performance-based assessment like LS public examination, developing a valid and reliable examination is very essential, and rubrics play an important role in enhancing the effectiveness of scoring process.

What is a Rubric?

According to Oxford online dictionary, rubric is derived from the Latin word, "rubrica terra", which means red earth or ochre as writing material and is referred to text written in red for emphasis. Nowadays, it means a set of instructions or rules. In education, It is a guide with established criteria, illustrations and rating scales that are used in scoring one or more dimensions of the performance tasks (Perlman, 2002; Westat, 2001), e.g., essay, project or presentation (Schultz, 2002).

According to Herman, Aschbacher, and Winters (1992), there are four characteristics of a rubric: criteria, examples, scale, and standards. First, there are one or more dimensions or criteria used to judge students' responses. Second, examples are provided to clarify the meaning of each trait or dimension. Third, a scale of values are assigned to rate each dimension. Fourth, a rubric has standards of excellence for specified performance levels with examples of each grade level.

By describing the characteristics of different levels of responses within each score category, a rubric guides evaluation of students' responses, so that ideally, every rater would give the same score to the same response (i.e., inter-rater reliability) (Gopinath, 2004).

Holistic Rubric

Students' responses are given scores in one scale in a holistic rubric. The scale is divided into levels in which descriptors and criteria for grading students' work are listed. There is a gradual difference found among each level for individual descriptor or criteria. One score is given to the answer as a whole. For example, a holistic rubric used by Miss Chan and Miss Lee in the previous examples would be like this:

Score	Description
3	Very creative, write personal message e.g., draw pictures to express their inner thoughts, draw something new
2	Average creative, write some words e.g., write "Happy mother's day", draw something popular with some modifications
1	Not creative, no message e.g., just draw a heart with pencil

Other examples include a holistic critical thinking scoring rubric by Facione, Sanchez, and Facione (1994) and a "Participation, Appearance, Cleanup, Engineering, and Safety" (P.A.C.E.S.) grading rubric for projects by Tufte (2005).

Analytic Rubric

An analytic rubric was firstly developed by the National Assessment of Educational Progress (NAEP) in U.S. (Lloyd-Jones, 1977), and the rubric was known as primary trait analysis. Students' responses are given scores according to different domains (at least 2) of the predetermined rubric. By summing up the scores obtained from different parts, students receive a total score of their responses to a particular question. Referring to our previous example of Miss Chan's and Miss Lee's performance assessment, their analytic rubric is illustrated as follows:

Score	Creativity	Message
3	Very creative Draw something new	Write personal message or draw pictures to express their inner thoughts
2	Average creative Draw something popular with some modifications, e.g., Mickey Mouse	Write some words e.g., write "Happy mother's day"
1	Not creative e.g., just draw a heart with pencil	No message

There are two components in assessing their students' writing, creativity and message conveyed. Miss Chan and Miss Lee will score their students twice on the two components independently. Two students who get the same total score might have a different performance on the two criteria. For example, both students A and B get a total score of 4. Student A may get a score 3 in creativity, and 1 in message, while student B might score 2 in both creativity and message. The teacher will know student A is very creative, but with room improvement in the message conveyed in his/her writing.

It is widely used in the educational settings, e.g., identifying students' writing proficiency levels for different purposes in EFL/ESL programs (Bacha, 2001), scoring postsecondary academic skills by Simon and Forgette-Giroux (2001), rating children's hypermedia "narratives" by Mott, Etsler, and Drumgold (2003), assessing students' performance in inclusive science by Finson and Ormsbee (1998), and measuring students' higher order thinking skills in Washington State University by Kelly-Riley, Brown, Condon, and Law (2001).

Criteria to Compare Analytic and Holistic Rubrics

The relation between rubrics and performance-based assessment is like a bridge connecting assessment and classroom or other settings (*impact* on teaching and learning, and communication with teachers, students, parents,

and employers, etc.). Moreover, it is also like a map which shows different students' responses to different scores (*discriminative power*). Every one will locate the same score by using the same rubric (*inter-rater reliability*) and there is no bias.

In addition to the above four criteria, *time saving* is added to the comparison. In large-scale examination, a lot of students' responses will be graded. However, human resources might be tight, compared with the workload. Therefore, time-saving is also a comparison criterion. In sum, there are five criteria in total: impact, discriminating power, inter-rater reliability, bias-free, and time-saving.

Analytic Versus Holistic Rubrics

Impact

Analytic rubrics provide more detailed information for future planning and improving instruction since it indicates the performance from different perspectives and students' different abilities in the performance assessment (Moon, Callahan, Brighton, & Tomlinson, 2002; Perlman, 2002). Criswell and Criswell (2004) suggested that in order to let students demonstrate their cognitive processes and thinking skills as requested in extended response questions, regular practice and understanding should be provided. Each question would measure different types of abilities. They argued that analytic scoring rubric can serve this purpose, but not holistic rubric.

Niemi, Wang, Steinberg, Baker, and Wang (2007) provided support for the above claim. They studied the instructional sensitivity of a standards-based ninth-grade performance assessment which was about writing an essay of conflict in a literary work. Teachers of 886 ninth-grade students were randomly assigned to one of three instructional groups: literary analysis, organization of writing, and teacher selected instruction. Students' performance assessment tasks were scored by both holistic and analytic rubric. Firstly, the overall quality of literary analysis and argumentation were rated by a holistic rubric, followed by an analytic score with eight

items on specific writing and analysis skills. They found that instruction on literary analysis significantly improved students' ability to analyze and describe conflicts in literature, and the organization of writing led to significantly higher scores on measures of coherence and organization.

Though instructional sensitivity is not the focus of this paper, an important message lies in the fact that without an analytic rubric, the instructional sensitivity could not be measured. What is going to be examined will be the focus of teaching, affecting teachers' instructional practices (Bacha, 2001) and classroom assessment (Thompson & Newsome, 2002). This is called a backwash effect (Russikoff, 1995). Therefore, if holistic rubric is only used in LS public examination, further researches on instructional sensitivity or backwash effects might be difficult to be implemented. Hence, the bridge which links large scale assessment and classroom teaching can hardly be formed.

Discriminating Power

The assumption behind a rubric is that if a student cannot do the lower level task, he or she cannot do higher level tasks (Biggs & Collis, 1982). In a holistic rubric, several abilities are collapsed into the same scale, but students' performance might be varied within the same scale. Therefore, in the holistic rubric students' different performance might not be reflected in their scores.

On the other hand, analytic rubrics typically have higher discriminating power (Mendelsohn & Cumming, 1987). Pomplun, Capps, and Sundbye (1998) investigated the relationship between rubric related features and the scores obtained by mathematics and reading assessment in which a holistic rubric was employed. They found that the scores for correct answers across content areas and grade levels accounted for only a bit over 50% of the holistic score variance. They also found that the longer the response length, the higher the score that teachers tended to give that response. These suggested that the discriminating power of holistic rubric was quite low.

Also, Sadler and Donnelly (2006) investigated how content knowledge and morality contributed to the quality of socio-scientific argumentation

among 56 high school students. Their multiple regression analyses revealed no statistically significant relationships among the three. This suggested that students' scores were different across different domains. In this case, only analytic rubric may capture this difference.

Inter-rater Reliability

Analytic rubric is found to have higher inter-rater reliability in performance-based assessment. Jonsson and Svingby (2007) reviewed 75 studies about using scoring rubrics in performance-based assessment. They found that analytic, topic specific rubric, accompanied by exemplars and with rater training enhanced the reliability of scoring performance assessments.

In Chi's study (2001), he compared holistic and analytic rubrics (he used the term "scoring method") to explore the differences between them for performance assessment using a many-faceted Rasch model. Forty-three students' reports for social studies were scored by four raters with both holistic and analytic rubrics. He found that analytic rubrics had better inter-rater reliability. Chi (2001) reported the significance level of chi square test for using holistic rubric and analytic rubric respectively, i.e. 0.03 and 0.39. He concluded that significant differences were found between raters using holistic scoring rubrics, but not analytic rubrics.

Klein et al. (1998) compared the scores obtained by using holistic and analytic rubric in scoring responses of students in Grades 5, 8, and 10 on three dimensions of science performance tasks developed by the California State Department of Education in 1992. There were 168 Grade 5 students, 98 Grade 8 students, and 102 Grade 10 students and 4–5 pairs of readers to grade the students' work. Each pair of readers was given same sets of answers under both methods. The inter-reader correlations were summarized in Table 1. Raters using analytic rubrics had consistently higher inter-reader correlations across both three grades and three domains. This indicated that the inter-rater reliability was higher in using analytic rubric than in holistic one.

Table 1 Inter-reader Correlations by Grade Level, Standard, and Scoring Method (Klein et al., 1998)

	Conceptual		Performance		Application	
Grade	Analytic	Holistic	Analytic	Holistic	Analytic	Holistic
5	.76	.56	.86	.48	.70	.49
8	.73	.43	.77	.24	.64	.43
10	.75	.65	.82	.38	.70	.63

Bias-free

Davidson, Howell, and Hoekema (2000) found that teachers' standpoints would affect the scores they given. They tended to give higher scores to students who shared the same view. For holistic rubrics, an overall score is given to each response. Thus, if the scorer disagrees with the argument in the response, they might overlook students' performance in other area. Hence scorers' personal biases can be hidden more easily when using holistic rubrics (Harlen, 2005).

In addition to standpoint bias, the ability of language use might affect the scorers' impressions. In a course writing, faculty expect "content" to be the most important indicator of competence, followed by organization, language use, and mechanics at approximately the same degree of emphasis with vocabulary as the least important. However, Russikoff (1995) found that language use turned out to be the only factor which reliably predicted holistic scores in all analyses across both analytic and holistic instruments.

In LS, the aim is not to assess students' language proficiency, Russikoff's (1995) finding raised the issue of fairness in using holistic rubric in performance-based assessment. However, this bias could be reduced by using analytic rubric, because standpoints and language use can be designed as separate components in the scoring rubric. Students could get a relative fairer grade across different components. Thus, analytic rubrics help promote a fair grade for evaluation of essay writing.

Time Saving

It is generally believed that holistic rubrics saved more time than analytic rubrics (Arter, 1993; Bainer & Porter, 1992). Bauer (1981) compared the

relative reliable uses and cost effectiveness of the analytic and the holistic scoring methods by scoring secondary school students' essays of 1973–74 writing assessment of the National Assessment of Educational Progress. Raters were assigned to one of three groups, each using one scoring method exclusively. He found that time needed to train and to grade the essays were two times and four times by analytic scoring method than holistic one respectively. Though it has been mentioned before, scoring methods were not the same as scoring rubrics. This study was discussed here for reference only as it is interesting to note that in both studies, by using analytic method or rubric, time was 3 – 5 times more needed than holistic rubric.

Klein et al. (1998) also compared the time needed for scorers to grade the science performance tasks in Grade 5 and Grade 8. In Grade 5, an average of 17.5 min. and 6.4 min. per student were spent by using analytic and holistic rubric respectively; in Grade 8, the corresponding time was 14.6 and 3.1 min. respectively.

Discussion

Which Rubric is Suitable for NSS LS Public Examination?

From the comparison between analytic and holistic rubrics in Table 2, analytic rubrics are seemed to outperform holistic rubrics in performance assessment, except more time is saved in using holistic rubrics. LS public examination is a high-stake and performance assessment. Students' achievement about the subject will be evaluated at that particular examination, and the grades will affect their further work and study. Besides, school's performance is also reflected from their students' results in the high-stake examination. For such an important examination with no objective answers, students, principals, teachers and parents are all highly concerned about the fairness of the examination. Therefore, enhancing higher inter-rater reliability, higher discriminating power, and bias-free during scoring process is very important. From the previous analysis, analytic rubric is better than holistic rubric in these areas.

Table 2 Summary of Comparison of Analytic and Holistic Rubrics

	Analytic Rubrics	Holistic Rubrics
Impact	Show students' different performance in different dimensions	
Discriminative power	✓ Better	
Inter-rater Reliability	✓ More evidence	
Bias	✓ Fewer	
Time Saving		✓ 3 – 5 times saved

Furthermore, the aim of large scale assessment is not only to certify students' level of attainment, it also communicates what educators value in the subject (Arter, 1993; Gilfert & Harada, 1992). As LS is a newly developed subject, both teachers and students may not fully understand the curriculum and related learning and teaching issues, more time spending on enhancing scoring process is worthwhile. A well designed rubric could provide information that teachers need for instructional decision making and tracking student progress toward important learning outcomes. In this sense, an analytic rubric is better than a holistic rubric as different dimensions of expected students' performance can be clearly seen and it could help increase teachers' awareness of students' different performances.

Although from the literature review, analytic rubric appears to be more suitable for LS assessment, it is also important for educators to make effort in working on the development of the rubrics. Suggestions on further researches on development of use of rubrics in LS are explained in the following session.

Suggestions and Further Researches on Development of Use of Rubrics in LS

Large scale research based project. With reference to the 7-year project of performance assessment development in UCLA (Niemi et al., 2007), a large scale research-based project should be done by collaboration between universities, government and school districts to connect the performance-based assessment in public examination and classroom learning.

At teacher/school level, *teacher-researcher model* is suggested. To increase reliability and validity, the rubrics of public examination should not be developed out of the context of real classroom setting. In addition to try out the rubrics designed by the examiners or expert scorers, more teachers should get a chance in participating in the rubric development process. Teachers' own developed rubrics which are used in their regular teaching should be collected and tried out by public exam rubric developers and vice versa. Therefore, teachers are encouraged to take active role to study the backwash effect of use of rubrics in performance-based assessment to their teaching.

Moreover, teachers should investigate how to engage students' learning to learn by use of analytic rubric. By using analytic rubrics, student might be easier to grasp the gist of they need to learn. Beaudry (1997) led a team to investigate how the use of holistic rubrics affecting student performance in reading and writing for the Maine Educational Assessment in 1992. He suggested that students' achievement could be improved by engaging them in classroom assessment by using holistic rubrics to evaluate their own work.

Accommodation of students with special learning needs. For students with special learning needs, e.g., dyslexia, they are disadvantaged by the written performance tasks assessment. Would there be other rubrics for scoring their work?

Holistic + Analytic Rubrics. Combination of holistic and analytic rubrics could be used in the same performance-based assessment, depending on the nature of the tasks being assessed. Baker (1994) launched a project to design and validate new assessments of history understanding and to learn about assessment design and validation in general. He did not simply use either holistic rubric or analytic rubric. There were six scoring elements in grading their students' history essays, which were overall content quality, prior knowledge, principles/themes, text detail, misconceptions, and argumentation. Some elements were scored holistically (e.g., overall content quality) while some were scored analytically (e.g., text detail).

Another example is a specific holistic rubric designed for assessing

argument presentation. Further researches could be done to see which type of rubrics work well on which type of questions. In the NSS LS curriculum guide, one of the students' performance items to be assessed is their argument presentation (p.121, table 5.1, e). If there is a separate domain in the scoring rubric, e.g., Toulmin Argument Model (0: No Justification; 1: Justification with no grounds; 2: Justification with simple grounds; 3: Justification with elaborated grounds; 4: Justification with elaborated grounds and a counter-position, Sadler, 2006), teachers could give the score specifically on students' argumentative performance. Toulmin Argument Pattern provides a framework for analyzing argument structure with references to features like claims data, warrants, backings, and rebuttals (Sadler, 2006). Students could learn from their scores to know how well they perform on making argument and will know the direction of improvement.

Raters' training. Every LS teacher should have a chance to get training in scoring the public examination since scoring experience enhanced their teaching. Goldberg and Roswell (2000) gathered teachers' reported ratings of knowledge about performance-based instruction and assessment among 37 teachers in Maryland school performance assessment program. They found that there is no significant differences in ratings between the 20 teachers who had scored for 1 year only and the 17 scorers who had 2 or more years' experience.

From their lesson observation, Goldberg and Roswell (2000) found that lessons developed by teachers without scoring experience demonstrate limited efforts to establish a context and purpose for performance tasks. Academic exercise was given instead. In contrast, teachers with scoring experience are able to design performance tasks more coherent to context in real world setting. However, the number of years of scoring experience that the teachers have was not mentioned in their paper. Furthermore, for teachers who had training in scoring, they applied the knowledge that they learnt in their daily assessment strategies. Moreover, the strategies were coherent to the rubrics that they used in the training session.

Conclusion

The rubric for the LS (NSS) examination in 2012 should be a fair and informative one. From the above review of past studies, an analytic rubric is found to be a fair and more informative grading tool than holistic rubric. Studies showed that an analytic rubric has higher inter-rater reliability, has higher discriminating power, gives more information, and reduces rater bias. Thus, an analytic rubric is more favorable than a holistic rubric. However, this might be costly in terms of large number of raters and the time needed to mark the examination papers, but this investment might be worthwhile.

This paper just gives a start to the road of exploration of the design of rubrics. Further studies are suggested to be done in implementing a large scale research project for facilitating learning from performance-based assessment, accommodation of students with special learning difficulty, combination of holistic and analytic rubric and comprehensive raters' training.

Acknowledgements

The author would like to thank Professor Ming Ming Chiu of the Department of Educational Psychology, CUHK and Ms. Wing-suet Ng of the University of Hong Kong for their many helpful comments, support and guidance.

References

Arter, J. (1993, April). Designing scoring rubrics for performance assessments: The heart of the matter. Paper Presented at the Annual meeting of the American Educational Research Association, Atlanta, GA. ERIC Document No. ED 358 143.

Bacha, N. (2001). Writing evaluation: What can analytic versus holistic essay scoring tell us? *System, 29*, 371–383.

Bainer, D., & Porter, F. (1992, October). *Teacher concerns with the implementation of holistic Scoring*. Paper presented at the annual meeting of the Midwestern Educational Research Association, Chicago. ERIC Document No. ED 355 511.

Baker, E. L. (1994). Learning-based assessments of history understanding. *Educational Psychologist, 29*, 97–106.

Baker, E. L. (2007). Model-based assessments to support learning and accountability: The evolution of CRESST's research on multiple-purpose measures. *Educational Assessment, 12*, 179–194.

Bauer, B. A. (1981). A study of the reliabilities and the cost-efficiencies of three methods of assessment for writing ability. Champaign: University of Illinosis.

Beaudry, J. (1997). *Maine education at a glance MEG-07: Does the use of holistic rubrics affect student performance in reading and writing?* Portland: University of Southern Maine, Center for Educational Policy, Applied Research and Evaluation.

Biggs, J. B., & Collis, K. F. (1982). The psychological structure of creative writing. *Australian Journal of Education, 26*(1), 59–70.

Black, P., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2004). Working inside the black box: Assessment for learning in the classroom. *Phi Delta Kappan,, 86*, 9–21.

Chi, E. (2001). Comparing holistic and analytic scoring for performance assessment with many-facet Rasch model. *Journal of Applied Measurement, 2*(4), 379–388.

Criswell, J. R., & Criswell, S. J. (2004). Asking essay questions: Answering contemporary needs. *Education, 124*(3), 510–516.

Curriculum Development Council Hong Kong Examinations and Assessment Authority [CDC & HKEAA]. (2006). *Provisional final draft of liberal studies in new senior secondary curriculum and assessment Guide S4–6*. Hong Kong: Government Printer.

Curriculum Development Council Hong Kong Examinations and Assessment Authority [CDC & HKEAA]. (2007). *Liberal Studies Curriculum and Assessment Guide*. Hong Kong: Government Printer.

Davidson, M., Howell, K. W., & Hoekema, P. (2000). Effects of ethnicity and violent content on rubric scores in writing samples. *Journal of Educational Research, 93*(6), 367–372.

Education and Manpower Bureau [EMB]. (2000). *Reform proposal for the education system in Hong Kong*. Retrieved Jan 12, 2008 from *www.hkeaa.edu.hk/files/pdf/hkdse_lib_std.pdf*

Facione, P. A., Sanchez, C. A., & Facione, N. C. (1994). Are college students disposed

to think? ERIC Document No. ED 368 311.

Finson, K. D., & Ormsbee, C. K. (1998). Rubrics and their use in inclusive science. *Intervention in School & Clinic, 34*(2), 79.

Gilfert, S., & Harada, K. (1992). Two composition scoring methods: The analytic vs. holistic method. *Bulletin of Faculty of Foreign Languages, 1*, 17–22.

Goldberg, G. L., & Roswell, B. S. (2000). From perception to practice: The impact of teachers' scoring experience on performance-based instruction and classroom assessment. *Educational Assessment, 6*(4), 257–290.

Gopinath, C. (2004). Exploring Effects of Criteria and Multiple Graders on Case Grading. *Journal of Education for Business, July/August*, 317–322.

Harlen, W. (2005). Trusting teachers' judgement: Research evidence of the reliability and validity of teachers' assessment used for summative purposes. *Research Papers in Education, 20*(3), 245–270.

Harmon, M., Smith, T. A., Martin, M. O., Kelly, D. L., Beaton, A. E., Mullis, I. V. S., et al. (1997). Performance assessment in IEA's third international mathematics and science study. Retrieved Jan 12, 2008, from http://timss.bc.edu/timss1995i/PAreport.html.

Herman, J. L., Aschbacher, P. R., & Winters, L. (1992). A practical guide to alternative assessment. Alexandria, VA: Association for Supervision and Curriculum Development.

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

Kelly-Riley, D., Brown, G., Condon, B., & Law, R. (2001). Washington State University critical thinking project. Retrieved December 7, 2006, from http://wsuctproject.wsu.edu.

Klein, S. P., Stecher, B. M., Shavelson, R. J., McCaffrey, D., Ormseth, T., Bell, R. M., Comfort, K., & Othman, A. R. (1998). Analytic versus holistic scoring of science performance tasks. *Applied Measurement in Education, 11*(2), 121–137.

Lloyd-Jones, R. (1977). Primary trait scoring. In C. R. Cooper & L. Odell (Eds.), *Evaluating writing: Describing, measuring, judging*(pp. 33–66). Urbana, IL: National Council of Teachers of English.

Mendelsohn, D., & Cumming, A. (1987). Professors' ratings of language use and rhetorical organization in ESL compositions. *TESL Canada Journal, 5*(1), 9–26.

Moon, T. R., Callahan, C. M., Brighton, C. M., & Tomlinson, C. A. (2002). *Development of differentiated performance assessment tasks for middle school classrooms*. Storrs, CT: National Research Center on the Gifted and Talented. (ERIC Document No. ED 476 371)

Mott, M. S., Etsler, C., & Drumgold, D. (2003). Applying an analytic writing rubric to children's hypermedia 'narratives'. *Early Childhood Research & Practice, 5*(1). Retrieved 10 December, 2006, from http://ecrp.uiuc.edu/v5n1/mott.html

Ni, Y. J. (1997). Performance-based assessment: Problems and design strategies. *Education Journal, 25*, 137–157.

Niemi, D., Baker, E. L., & Sylvester, R. M. (2007). Scaling up, scaling down: Seven years of performance assessment development in the nation's second largest school district. *Educational Assessment, 12*, 195–214.

Niemi, D., Wang, J., Steinberg, D. H., Baker, E. L., & Wang, H. (2007). Instructional sensitivity of a complex language arts performance assessment. *Educational Assessment, 12*, 215–237.

Perlman, C. (2002). An introduction to performance assessment scoring rubrics. In C. Boston (Ed.), *Understanding scoring rubrics: A guide for teachers*. College Park, MD: University of Maryland, ERIC Clearinghouse on Assessment and Evaluation. (ERIC Document No. ED 471 518)

Pomplun, M., Capps, L., & Sundbye, N. (1998). Criteria teachers use to score performance items. *Educational Assessment, 5*(2), 95–110.

Russikoff, K. A. (1995, March–April). A comparison of writing criteria: Any differences? Paper Presented at the annual meeting of the Teachers of English to Speakers of Other Languages, Long Beach, CA. (ERIC Document No. ED 386 025)

Sadler, T. D. (2006). Justification of socioscientific claims as the basis for assessing argumentation. In *Proceedings of the 7th international conference on learning sciences* (pp. 978–979). Bloomington, IN: International Society of the Learning Sciences.

Sadler, T. D., & Donnelly, L. A. (2006). Socioscientific argumentation: The effects of content knowledge and morality. *International Journal of Science Education, 28*(12), 1463–1488.

Schultz, R. A. (2002). Teachers as learners: Studying a three-phased rubric assessment plan. *Gifted Child Today, 25*(4), 38–45.

Simon, M., & Forgette-Giroux, R. (2001). A rubric for scoring postsecondary

academic skills. *Practical Assessment, Research & Evaluation, 7*(18). Retrieved December 7, 2006, from http://PAREonline.net/getvn.asp?v=7&n=18.

Thompson, T., & Newsome, K. (2002). Is state assessment a viable tool for reflection of classroom assessment? (ERIC Document No. ED 471757)

Tufte, R. (2005). The P.A.C.E.S. grading rubric: Creating a student-owned assessment tool for projects— the design brief brings out all kinds of "out of the box" thinking, with many correct answers to solve the problem. *Technology Teacher, 64*(5), 21–2.

Westat. (2001). *ABC on scoring rubrics development for large scale performance.* Retrieved Jan 12, 2008, from http://lsc-net.terc.edu/do.cfm/paper/8315/show/use_set-stud_assess/page-1.

Appendix

The 15 assessment objectives of LS listed in the *Liberal Studies Curriculum and Assessment Guide* (CDC & HKEAA, 2007, pp. 123–124)**:**

1. demonstrate a sound understanding of the key ideas, concepts and terminologies of the subject;
2. make conceptual observations from information resulting from enquiry into issues;
3. apply relevant knowledge and concepts to contemporary issues;
4. identify and analyze the interconnectedness and interdependence amongst personal, local, national, global and environmental context;
5. recognize the influence of personal and social values in analyzing contemporary issues of human concern;
6. draw critically upon their own experience and their encounters within the community, and with the environment and technology;
7. discern views, attitudes and values stated or implied in any given factual information;
8. analyze issues (including their moral and social implications), solve problems, make sound judgments and conclusions and provide suggestions, using multiple perspectives, creativity and appropriate thinking skills;
9. interpret information from different perspectives;
10. consider and comment on different viewpoints in their handling of different issues;
11. self-manage and reflect upon the implementation of successive stages of the enquiry learning process in terms of time, resources and attainment of the objectives of the enquiry;
12. communicate clearly and accurately in a concise, logical, systematic and relevant way;
13. gather, handle and analyze data and draw conclusions in ways that facilitate the attainment of the objectives of the enquiry;
14. demonstrate an understanding and appreciation of different cultures and universal values; and
15. demonstrate empathy in the handling of different issues.