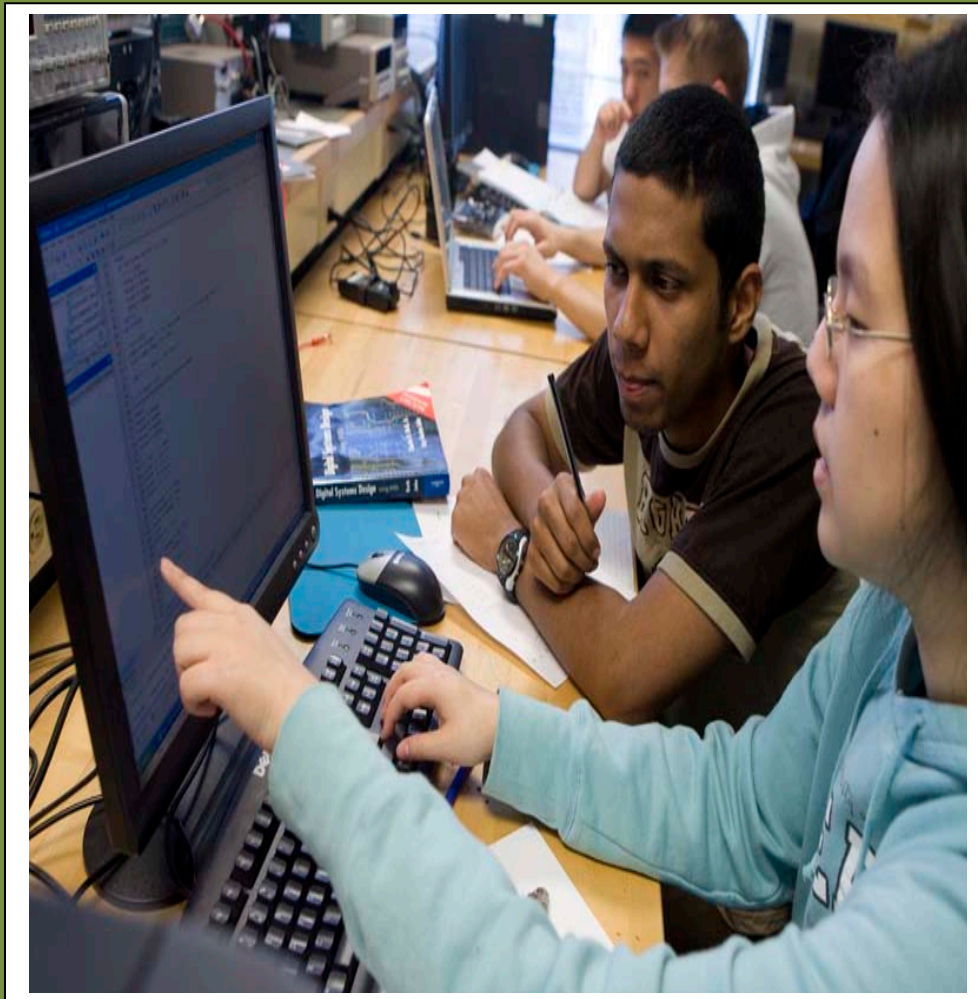


Office of Institutional Research & Assessment
August 2011

Results of Math-Reasoning Rubric Analysis



BINGHAMTON
UNIVERSITY
STATE UNIVERSITY OF NEW YORK

Executive Summary

The Office of Institutional Research & Assessment recruited 5 instructors to apply a Faculty Senate-approved rubric to evaluate overall student performance. Each instructor chose one course and a representative assignment in each course, used the rubric, and submitted the scores to OIRA for analysis. 141 student scores were recorded. We found that students performed satisfactorily on all five elements of the rubric (median scores were all 3.0 on a 4-point scale on all elements) and that there were no significant differences between transfer (n = 24) and native students' (n = 116) scores or between scores of upper-level and lower-level students. However, the quantitative findings and follow-up interviews with the instructors who used the rubric suggested that students might improve with regard to using mathematical and logical information beyond merely churning out formulas. These findings will be communicated to the assessment category team in mathematics-reasoning, who will write a report regarding the student learning outcomes in math-reasoning after using these results, the results of the ETS Proficiency Profile®, information gleaned from general education course portfolios, and alumni, senior, and faculty survey results. The report is due in Fall 2011.

Introduction

In 2006, Binghamton University participated in SUNY's "Strengthening Campus-Based Assessment" program, in which the campus agreed to apply rubrics in three subject areas: composition, critical thinking, and mathematics-reasoning. The rubrics were developed by SUNY faculty throughout the system and were approved for use at the campus level through campus-based faculty senate processes (Francis, 2006). In 2006, the Binghamton University Faculty Senate Executive Committee approved the use of the math-reasoning rubric.

Since 2006, the math-reasoning rubric was used against a small sample of student papers collected from Math 130 ("Mathematics in Action") courses. The results revealed that students needed more practice recognizing the limitations of formulas used in solving mathematical problems and moving beyond rote memorization toward application of mathematical information in different situations. Interestingly, these findings reflected the findings of a similar rubric exercise in critical thinking that students needed to be challenged more in the areas of evaluating and synthesizing information.

Since that time, the Provost's Office and Office of Institutional Research & Assessment have communicated these findings to various units on campus (e.g., Deans' offices, the Institute of Student-Centered Learning) in the hopes of enhancing student learning in

this area. The purpose of this white paper is to present the findings of the second cycle of assessing student learning in math-reasoning.

Procedure

As stipulated by the Faculty Senate-approved procedure, the Office of Institutional Research & Assessment (OIRA) recruited instructors teaching Math 130. We also included one Math 108 course (Trigonometry) and one Math 220 course (Business Calculus) because they are considered beginning courses for students in various fields of study. We deviated from the original Faculty Senate procedure of limiting our collection of rubric results to Math 130 because a number of schools and colleges (and departments) at the university have suggested over the years that a more representative sampling of beginning level math courses would make the results of this study more meaningful. We also made a decision to encourage instructors to apply the rubric to their own students' work in order to train faculty on the use of rubrics and to ground a follow-up focus group discussion in the instructors' own experiences. At the beginning of Spring 2011 we asked five instructors to participate and all agreed to do so.

The Assistant Provost & Director of OIRA met with each instructor to discuss the rubric and train him or her on how to utilize it. Each instructor described the type of assignment that would be used for the exercise and discussion ensued regarding which assignments were appropriate for use with the rubric. During the Spring semester (2011), each instructor applied the rubric to an appropriate assignment and then submitted each assignment, with the ratings, to OIRA. OIRA staff then inputted the rubric results onto a database using SAS© software to conduct the rubric analysis for this report.

Unlike the rubric exercise in math-reasoning in 2008, instructors were asked to include student identifiers on each assignment so that student ratings on the rubric could be matched with information we have for each student in the university's student information system. This was beneficial to the overall analysis for a number of reasons. First, by merging student information with rubric results we were able to generate a table of student characteristics including data on gender, parental educational background, native vs. transfer student status, and SAT scores. Second, we were able to conduct deeper analyses (e.g., logistic regression analyses) with several variables gleaned from the university's student database system.

We also collected qualitative information to gain a better understanding of trends instructors were seeing in student performance. We asked instructors to write their general impressions of student performance, using the rubric elements as reference

points. In addition, we interviewed 3 of the 5 rubric evaluators after they evaluated students about their overall impressions of students' strengths and weaknesses in performing well on the five elements of the rubric. This process greatly assisted us in making conclusions about what the analysis meant with regard to the overall student learning goals in mathematics/reasoning.

Research Questions

The research questions for this analysis are as follows:

1. What are the basic results of the rubric evaluations, and how do they compare to the 2008 rubric analysis in mathematics/reasoning?
2. How do transfer and native students' scores compare?
3. How do the scores of first year and sophomore students compare to those of junior and senior students?

Math Reasoning Rubric

The math reasoning rubric (see Appendix 1) requires instructors to evaluate student work samples with regard to the following elements:

- *Interpreting* mathematical models
- *Representing* mathematical information
- *Employing* quantitative methods
- *Estimating/checking results* for reasonableness
- *Recognizing limits* of mathematical methods

Each assignment was evaluated on a four-point scale:

- Completely correct
- Generally correct
- Partially correct
- Incorrect

When inputting information onto a spreadsheet, "completely correct" was denoted as 4 points, "generally correct" as 3, "partially correct" as 2, and "incorrect" as 1 point. In training conversations with each instructor, emphasis was placed on understanding the differences between each rating level. Completely correct meant no errors, generally correct meant that there were some errors but that these did not interfere with students' generating a correct final solution to each problem, partially correct meant that there were significant errors but the student was able to demonstrate partial ability

to come to a mathematical conclusion, and incorrect meant that a student could not demonstrate abilities suggested by each element. Follow-up interview discussions revealed few if any problems associated with applying the rubric to the student work samples that were self-selected by the instructors.

Findings

General Findings. A total of 141 student work samples were collected and evaluated using the math-reasoning rubric. As shown in Table 1, 83% of the students whose work was evaluated were native students, 44% were female, and approximately 26% were under-represented minorities. With regard to class level, 61% of students who were evaluated were first-year and sophomore students, 27% came from homes whose fathers had little or no collegiate experience, and 36% whose mothers came from this same category. Native students' composite SAT scores were approximately 1140, lower than the average for Binghamton University students.¹ Cumulative GPA was 2.9.

¹ SAT scores were not available for transfer students because they are not required to submit SAT scores as part of the admissions process like native students are.

Table 1. Demographic Characteristics

Characteristic	Variable	n	percent	M	Min	Max	SD
Students	Native	116	82.86				
	Transfer	24	17.14				
Sex	Male	79	56.03				
	Female	62	43.97				
Origin	US	125	88.65				
	International	16	11.35				
Ethnicity	Asian/PI	19	13.48				
	Black/NH	15	10.64				
	Hispanic	21	14.89				
	NRA	18	12.77				
	Unknown	15	10.64				
	White/NH	53	37.59				
Class	First-Year	46	32.62				
	Sophomore	40	28.37				
	Junior	31	21.99				
	Senior	24	17.02				
Father's Educational Background	1 or 2	29	27.10				
	3 or 4	78	72.90				
Mother's Educational Background	1 or 2	39	35.78				
	3 or 4	70	64.22				
Age		141		19.87	18.00	48.00	2.96
SAT Verbal		126		553.29	320.00	760.00	96.12
SAT Math		126		585.60	360.00	790.00	92.65
Cumulative GPA		137		2.90	0.00	4.00	0.68
Family Size		109		3.74	1.00	6.00	1.16

Raters' consistency when using the rubric was evaluated using Cronbach's alpha tests of consistency. No alpha was lower than .80 on all elements of the rubric, suggesting acceptable levels of consistency (D'Antoni et al., 2009; Jonsson & Svingby, 2007).² We also observed correlations between students' scores on the rubric and the course grades faculty raters gave to students in the courses from which their work samples were collected. Given statements by faculty raters in follow-up discussions that they felt

² Generally, a Cronbach's alpha higher than .70 establishes acceptable levels of reliability (Bresciani, et al., 2009).

students' performance on the rubric was associated with their overall performance in the math class they took, these correlations suggest that the rubric scores were valid and reliable estimates of students' math-reasoning skills. As Table 2 indicates, these correlations ranged from a low of .33 to a high of .86, with no inverse (negative) correlations.

Table 2. Correlations between Course Grades and Question Items

	<u>Interpreting</u>	<u>Representing</u>	<u>Employing</u>	<u>Estimating</u>	<u>Recognizing</u>	<u>Total Avg.</u>
Course Grade	0.61	0.49	0.46	0.42	0.48	0.62

Table 3 indicates that students' performance on their work assignments was between "generally correct" and "partially correct" on most elements of the rubric. All medians were 3.0. Students performed best on "recognizing limits of mathematical methods," and not as well on "interpreting mathematical models."

Table 3. Math Reasoning Rubric Results

	<u>N</u>	<u>Mean</u>	<u>SD</u>
Interpreting mathematical models	141	2.65	1.1
Representing mathematical information	141	2.78	1.21
Employing quantitative information	141	2.89	0.969
Estimating and checking for reasonableness	141	2.98	1.09
Recognizing limits of mathematical methods	141	3.14	0.975
Overall Rubric Average	141	2.89	0.836

Table 4. Comparisons of 2008 and 2011 Math Reasoning Rubric Results—Percentage Scoring ‘Completely Proficient’ and ‘Generally Proficient’

	<u>2008</u>	<u>2011</u>
Interpreting mathematical models	96.7%	57.5%
Representing mathematical information	86.7%	58.9%
Employing quantitative information	60.0%	63.1%
Estimating and checking for reasonableness	53.3%	66.7%
Recognizing limits of mathematical methods	37.9%	74.5%

The results depicted in Table 4 show differences between the 2008 and 2011 results of students who work was evaluated using the math-reasoning rubric who scored either completely or generally proficient. Although there were significant concerns with students’ abilities to recognize limits of mathematical methods, this is less of a concern now. In contrast, this year’s evaluators are more concerned with students’ abilities to interpret mathematical models than evaluators were in 2008, and 2011 evaluators appear to be more concerned with the extent to which students represent mathematical information when solving word problems in math.

In follow-up discussions with the evaluators, it became clear that they were mostly concerned with students’ ability to successfully use mathematical information in different situations. That is, in their view students tend to perform well with regard to applying formulas and churning out answers, but when asked to apply formulas in different situations, they perform less well.

Transfer Students. Understanding differences in the performance of transfer students has increasingly become a salient issue because of recent SUNY initiatives to increase this important sub-population. Because we have been able to increase the sample size for this rubric evaluation when compared to prior years, we are able to address the issue of whether or not the performance of native versus transfer students differs significantly based on the rubric evaluations used for this analysis.

Table 5. Comparison of Native (n=116) and Transfer Student (n=24) Rubric Results

	<u>Native Students</u>		<u>Transfer Students</u>	
	<u>Mean</u>	<u>SD</u>	<u>Mean</u>	<u>SD</u>
Interpreting mathematical models	2.72	1.09	2.42	1.14
Representing mathematical information	2.82	1.21	2.67	1.20
Employing quantitative information	2.90	1.01	2.92	.78
Estimating and checking for reasonableness	3.00	1.09	2.79	1.06
Recognizing limits of mathematical methods	3.14	.98	3.17	1.01
Overall Rubric Average	2.92	.84	2.79	.81

Table 5 depicts the results of the analysis. Native students outperformed transfer students in three of the rubric elements while transfer students outperformed native students on two of the rubric elements. Statistical tests of differences between these two groups reveal no significant differences, although it could be said that a lack of differences is due to the difference in sample sizes (the sample size for native students is much larger than the sample size for transfer students).

Class Levels. Because the data collection process association with this analysis could not practically divide students by class level (first-year student, sophomore, etc.) within each course chosen for this rubric assessment, it was important to investigate the extent to which performance on the rubric elements differed by class level. We maximized sample size by dividing the sample into two levels: upper-level students (junior and senior students) and lower-level students (first-year and sophomore students).

Table 6. Comparison of Lower Level (n = 86) and Upper-Level Students (n = 55)

	<u>Lower-Level Students</u>		<u>Upper-Level Students</u>	
	<u>Mean</u>	<u>SD</u>	<u>Mean</u>	<u>SD</u>
Interpreting mathematical models	2.64	1.14	2.67	1.06
Representing mathematical information	2.80	1.23	2.75	1.21
Employing quantitative information	2.84	1.06	2.98	0.80
Estimating and checking for reasonableness	2.90	1.14	3.11	1.01
Recognizing limits of mathematical methods	3.06	1.04	3.27	.85
Overall Rubric Average	2.85	0.92	2.96	0.68

As depicted in Table 6, upper-level students performed better on all elements of the rubric and on the overall rubric average except for “representing mathematical information.” However, tests of differences between medians did not demonstrate significant differences between these two groups.

Limitations

In 2008, inter-rater reliability was moderate to high, and because student work samples were randomly selected and not associated with the raters themselves, there were few issues associated with moderator effects, meaning that there were few presumptions that raters would rate students higher on the rubric simply because they were grading their own students’ papers. However, the primary problem three years ago was that efforts to establish inter-rater reliability were expensive, which required us to work with a small sample of students. Given the funding for the project, only forty work samples could be graded by three instructors who were contracted to evaluate the same student work samples.

In this exercise, five instructors were recruited who submitted 141 samples, substantially increasing the sample size from that of three years ago. The primary limitations associated with this approach are two-fold. First, because instructors were

asked to evaluate their own students' work, we cannot assume that there were few moderator effects. Second, in order to increase the overall sample size, and given funding limitations, we could not establish inter-rater reliability as we did before.

However, as we describe above, we endeavored to address both validity and reliability issues by using a few well-established statistical procedures. We used Cronbach's alpha scores to measure overall reliability and associated instructors' scores on the rubric with course grades given to each student in the courses from which the work samples were collected.

Finally, it is beyond the scope of this study to compare grades given to transfer and native students in all m-designated courses. This study reviews math competencies with regard to the elements of the specific rubric employed to evaluate student work samples and does not include students in courses above Business Calculus (Math 220). Generally, grades are an assessment of many various performance criteria, including participation, attendance, quiz score performance, among others. In the future, it might be interesting to look at these differences, but it is not the purpose of this study to make inferences about course performance as opposed to performance in specific outcomes that relate to the math-reasoning rubric.

Discussion

Generally, it appears that students perform satisfactorily on all elements of the rubric if we declare a median of 3 ("generally correct") as the benchmark expectation for performance. Means give us more ability to interpret variance between the five elements of the rubric, but because student work samples were based on an ordinal scale, judging student performance using medians as a basis is more appropriate.

That said, these findings suggest that further action might be needed with regard to assisting students in general education math courses in their application of mathematical information and moving beyond formulaic responses to problem solving. This is consistent with the findings of the critical thinking rubric analysis during the 2009-10 academic year. This analysis found that students need assistance with regard to gathering, analyzing, and synthesizing information and moving beyond summarizing information. It appears that a review of rubric analyses in both math-reasoning and critical thinking suggests that some students may need to improve in their ability to use information outside self-imposed parameters—using information in ways that move beyond formulaic, logical calculations.

With regard to comparing performance in math reasoning of transfer and native students, the above findings are consistent with a sample of self-selected seniors who completed the ETS Proficiency Profile© examination. Although the sample was not random and was based on a convenience sample, an analysis of both math scaled scores and estimates of proficiency on that examination revealed that there are no significant differences between transfer and native students.³ In addition, when we performed a logistic regression of the math-reasoning rubric results, with pass/fail on the rubric as the dependent variable⁴ and transfer student status and undergraduate GPA as independent variables, we could not find any significant differences.

It was also interesting to note that we could find no differences with regard to students in their first and second year of study (defined by number of credits earned) versus students who were classified as juniors and seniors. Ordinarily, one would expect that students with more collegiate experience would score better on a rubric exercise such as this, but we note also that this sample includes students taking primarily 100-level courses, so it is likely that students did not differ by class standing because they were more equal with regard to their overall level of math ability.

Conclusion

Although this study is limited in scope because it does not include math courses beyond the basic level, its results reflect prior findings that some students might benefit from being challenged to use information beyond their comfort zone. The message of this analysis is that students need to move beyond formulaic responses in their efforts to solve problems, a finding that is not unique to Binghamton University students.

The findings of this study will be shared with the math-reasoning assessment category team (ACT), which will write a report on the student learning goals in math-reasoning as part of the university's general education program of study. This study is only one part of additional information the ACT will use in writing its report, but our hope is that this analysis will augment discussions that the ACT will have about overall student performance.

³ 121 native and 78 transfer students completed the short version of the ETS Proficiency Profile© over a two-year period. The results demonstrated that 89% of native students were highly proficient in the most basic levels of mathematics proficiency (the first of three levels) compared to 78% of transfer students.. On the scaled score, the difference was not found to be significant ($p < .07$). Although the significance level was below .10, it is higher than the expected .05 level of significance before controlling for SAT score and background variables such as ethnicity and parental educational background.

⁴ Pass means that a student on average received a 4 or 3 on all the elements on the rubric; failed means they received a 1 or 2.

Student Name:

Evaluator Name:

NOTE: Student name is confidential, and is only used for data purposes.

Mathematics and Reasoning Rubric				
	<u>Completely Proficient</u>	<u>Generally Proficient</u>	<u>Partially Proficient</u>	<u>Not Proficient</u>
1. Interpreting mathematical models				
2. Representing mathematical information				
3. Employing quantitative methods				
4. Estimating/checking results for reasonableness				
5. Recognizing limits of mathematical methods				

NOTE: Please return to Sean McKittrick, Office of Institutional Research & Assessment, AD-305. If desired, please make comments on other side of this evaluation sheet.

References

- Aberdeen, S.M., Leggat, S.G., & Barraclough, S. (2009). Validating a marking rubric for evaluating staff knowledge of dementia for competency in residential aged care. *Journal of Vocational Education and Training*. 61(4), 535-552.
- Ammons, J. & Mills, S. (2005). Course-embedded assessments for evaluating cross-functional integration and improving the teaching-learning process. *Issues in Accounting Education*. 20(1), 1-19.
- Banta, Trudy W. (2007). Can assessment for accountability complement assessment for improvement? *Peer Review*, Spring, 9-12.
- Beyreli, L., & Ari, G. (2009). The use of analytic rubrics in the assessment of writing performance: Inter-rater reliability concordance study. *Educational Sciences: Theory & Practice*. 9(1), pp. 105-125.
- Blommel, M.L. & Abate, M.A. (2007). Instructional design and assessment: A rubric to assess critical literature evaluation skills. *American Journal of Pharmaceutical Education*. 71(4), pp. 1-8.
- Bresciani, M.J., Oakleaf, M., Kolkhorst, F., Nebeker, C., Barlow, J., Duncan, K., & Hickmott, J. (2009). Examining design and inter-rater reliability of a rubric measuring research quality across multiple disciplines. *Practical Assessment, Research, & Evaluation*. 14(12), 1-7.
- Choinski, E., Mark, A.E., & Murphey M. (2003). Assessment with rubrics: Efficient and objective means of assessing student outcomes in an information resources class. *Libraries & the Academy*. 3(4): pp. 563-575.
- Ciorba, C.R. & Smith, N.Y. (2009). Measurement of instrumental and vocal undergraduate performance juries using a multidimensional assessment rubric. *Journal of Research in Music Education*. 57(1), 15-15.
- Connors, P. (2008). Assessing written evidence of critical thinking using an analytic rubric. *Journal of Nutrition Education & Behavior*. 40(3), 193-194.
- Daiker, D.A. & Grogan, N. (1985). The selection and use of sample papers in holistic evaluation. Oxford, OH: Miami University. (ERIC Document Reproduction Service No. ED305391).
- D'Antoni, A., Zipp, G.P., & Olson, V.G. (2009). Interrater reliability of the mind map assessment rubric in a cohort of medical students. *BMC Medical Education*. 9, 1-8.
- Englehard, G., Jr. (1996). Evaluating rater accuracy in performance assessments. *Journal of Educational Measurement*. 33, 56-70.

- Gadbury_Amyot, C.C., Kim, J., Palm, R.L., Mills, G.E., Noble, E., & Overman, P.R. (2003). Validity and reliability of portfolio assessment of competency in a baccalaureate dental hygiene program. *Journal of Dental Education*. 67(9), 991-1002.
- Gerretson, Helen & Emily Golson. (2004). Introducing and evaluating course-embedded assessment in general education. *Assessment Update*. 16(6), 4-6.
- Flowers, C. (2006). Confirmatory factor analysis of scores on the clinical experience rubric: A measure of dispositions for preservice teachers. *Educational and Psychological Measurement*. (66), 478-488.
- Francis, P.L., Salins, P.D., & Huot, A.E. (2006). The SUNY assessment initiative: Meeting standards of good practice. *Assessment Update*. 18(1), pp. 1-2, 13.
- Jonsson, A. & Svingby, G. (2006). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review*. 2, 130-144.
- Kan, A. (2007). An alternative method in the new educational program from the point of performance-based assessment: Rubric scoring scales. *Educational Sciences: Theory & Practice*. 7(1), 144-152.
- Kerby, D. & Romine, J. (2009). Develop oral presentation skills through accounting curriculum design and course-embedded assessment. *Journal of Education for Business*. (85), 172-179.
- Mansilla, V.B., Duraisingh, E.D., Wolfe, C.R., & Haynes, C. (2009). Targeted assessment rubric: An empirically grounded rubric for interdisciplinary writing. *The Journal of Higher Education*. 80(3), 334-353.
- McReal, T.L. (1990). The use of rating scales in teacher education: Concerns and recommendations. *Journal of Personnel Evaluations in Education*. 4, 41-58.
- Moskal, B.M. & Leydens, J.A. (2000). Scoring rubric development: Validity and reliability. *Practical Assessment, Research, & Evaluation*. 7(10). Retrieved July 20, 2011 from <http://PAREonline.net/getvn.net/getvn.asp?v=7&n=10>.
- Osborn Popp, S. E. & Thompson, M.S. (2009). The critical role of anchor paper selection in writing assessment. *Applied Measurement in Education*. 22, 255-271.
- Peat, B. (2003). Integrating writing and research skills: Development and testing of a rubric to measure student outcomes. *Journal of Public Affairs Education*. 12(3), 295-311.
- Petkova, O., Petkov, D., & D'Onofrio, M.J. (2008). Interweaving rubrics in information systems program assessments-Experiences from action research at two universities. *Issues in Informing Science and Information Technology*. 5, 423-432.

- Reddy, Y.M. & Andrade, H. (2010). A review of rubric use in higher education. *Assessment & Evaluation in Higher Education*. 3(4), 435-448.
- Rezaei, A. R. & Lovorn, M. (2010). Reliability and validity of rubrics for assessment through writing. *Assessing Writing*. 15, 18-39.
- RiCharde, R.S. (2009). Response to Arend Flick. *Assessment Update*. 21(1), 3.
- Shavelson, Richard J. (2007). Assessing student learning responsibly: From history to audacious proposal. January/February, 26-33.
- Stanford, M.R., Gras, L., Wade, A., & Gilbert, R.E. (2002). Reliability of expert interpretation of retinal photographs for the diagnosis of toxoplasmosis and retinochoroiditis. *British Journal of Ophthalmology*. 86(6), 636-639.
- Stellmack, M.A., Konheim, K. Kalkstein, Y.L., Manor, J.E., Massey, A.R., & Shmitz, A.P. (2009). Assessment of reliability and validity of a rubric grading APA-style introductions. *Teaching of Psychology*. 36, 102-107.
- Thaler, N., Kazemi, E. & Huscher, C. (2009). Developing a rubric to assess student learning outcomes using a class assignment. *Teaching of Psychology*. 36, 113-116.
- Tierney, J, & 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).