The WAC Journal 23 (2012). © 2013 by Clemson University.
Copies may be circulated for educational purposes only.

Crossing the Measurement and Writing Assessment Divide: The Practical Implications of Inter-Rater Reliability in Faculty Development

JENNIFER GOOD

I AM A HYBRID. With my terminal degree in educational psychology, which included a literacy cognate piecemealed from writing and pedagogy courses in both English and education, I live with one foot in the world of measurement and assessment and the other in the world of writing theory. Like Brian Huot (2009), whose "colleagues were concerned about hiring him because he had all this assessment stuff on his vita" (Huot & Dillon, p. 207), I have often felt that any efforts to provide quantifiable measures of writing assessment have been looked at with suspicion among my rhetoric and composition colleagues. This has created a tenuous balancing act for me. While I preach the power of the written word and understand its complexities and nuances, I also believe in the strength of numbers to document growth, change, and program value. As a result, my goal is to demonstrate how statistical analysis as part of an ongoing writing assessment can have practical benefits in a writing across the curriculum (WAC) program, specifically that of improving WAC faculty development offerings.

Because of my divided existence, I seek the marriage of writing assessment through authentic samples of student writing with quantifiable and psychometrically sound measurement methods. While I believe in the importance of writing measures that rely upon human raters who read and rate authentic writing samples in response to locally developed assignments (Conference on College Composition and Communication, 2009), I also believe it is essential to demonstrate the reliability of those measures, which means that any writing assignment that is rated, regardless of who completes the evaluation, will yield consistent scores. Thus, I present a model in use at my home university that demonstrates how an analysis of inter-rater reliability can be a bridge connecting the technical expectations of sound psychometric properties in measurement with assessment of authentic writing samples. Through this

DOI: 10.37514/WAC-J.2012.23.1.02

The WAC Journal 23 (2012). © 2013 by Clemson University.
Copies may be circulated for educational purposes only.

model, I also demonstrate how the results can be used to suggest practical ongoing reform for WAC faculty development at both the group and individual levels.

Colleagues of mine in the composition program have told me that they "don't do numbers." Trusting their own expertise in writing pedagogy and theory, they feel confident that the feedback they provide on student papers is sound, fair, and equitable. O'Neill and Moore (2009) explain that most college professors recognize the social, contextualized nature of writing and, in an effort to protect that understanding, resist any measure that minimizes writing achievement to mere numbers: "Misfires can happen in writing assessment when tests are not sensitive to the particular students and their contexts" (p. 37). Like O'Neill and Moore, I value the importance of individual feedback and response to student writing, yet this realization does not make the development of common writing assessments with proven psychometric properties an easy task to accomplish. O'Neill and Moore continue their argument by stating the following: "Because of its complexity, writing cannot be researched—or measured—in the same way that physical traits such as height or weight might be measured" (p. 40).

In his historical overview of writing assessment, Condon (2010) chronicles the trends and changes that have occurred over the past decades, from debates and arguments about the actual measurement tool to issues of reliability and validity. He criticizes initiatives sparked by writing theorists that describe "concepts such as validity and reliability as hegemonic forces of the commercial enterprises that inevitably undermine attempts to establish better assessments" (p. 176). In response, Condon promotes the importance of statistics and sound measurement practices in this ongoing debate, while also emphasizing the need for authentic and contextualized writing production through meaningful prompts and assignments. In agreement with Condon, I feel it is important to create and find models to balance sound measurement with writing samples produced in response to discipline-specific assignments and problems. A number of other writing assessment theorists have emerged in this ongoing conversation (Gallagher, 2010; Huot, 1996), also extolling the benefits of considering properties such as reliability and validity when measuring authentic writing samples. Huot and Dillon (2009) provide the following summative argument: "Writing teachers and program administrators should make an effort to become more familiar with the terminology and beliefs of educational measurement" (p. 216).

Acknowledging that "good writing is a complex process that varies by discipline" (Brockman, Taylor, Crawford, & Kreth, 2010), it can be assumed that good assessment of writing is equally as complex, particularly for faculty from different disciplines, with different levels of expertise regarding writing pedagogy and writing assessment. In response to the need to measure authentic writing samples with a

The WAC Journal 23 (2012). © 2013 by Clemson University.
Copies may be circulated for educational purposes only.

degree of consistency, rubrics are often offered as a solution, as they capture writing outcomes at various levels of performance (Spandel, 2006). Yet, even though a rubric may measure writing skills at various stages of competencies, an abundance of types of writing rubrics and scoring guides exist, and selecting or understanding the use of a particular rubric as it aligns appropriately with an assignment becomes necessary for effective writing instruction (Moskal, 2000).

Even when using rubrics or other scoring guides, Stern (2009) notes that comments regarding assessment of writing samples varied. Leckie and Baird (2011) argue that many factors, including rater severity, rater expertise, and central tendency of a rater can create bias and variation in writing assessment. They noted that rater severity in particular was unstable and changed over time. According to writing experts, administrators must "consider whether we are studying what we think we are studying and whether the measures we use are consistent" (Writing @CSU Guide). In response to the importance of reliability, the training and follow-up statistical model that I have incorporated as part of our regular WAC program operations allows for open discussion of the effect raters can have on evaluation, recognizing that, ethically, the consistency of writing evaluation must be determined when used in high-stakes decision-making or in overall writing program evaluation.

Because a primary aim of many university-wide WAC programs is to ensure that students improve in writing outcomes, consistent assessment of writing and understanding for faculty in all disciplines of the characteristics or dimensions of effective writing becomes imperative. For this reason, inter-rater reliability, or the degree of agreement among two or more raters based upon the relationship in their scores, is often considered. This is not to imply that the students should write in response to common university-wide assignments using the same organizational and stylistic expectations in every class regardless of discipline. Rather, the raters of the students' writing should have some common understanding of the complex components or dimensions of writing that they evaluate, which can then be considered at a discipline-specific level. An analysis of intra-class correlation, which means how closely individual ratings resemble each other within a group, can be used to determine overall inter-rater reliability when using a rubric or scoring guide. What makes this model unique, however, is the follow-up emphasis on individual interclass ratings, or the single measure of one set of ratings to another individual set of ratings within a group. In the case of this model, I use my ratings—as both facilitator of faculty development and program administrator who is responsible for bringing a comprehensive unification to the program goals—as the point of reference or expertise for assessing individual faculty member rubric ratings. Because I compare faculty ratings to my own ratings, two primary benefits emerge when adopting this statistical approach to inter-rater reliability: (a) the first consideration of data can improve group training

The WAC Journal 23 (2012). © 2013 by Clemson University.
Copies may be circulated for educational purposes only.

materials and emphasis of training that is designed and facilitated by me; and (b) the second consideration of data can suggest follow-up training at the individual faculty level to continue to allow me to provide ongoing support to help faculty members understand the university's WAC program and its learning outcomes. This article provides a faculty professional development model that integrates a statistical analysis to help inform further reform in faculty training. The model presented can be easily adapted for use at other institutions.

The Training Model: Connecting Numbers and Words

Before a statistical analysis of ratings can occur, ratings must be collected, and it is within the WAC program faculty development sessions that I have faculty generate these. Our WAC program, and the faculty training integral to the program's success, were developed in response to our accrediting agency's requirement of a quality enhancement plan. For this reason, the program had the support of both administration and faculty in its early stages of design. The content-area faculty at my institution engage in 30 hours of extensive professional development to prepare them for and support them in writing-intensive instruction; these professional development hours are spread throughout 10 sessions over two academic semesters. According to our WAC program procedures, training is required and provided each semester to faculty members who are interested in teaching writing-intensive, content-area courses from different disciplines. Because the program is highly incentivized by the university administration, including $100 per training session that is transferred into appropriate departmental accounts for individual use and the potential of a course release for research after teaching three writing-intensive courses, it is a popular program. Approximately one third of our tenured or tenure-track faculty members, representing all departments and academic schools, have participated within the past three years.

The first four sessions of faculty training, offered prior to writing-intensive instruction, focus on overall program goals and procedures, including the writing assessment requirements that measure student learning outcomes and inform the program's development. It is within the third session, after the WAC program's objectives have been discussed in this faculty development model, that writing assessment methods are shared in detail. An informal inter-rater reliability exercise is included within the session. Although no statistical analysis of the ratings provided by faculty occurs at this point, this exercise allows for an introductory discussion among faculty of the five unique writing dimensions that are collected and measured within the WAC program's assessment system to measure student outcomes of writing growth.

As defined by common university-wide objectives and fleshed out in a university-wide rubric, these five dimensions are *Focus*, *Content*, *Organization*, *Style*, and

The WAC Journal 23 (2012). © 2013 by Clemson University.
Copies may be circulated for educational purposes only.

Language Conventions. Training first introduces content-area faculty to these five dimensions of writing and, after establishment of common understandings of terms and expectations per dimension, discussions of the unique indicators that help to define each dimension at the discipline level also ensue. Although I will be providing sample ratings and correlations on each of these dimensions, it is important to note that *the method* of assessing and analyzing the rubric, more so than the actual rubric, is the centerpiece of this model.

When completing the first inter-rater reliability activity in the third session of training, faculty members must grapple with the meaning of good writing as it aligns with the university's writing program objectives as well as their specific discipline expectations. In order to do that, they read and evaluate three different student-generated written products, rating each of the three products on each dimension of the rubric. The three written products are all in response to a single persuasive prompt and have been intentionally selected to present weak, acceptable, and exceptional writing on a number of the writing dimensions.

After completing the evaluation, the ratings per faculty member are revealed and shared, allowing individuals to defend, question, and discuss the ratings they provided; the faculty members are asked to use specific indicators within each writing dimension and language from the rubric template to support their points during the discussion. This initial interaction with the university's rubric opens conversations that help develop faculty members' understandings of the complexities of writing and some of the difficulties of writing assessment, such as rater bias, or raters' judgments and perceptions that cloud their evaluation of writing, and severity, which means the tendency of some raters to be harsh or lenient relative to other raters.

The final six sessions of professional development occur simultaneously with the faculty member's first semester of writing-intensive instruction, with one session each dedicated to allowing faculty to discuss and understand a different writing dimension within the rubric. During each of these sessions, faculty members closely study the indicators that define each of the writing dimensions being assessed for the university-wide writing program. Views and definitions of each dimension are demanded and challenged through prompts and discussion boards, while instructional strategies that can be used in the classroom to help students improve in a particular dimension are offered, generating thoughts on instructional strategies and feedback to help the faculty members move from the rubric ratings to improvement of writing at the individual student level. In these sessions, faculty members are asked to refine the indicators that define each rubric dimension to help them tease out their own discipline-specific expectations.

In the sixth session of the second part of training, faculty members are asked to participate in a second inter-rater reliability exercise. Unlike the first inter-rater

The WAC Journal 23 (2012). © 2013 by Clemson University.
Copies may be circulated for educational purposes only.

reliability experience, this final rating activity occurs at the end of a full semester of writing-intensive instruction and at the completion of intensive in-depth consideration of both the writing dimensions and the indicators that inform and define each of the dimensions. Because the faculty members have practiced using the rubric in response to authentic writing samples generated in their classes, they are able to hold deeper conversations than offered in the initial introductory sessions regarding assessment. The data from this exercise are collected and analyzed each semester.

After the completion of the final faculty training session and the collection of outcome ratings, inter-rater reliability is determined, and an inter-item correlation matrix is created. As the university's WAC program director, I first look at intra-class correlation coefficients, selected over Cronbach's alpha due to the statistical analyses' ability to tease out complexities such as interaction effect which strengthen the consideration of the extent to which the raters agree, per each of the analytic rubric's writing dimensions; these intra-class correlation coefficients per dimension are analyzed to establish overall technical merit of the university's rubric. I then consider interclass correlation coefficients, or the relationship of one rater to another rater using my ratings, to help determine the individual faculty members who may need additional support in understanding and completing assessments and ratings that are consistent with the university's expectations. Because the director of the program facilitates the faculty training in alignment with the WAC program objectives, I decided to use my ratings for comparison against the individual faculty ratings that are generated in the correlation matrix, as I will also be the facilitator of follow-up training.

The Model in Action: Two Semesters of Lessons Learned

For the initial analysis of inter-rater reliability, five of us rated six unique essays on a scale of 1 (*Inadequate*) to 5 (*Excellent*) per each of the five writing dimensions of the rubric. This initial cohort included me, the WAC senior program associate, and three faculty members from departments in three of the five academic schools on campus: Liberal Arts, Sciences, and Education. For this first analysis of data, I asked each of the faculty members to submit two anonymous and brief writing samples collected during their first semester of writing-intensive content-area instruction, one perceived as a sample of strong or effective writing and the other perceived as a sample of weak writing in their disciplines.

During the first data collection for inter-rater reliability, the overall intra-class correlation coefficients per writing dimension appeared sufficient, as noted by .70 alpha levels or higher, and demonstrated the writing rubric's reliability (Focus=.83; Content=.71; Organization=.81; Style=.80; Language Conventions=.70). Practically speaking, however, the overall intra-class correlation coefficient tells us little to

The WAC Journal 23 (2012). © 2013 by Clemson University.
Copies may be circulated for educational purposes only.

nothing helpful in terms of improving an individual faculty member's ability to assess writing dimensions with consistency relative to other faculty members. It is this consistency of raters we seek in order to have common understanding of expectations and outcomes across the entire university. How can these data regarding an individual's rater reliability be teased out of the available data that is generated during this professional development exercise? As a follow-up step, I considered each individual faculty members' ratings per dimension and looked for differences from my ratings in each dimension to discover where rating issues may exist. Table 1 provides the intra-class correlation coefficients, or the similarity of individuals within a group to determine ratings, using my ratings for comparison purposes.

Table 1. Inter-item correlation coefficients between expert rater and faculty training participants

	Focus	**Content**	**Organization**	**Style**	**Language Conventions**
Rater 1	.343	.000	.200	-.108	.791
Rater 2	.691	.698	.586	.412	.421
Rater 3	.343	-.185	.067	.108	.281
Rater 4	.788	.982	.890	.692	.750

The relationship of individual faculty members' rating scores to my rating scores was disappointing in this first iteration of data collection. For instance, using .70 as an acceptable correlation coefficient, only Rater 4 appeared to be consistent in her ratings per dimension with mine, yielding positive and strong correlation coefficients. Both Rater 1 and Rater 3 actually had negative correlation coefficients on at least one of the dimensions in comparison to my ratings, which means that these two raters were actually moving in opposite directions from my ratings (i.e., if my rating increased for a dimension, their ratings would decrease). Faculty comments at the completion of this exercise supported the data showing a lack of understanding of certain dimensions on the rubric, as faculty noted that evaluating writing samples from disciplines outside of their areas of expertise was difficult, particularly in the writing dimensions of *Content* and *Style*. The data from the analysis supported and confirmed their verbalized concerns. The inconsistency of the ratings between me and most of the faculty suggests that I need to revise some of the professional development sessions and emphasize a better understanding of what we are actually measuring and evaluating related to content or style in student writing. These two dimensions were not consistent in rating responses, nor did they match my understanding, as the writing program administrator, of the dimensions. The most important kernel of wisdom that I gleaned from this initial analysis was that the acceptable

The WAC Journal 23 (2012). © 2013 by Clemson University.
Copies may be circulated for educational purposes only.

overall reliability coefficients for each dimension (range of .70 to .83), although sufficient by textbook standards, were not sufficient when trying to understand the individual needs and understanding of faculty members from different disciplines. Some faculty members simply were not in alignment with me or other members of the university when rating student writing, which, if not addressed, could potentially limit the success of the WAC program overall.

These disappointing individual inter-item correlation coefficients prompted the need for change. Specifically, although faculty members again provided two anonymous samples that had not yet been rated, the following areas were addressed for the next cohort of faculty undergoing training: (a) I required sample length of three pages or less to encourage brevity for the inter-rater reliability exercise in the final training session; (b) Because the majority of disagreement and debate during discussion centered around defining ratings of 3 in the rubric continuum, instead of requesting weak and strong writing samples, I asked the faculty members to provide papers they considered to be 3-rated on the majority of the writing dimensions, as well as one that was primarily 5-rated, and the six new writing samples were selected from among these papers; and (c) I requested the actual writing assignment handout of instructions to accompany the writing sample in order to help the faculty engaged in rating papers to understand the focus, purpose, and criteria of the assignment as defined by the instructors.

After making these programmatic changes to the faculty training, the final rating activity and statistical analysis was repeated with 11 faculty members who were completing training in the next semester of writing-intensive instruction; these faculty members represented seven departments or disciplines within the Schools of Business, Liberal Arts, and Sciences. With the change in emphases in both instruction during professional development and its impact on the final inter-rater reliability exercise, the second round of data generated revealed improvement in understanding of and consistency in rating the dimensions of writing. Intra-class correlation coefficients increased for all five dimensions of the writing rubric (Focus=.94; Content=.94; Organization=.95; Style=.93; Language Conventions=.93). Because of the improved coefficients on this second analysis of data, I now continue to use those same writing samples, which represent three different academic disciplines, with all new training cohorts, enabling me to analyze data across an entire year and make longitudinal decisions about program improvement.

In spite of the improvement in intra-class correlation, a consideration of the inter-item correlation coefficients from the faculty member to my ratings revealed a need for additional revision to faculty training. For instance, if using a .70 coefficient as a goal for consistency in ratings from the individual faculty member to my ratings, only one faculty member, Rater 3, achieved acceptable reliability ratings on all five

The WAC Journal 23 (2012). © 2013 by Clemson University.
Copies may be circulated for educational purposes only.

writing dimensions (see Table 2.) Possibly, for future and follow-up training, this faculty member could act as an additional facilitator of professional development or a mentor for other faculty members, particularly within her discipline, as faculty continue to grapple with and learn about the writing program and the program's stated objectives and expectations of student writing.

Table 2. Inter-item correlation coefficients between expert rater and faculty training participants

	Focus	**Content**	**Organization**	**Style**	**Language Conventions**
Rater 1	.657	.463	.354	.768	.722
Rater 2	.853	.271	.802	.674	.459
Rater 3	.945	.793	.913	.768	.702
Rater 4	.316	.492	.267	.598	.702
Rater 5	.853	.800	.956	.676	.631
Rater 6	.433	.836	.433	.559	.411
Rater 7	.866	.800	.500	.632	.791
Rater 8	.739	.922	.491	.860	.770
Rater 9	.426	.768	.640	.586	.884
Rater 10	.632	.812	.673	.950	.702
Rater 11	.562	.897	.682	.632	.554

Diagnostically, these data can be used to inform more changes to the professional development program. For instance, a global overview per writing dimension reveals that the *Content* and *Language Convention* dimensions yielded the most consistency in ratings per individual participants to my ratings, suggesting that faculty have a common understanding of these dimensions, which allows them to rate writing samples in these areas reliably. In contrast, when looking at the correlation coefficients for *Organization* and *Style*, only Raters 3 and 4 of the 11 faculty members rated in close relationship to my ratings (at .70 coefficient or above), intimating that these two writing dimensions need further attention in group training to support faculty members' understanding of the elements or indicators that define these particular dimensions.

What is most beneficial is that the data helps all faculty members understand that additional faculty development can be helpful as they continue to navigate writing-intensive instruction in their disciplines. For instance, Rater 4 yields consistently low inter-item correlation coefficients when compared to my ratings, from a range of .267 to .702. This finding intimates that I need to give additional attention and individualized instruction to Rater 4. For instance, informal sessions that allow co-rating

The WAC Journal 23 (2012). © 2013 by Clemson University.
Copies may be circulated for educational purposes only.

of other papers could open conversations and discussion about interpretations of the dimensions in the rubric, possibly confronting misunderstandings or deepening knowledge of writing elements. In other cases, the matrix of data suggests that I may want to discuss only one or two dimensions of the rubric with the faculty member. For instance, Rater 5 has strong inter-item correlation coefficients on three of the writing dimensions; however, *Style* and *Language Convention*s tend to be more problematic. Similarly, Rater 8 has strong inter-item correlation coefficients when compared to my ratings on everything except *Organization*.

This matrix of correlation coefficients acts as a map of individual needs for the faculty members at the conclusion of the program. The intent of this analysis is not to point out individual weaknesses of faculty in understanding writing elements or in their ability to rate reliably. Rather, the point of the analysis is to seek areas to focus support in faculty development. The data inform me about overall faculty program success as well as individual faculty who may need more attention or assistance.

To date, the faculty members have been responsive to informal follow-up requests to meet and discuss the rubric, assessment methods, and assignments as a result of this analysis. When first introducing the activity during training, some faculty members have expressed concern over publicly sharing and discussing their ratings, fearing they would be embarrassed if they didn't align with other faculty members. They also indicated that they didn't want ratings that may reflect their performance in the WAC program, particularly if ratings were not in alignment with other faculty, to get shared with departmental administrators who complete faculty performance evaluations. I had to ensure that the individual ratings would remain inside the WAC program and would not be shared with administrators for any personnel evaluative purpose. Because I have remained true to this promise, faculty members have learned to trust the professional development benefits of the inter-rater reliability experience. Due mostly to open communication in the previous nine sessions within the cohort, they accept that the objective of the inter-rater reliability activity, in alignment with the overall faculty development experience, is to enhance their ability to teach and assess writing in their discipline-specific courses, not to judge or rate them as teachers.

At the end of the full training experience, faculty members are asked to provide anonymous written responses and feedback regarding training. The prompts ask that they consider training program strengths, weaknesses, and topics for future training sessions. Other than comments requesting more emphasis on how to teach online writing-intensive courses, the feedback regarding the training experience was positive. Sample comments from these final evaluations parallel my perception of the content-area faculty members' willingness to continue to grow and learn in writing pedagogy and assessment:

The WAC Journal 23 (2012). © 2013 by Clemson University.
Copies may be circulated for educational purposes only.

- I would like to see more on grading strategies.
- More opportunities to assess writing in training before we actually implement our writing-intensive course.
- I'd like to see us have more work and discussion on inter-rater reliability.
- The strengths include focusing on peer review, evaluating using the assessment rubric, sharing classroom problems and experiences, and focusing on kinds of writing assignments that we can give.
- A strength is bonding and learning with other faculty across disciplines. I loved it and learned so much about how to collaboratively learn with professionals in multiple disciplines.
- Working with other faculty to compare and contrast was a strength.

Having the opportunity to share their ratings on student-produced writing assignments from colleagues' discipline-specific classes and discuss and defend the ratings they provided in training created an opportunity for interdisciplinary growth and teamwork.

Conclusions

By looking at the inter-item correlation coefficients per individual faculty member and comparing them to my own, I am better able to ascertain which faculty members continue to struggle with specific writing dimensions. In essence, the matrix becomes my blueprint that enables me to communicate individually with a faculty member and provide additional support or practice in evaluating writing samples on specific writing dimensions. Essentially, the benefits of individual conferencing with students regarding their papers can be modeled with faculty when I individually conference with them regarding their writing assessment.

Certainly, checking the psychometric properties of a writing rubric, such as validity and reliability, is valuable for any program that uses a rubric to measure growth in student outcomes. Yet, on a practical level, many administrators and trainers stop at the first broad sweep of the data when they feel it has produced sufficient results. Few bother to look at the data at a deeper level and consider the story it tells about the various faculty training components of the writing program and what it reveals about the need for additional follow-up in faculty training, both at the group level and the individual faculty member level.

The inclusion of a regular check of reliability for the writing rubric used at my institution developed into a more detailed overview of reliability at the individual faculty level, which then created an abundance of ideas for reforming faculty training and a mechanism for identifying individual faculty members who may need more support and assistance. From research and regular analysis of reliability of writing rubrics, administrators can prompt action and revision in the practical world

The WAC Journal 23 (2012). © 2013 by Clemson University.
Copies may be circulated for educational purposes only.

of faculty training. The practical implications discovered through rigorous analysis of inter-rater reliability can improve both faculty understanding of good writing and faculty development offerings.

REFERENCES

Brockman, E., Taylor, M., Crawford, M. K., & Kreth, M. (2010). Helping students cross the threshold: Implications from a university writing assessment. *English Journal, 99*(3), 42-50.

Condon, W. (2010). Reinventing writing assessment: How the conversation is shifting. *WPA: Writing Program Administration, 34*(2), 162-82.

Conference on College Composition and Communication. (2009, March). *Writing assessment: A position statement* (Rev. ed.). Retrieved from http://www.ncte.org/cccc/resources/positions/writingassessment

Gallagher, C. W. (2010). Assess locally, validate globally: Heuristics for validating local writing assessments. *WPA: Writing Program Administration, 34*(1), 10-32.

Huot, B. (1996). Toward a new theory of writing assessment. *College Composition and Communication, 47*, 549-66.

Huot, B., & Dillon, E. (2009). WAC and writing program assessment take another step: A response to Assessment of Writing. In M. C. Paretti & K. M. Powell (Eds.). *Assessment of Writing* (pp. 207-18). Tallahassee, FL: Association for Institutional Research.

Leckie, G., & Baird, J. (2011). Rater effects on essay scoring: A multilevel analysis of severity drift, central tendency, and rater experience. *Journal of Educational Measurement, 48*, 399-418.

Moskal, B. M. (2000). Scoring rubrics: What, when, and how? *Practical Assessment, Research and Evaluation, 7*(3). Retrieved July 15, 2011 from http://goo.gl/ep8pe

O'Neill, P., & Moore, C. (2009). What college writing teachers value and why it matters. In M.C. Paretti & K. M. Powell (Eds.), *Assessment of Writing* (pp. 35-47). Tallahassee, FL: Association for Institutional Research.

Spandel, V. A. (2006). In defense of rubrics, *English Journal*, *96*(1) 19-22.

Stern, L. A. (2006). Effective faculty feedback: The road less traveled. *Assessing Writing*, *11*, 22-41.

Writing@CSU. (2012). Reliability and validity. Colorado State University. Retrieved 19 June 2012 from http://writing.colostate.edu/guides/research/relval/pop2a.cfm