Paper 908

Using The SAS® System To Examine The Effect Of Augmentation On The Inter-Rater Reliability Of Holistic Ratings

Jim Penny, Center for Creative Leadership, Greensboro, NC
Robert L. Johnson, University of South Carolina, Columbia, SC
Belita Gordon, University of Georgia, Athens, GA

ABSTRACT

A two-stage process by which an holistic rubric is applied to the assessment of open-ended items, such as writing samples, is defined. The first stage involves scoring a performance by the assignment of an integer rating that is congruent with the proficiency level that is exhibited in the performance. The second stage is the subsequent assignment by the rater of an augmentation that indicates whether or not the writing competency reflected in the paper is a bit higher or lower than the competency level reflected in the benchmark paper for the given proficiency level. If the rater feels that the paper represents benchmark proficiency for the given level, no augmentation is assigned to the rating. The results of this study indicate that the use of rating augmentation can improve the inter-rater reliability of holistic assessments, as indicated by generalizability phi coefficients, correlation coefficients, and percent agreement indices. Implications and suggestions for follow-up research are discussed.

INTRODUCTION

During the 1996-1997 school year, 37 of 47 states used some form of performance-assessment in their testing programs (Johnson, Penny, & Johnson, 1998). These open-ended items usually are assessed using a holistic or analytic rubric with 4- to 6-point scales, providing a metric and a framework for the scoring process that is designed to minimize measurement errors due to rater judgment when scoring of essays and similar open-ended assessments (Cooper, 1984; Johnson et al., 1998).

MEASURES OF INTER-RATER RELIABILITY

There are many indices of inter-rater reliability. Among those that are examined in this research are (1) percent exact agreement, (2) percent adjacent agreement, (3) Pearson product-moment correlation coefficient, and (4) the phi index of dependability from Generalizabiltiy Theory. Percent exact and adjacent agreement indices are computed just as one would expect from the names. That is, percent exact agreement is the percentage of times that two raters agree exactly on the score given to a performance, and percent adjacent agreement is the percentage of times that two raters agree to within one unit on the score given to a performance. For example using a 4-point integer scale, if one rater assigns a score of 2 and the second rater assigns a score of 3, then the ratings are not in exact agreement, but they are in adjacent agreement.

The Pearson product-moment correlation coefficient is the index of which many commonly think when they hear "correlation." Because this index is formulated with the assumption that the data are normally distributed in the population, other indices of correlation have been derived for use in situations where distributions may be non-normal. One example of an alternative correlation coefficient is Spearman's rho which is used to co-relate two sets of rankings.

The phi coefficient from Generalizability Theory (Brennan and Kane, 1977; Shavelson and Webb, 1991) is a special case of the intra-class correlation coefficient that can be used as an index of inter-rater dependability (Cherry & Meyer, 1993). This index is computed from ratios of apportioned variances that are estimated using specialized software such as GENOVA (Crick and Brennan, 1983).

BOOSTING INTER-RATER RELIABILITY

THE LENGTH OF THE SCALE BEING USED

In studies of the assessment of performance tasks, high levels of inter-rater reliability have been achieved with scales defined by up to 15 points. For example, when using a 10-point scale, inter-rater reliability estimates using correlation coefficients for Advanced Placement (AP) examinations ranged from .79 to .85 in psychology and from .87 to .96 in computer science (Longford, 1994). Gosling (1966) reported high levels of inter-rater reliability for a 15-point scale in which raters assigned grades from A+ to E-. In this study, ten raters who were assigned to randomly formed teams scored three essays. The correlations between the scores awarded by a chief examiner and the averaged scores of the raters ranged from .96 to .98.

In a simulation study, Coffman (1971) demonstrated that higher inter-rater reliability is possible when scores are based on a 15-point scale instead of than a 5-point scale. He noted, however, that a scale that is based on such refined intervals may not accrue higher inter-rater reliability in practice. He suggested that it was possible, as seen in Longford (1994) and Gosling (1996), to achieve sufficiently inter-rater reliable using a scale of between 7 and 15 intervals. However, Coffman also stated that the greater number of intervals required "clear notions about the characteristics of answers falling at each point" (p. 34).

The task to succinctly and uniquely define the characteristics of benchmark performance for each level of the scale with many points, however, can present some difficulty when implemented. Not only can it be problematic to clearly define what is representative of proficiency at so many levels, but it also can be difficult for the raters to discern, with adequate inter-rater reliability, among eight or more proficiency levels. Hence, it seems possible, and, moreover, it seems likely, that the length of a scale may affect measurement error, serving to increase the error component of variance when the scale length surpasses the ability of raters to discriminate between levels of proficiency.

THE PERFORMANCE BEING ASSESSED

The ability to make ever-increasing refinements in judgments of quality and proficiency appears to be related to the specificity of responses expected in the product (Elbow & Yancey, 1994). In contrast to the high level of inter-rater reliability achieved in the assessment of AP examinations in psychology and computer science, Longford (1994) found that the scoring of AP composition exams was associated with lower inter-rater reliability, with correlations between raters that ranging from a low of .50 to a high of .73. Coffman (1971) indicated that inter-rater reliability tends to decrease as the subject area becomes more fluid, with the assessment of mathematics problems resulting in the highest inter-rater reliability, followed by incremental reductions of inter-rater reliability in chemistry, history, and general composition. It seems reasonable to suppose that the complexity of performance tasks in mathematics has substantially increased since 1971, resulting in the potential for reduced inter-rater reliability of these assessments. Yet, the point remains that inter-rater reliability is likely to decrease as the complexity of the performance increases.

EXTENDING THE LENGTH OF A SCALE

As previously noted, the level of specificity that can often be achieved in the sciences and in mathematics is not always possible in subject areas where diversity in responses is more acceptable, and perhaps even valued (Elbow & Yancey, 1994). To improve inter-rater reliability in the scoring of

performance assessments where a variety of correct responses is expected, Cronbach, Linn, Brennan, and Haertel (1995) proposed allowing raters to augment integer-level scores by the use of an additional decimal. Thus, a rater can assign a 2.4 rather than a 2 (1995, p. 7) with the hope of the researchers being that if the rater feels that a response is a bit superior, or inferior, to the benchmark response for a given level, but not sufficiently different to warrant a different integer-level rating, then the rater can augment the integer with the fraction. Moreover, it is also suggested that permitting such flexibility in the scores that are given by raters is likely to improve inter-rater reliability.

DEFINITION OF RATING AUGMENTATION

The augmentation of scores by raters can be described as a two-stage process. In the first stage, the rater assigns the integer-level rating that best describes the level of proficiency that is represented by the paper. This integer-level rating is reached by comparing the paper that is being evaluated with the benchmark papers that are considered representative for each level of proficiency in the scoring rubric. One might expect, however, that there is a dispersion of true proficiencies that underlies each of these integer-level ratings. Moreover, it is likely that a paper could contain a few elements of an adjacent level of proficiency, but not in contain them sufficient abundance as to warrant the adjacent rating. Hence, it seems reasonable to expect that some papers at a given integer-level of proficiency, may actually strike the rater as reflecting a slightly higher, or perhaps lower, level of competency than the benchmark paper.

If we allow a rater a second stage in which to indicate that a paper appears markedly different from the benchmark, though not sufficiently different as to warrant a different integer-level rating, then we should have an indicator of "lean" within that distribution. Does the paper "lean" to the higher, or lower, end of the distribution, away from the center that is defined by the benchmark paper? If we allow the rater to augment an integer-level rating, say the rating is a "2," using a "+" if the paper appears superior to the benchmark paper and a "-" if the paper appears inferior to the benchmark paper, then it seems reasonable to think such an augmented measure would provide additional information about the level of proficiency that is represented by the paper.

RESEARCH QUESTIONS

Five research questions guided the investigation between rating augmentation and inter-rater reliability. The first question involved the propensity of raters to use augmentation. If raters augment scores in a limited fashion, then it is unlikely that either the scale expansion will occur or that the potential benefits to inter-rater reliability will occur.

The second research question involved the effect of the underlying distribution of proficiency on the selection of positive and negative augmentation. Based on the assumption that the underlying distribution of writing proficiency is centrally mounded and symmetrical (Diederich, 1974; Myers, 1980; Smith, 1993), one might expect that the use of augmentation would more likely move a rating toward the center of the scale.

The third research question involved the effect of augmentation on the observed distribution of ratings. Our expectation was that the use of augmentation was unlikely to change the mean rating given by the raters, but that it was likely to reduce the variance of the ratings by a small amount.

The fourth question involved the expected changes in the indices of exact and adjacent agreement. Our expectation was that, given the increase in the number of possible scores, the percent of exact and adjacent agreement between raters was likely to be substantially reduced. However, such a comparison of percent agreement is not a fair contrast of the two indices because the adjacent augmented ratings are separated by less distance than are the integer ratings. It was our expectation that the proportion of adjacent ratings would be comparable to the proportion of augmented ratings that differ by one unit or less.

The fifth research question involved the effect of the use of augmentation on the measures of association, such as correlation coefficients and indices from generalizability theory, between the scores given by raters (Cherry & Meyer, 1993). As mentioned earlier, Cronbach et al (1998) predicted that the use of augmentation would improve such measures. Implicit in the anticipation of improved agreement between raters as they assess level of exhibited proficiency is the idea that augmented ratings could improve the accuracy of holistic ratings, addressing the fundamental concerns involving the validity of holistic scoring (Huot, 1993; Pula & Huot, 1993).

METHODOLOGY AND DATA SOURCES

The data for this investigation are 120 essays drawn from the Georgia Writing Assessment for 5th-grade students. Although this assessment is not considered a high stakes assessment by the state of Georgia, one school system did use the results as part of the promotion criteria for a student to advance to the sixth grade. Two raters scored each essay. The papers that received ratings that were not in exact agreement at the integer level were re-scored by two adjudicators who had more expertise in scoring essays than the raters.

The papers that were selected for this study were not chosen completely at random; rather, they were selected by virtue of having been found "difficult" in prior assessments. By difficult, it is meant that prior raters had difficulty coming to agreement over the degree of proficiency that was exhibited in the papers.

Raters independently scored the writing samples essays with an holistic rubric that is designed to represent six developmental stages. These stages are (1) The emerging writer, (2) The developing writer, (3) The focusing writer, (4) The experimenting writer, (5) The engaging writer, and (6) The extending writer (Georgia Department of Education, 1993). After recording the integer-level score, the raters were also asked to indicate whether the integer score should be augmented with a plus (+) or a minus (-). Raters were instructed to assign a "+" if the paper appeared to be higher than the benchmark paper at that proficiency level and to negatively augment the paper if it appeared to be lower than the benchmark paper at that level.

The adjudicators were aware that score discrepancies occurred in the scoring of the essays; however, they were not aware of the actual scores until their ratings were completed. Working independently, adjudicators recorded integer-level ratings and indicated whether their integer-level scores should be augmented.

RESULTS

Were the raters and experts inclined to use augmentation? As shown in Tables 1a and 1b, both raters and experts chose to augment many of their ratings. Of the 240 ratings given by the raters, 53 percent were integers, 25 percent were given negative augmentations, and 23 percent received positive augmentations. Of the 114 augmentations given by raters, 52 percent were negative and 48 percent positive. The story for the distribution of augmentations by experts is a bit different. Of the 98 ratings that were given by the experts, 55 percent were integers, 19 percent received a negative augmentation, and 25 percent received a positive augmentation. Of these 44 augmentations, 43 percent were negative, and 57 percent were positive. Although the split of augmentations by experts is just the opposite of the split seen in the data from the raters, the difference in the proportions of negative and positive augmentations is not statistically significant.

Is there evidence to suggest that the underlying distribution of proficiency affects the use of augmentation? Although there is evidence to suggest that the numbers of positive and negative augmentations are well balanced, there is additional evidence in Tables 1a and 1b to suggest that the choice of augmentation is dependent on the integer rating. First, note that in this 6-point rubric, the scores of "1," "2," and "3" can be considered to be below the center of the scale whereas the scores of "4," "5," and "6" are above the center of the scale. In every instance of integer scores that were below the center, there were more positive augmentations than negative augmentations. For those ratings that were above the center of the scale, just the opposite proved true; that is, for the ratings that are above the center of the scale, there are more negative augmentations than

there are positive. It is apparent that these raters and experts tended to choose augmentations that were toward the center of the scale.

In order to compute the mean and standard deviation of the augmented ratings, we needed to assign a numerical value to the augmentations. We chose to add one-third of a point to the integer when the augmentation was positive, and to subtract one-third when the augmentation was negative. We chose the additive constant of one-third to represent augmentation because that gave us numerically equal distances between the steps in the augmented rating scale. For instance, the distance between from 3+ to 4- is one-third, just as is the distance between 1 and 1+.

The shifts that were seen the means were small and never statistically significant. As well, the changes in the standard deviations were small, though achieving statistical significance only for the complete set of 120 ratings by raters ($F=1.27$, $df=(240,240)$, $p=.0343$). These results, shown in Tables 2a and 2b, are not surprising in view of the fact that both raters and experts tended to choose augmentations that were toward the center of the scale.

To examine the degree of congruency between raters and experts, the percentage of agreements and the percentage of disagreements were computed for both the integer-level scores and the augmented scores. Though rare, there are a few assessment programs that require exact agreement between raters before a decision is made about the paper. While this strict requirement is easily defended in both public and professional forums, it does tend to increase the need for follow-up review by experts who must resolve the often occurring discrepancies between the raters.

As shown in Table 3a the percent of exact agreement was 59 percent for raters in this study. Hence, 41 percent of the papers, for a total of 49, were sent to experts for follow-up review. Had adjacent disagreement been designated as a sufficient level of agreement, only a single paper would have been submitted to the experts, increasing the effective agreement rate from 59 percent to 99 percent. From these data, it is easy to see the motivation of an assessment organization to consider the inclusion of adjacencies as sufficient agreement.

Because part of the present investigation focused on the measurement issues that are embedded in the various methods of resolving rater disagreement, we chose to send all the papers over which the raters disagreed to the experts for additional review. Doing so allowed us the opportunity to examine the percentage of agreement between experts. As shown in Table 3b, experts in this study did tend to agree more often than the raters even though the experts were faced with a more difficult subset of this sample of intrinsically difficult papers. The experts agreed exactly on 63 percent of the papers. This percentage is greater than that found with the raters, but the difference was not statistically significant ($Z=.494$, $p=.3106$) . If the requirement for sufficient agreement includes adjacent scores, then the agreement level of the experts rises to 96 percent, a similar, though slightly lower, agreement level to that achieved by the raters and representing only 2 of the 49 papers submitted for expert review.

One might expect that the use of augmented ratings, because of the increased number of ratings that a paper could receive, would have the effect of reducing the amount of exact agreement that is exhibited between raters, and this is exactly what was found. As shown in Table 4a the level of exact agreement between raters fell to only 25 percent.

If one extrapolates from the existing assessment procedures that permit adjacent agreement in ratings, then one could reason that a difference of 1 or less in the ratings also constitutes sufficient agreement and precludes the need for follow-up review by the expert raters. If such a decision ruled were applied to the augmented data that are summarized in Table 4a, then 96 percent of the papers would have received ratings that were in sufficient agreement. This change in agreement by the raters from 59 percent to 96 percent is statistically significant ($Z=6.6$, $p=.0000$), and, more importantly, practically significant. Indeed, had this decision rule been applied, only 6, instead of 49, of the papers would have been passed to the experts for follow-up review.

If the decision rule were to require a difference in rating of less than 1, but not inclusive of 1, agreement is 85 percent, an improvement from the original 59 percent that is still statistically significant ($Z=4.3$, $p=.0000$), and also practically significant in that only 19, not 49, of the papers would have been submitted for adjudication.

The effect that using the augmented ratings had on the agreement level of the experts is presented in Table 4b, and is similar to that which was seen with the raters.

To further investigate the effect of augmentation on agreement between the raters, the correlations of the ratings given by the pairs of judges were computed. Although correlation coefficients are not necessarily good estimates of inter-rater reliability when the mean scores differ between across raters, correlation coefficients can provide a reasonable index of the consistency in rank order. Moreover, the similarity of mean scores for the raters and the experts in this particular study indicates that, in this instance, the correlation coefficient could provide a satisfactory index of inter-rater reliability between the pairs of judges.

In every case, as shown in Tables 2a and 2b, the correlation coefficient from the integer ratings to the augmented ratings, for both raters and expert judges increased, with the change in correlation between the two types of ratings for the discrepant ratings by raters and for the follow-up ratings by the experts both achieving statistical significance, ($F=1.92$, $df=(98,98)$, $p=.0030$) and ($F=1.42$, $df=(98,98)$, $p=.0407$), respectively. If one considers only those papers on which the raters disagreed, the correlation using integer ratings is .31 and improves to .43 after augmentation, a change that is a statistically significant ($F=1.92$, $df=(49,49)$, $p=.0119$). The correlation of the ratings by the experts improved from value of .62 with integer ratings to a value of .74 with augmentation, a change that is not statistically significant.

The generalizability phi-coefficient estimates, defined in Brennan and Kane (1977) and discussed in Shavelson and Webb (1991), of the inter-rater reliability of the raters and the experts were computed using the GENOVA program by Crick and Brennan (1984) from the 49 papers on which the original raters disagreed. These coefficients are computed for four different methods of discrepant score resolution and presented in Table 6. These four methods are (1) the average of the two raters, (2) the average of the two experts, (3) the average of the two raters and each expert, and (4) the table score which is produced by averaging the rating given by an expert with the closest rating given by a rater. There are two experts, hence the two table scores. Moreover, Table 5 gives the increment in the inter-rater reliability for each of the components of a given average. For instance, when the discrepancy is resolved by using the average of Rater 1, Rater 2, and Expert 1, three phi-coefficients are given. The first phi coefficient estimates inter-rater reliability for scores based on the judgment of one rater, the second for scores based on two raters, and the third for scores based on three raters.

In every instance, the use of score augmentation produces a higher value of the phi-coefficient, indicative of improved levels of inter-rater reliability in these data. The mean phi-coefficient that is computed using the integer data is 0.60 with a standard deviation of 0.14, while the mean that is computed using the augmented scores is 0.69 with a standard deviation of 0.13. The variances of these two groups are very similar, and, as expected, the difference between the two variances is not statistically significant. The means, however, are another story and differ by 15 percent, a statistically significant difference ($d=0.09$, $s= .027$, $t=11.3$, $p=.0000$, $df=10$) that we also see as practically significant.

DISCUSSION

The results of this study are encouraging. While the implementation of augmentation does require some additional training of raters and experts, it does seem to be easily accomplished, with few questions on the part of the raters; indeed, in the present study, it became apparent very quickly that the raters wanted the ability to record the augmentation in order to

indicate that particular papers were somewhat different from the exemplars, but not sufficiently so as to warrant a different score.

We are intrigued by the phenomenon exhibited by both the raters and the experts in which they appeared more likely to assign an augmentation towards the center of the scale, and we see this as support for the expectation that true proficiency lies on a continuum which underlies the rating scale. Moreover, it seems reasonable to argue that the existence of this phenomenon confirms our expectation that raters will use augmentation to indicate those papers that represent a proficiency that differs slightly from the chosen benchmark, though not enough so as to receive a different integer score.

That the mean and standard deviation of the ratings before and after augmentation did not change substantially came as little surprise, and it appeared to be the result of the thoughtful application of augmentation by the raters and the experts. It would appear, then, that the use of augmentation had little or no effect on the distribution of the ratings, other than to increase the number of scale points that span the underlying proficiency distribution.

We find the improvement in rater agreement that was seen in the augmented scores particularly relevant to scoring practices in testing agencies that use holistic rubrics. It was evident from these data that the use of augmented scoring could substantially reduce the need for follow-up review by expert raters.

Finally, it is the improvement in inter-rater reliability as seen in the higher values of the phi and correlation coefficients that presents the strongest argument that augmentation offers a method for improving inter-rater reliability. The use of augmentation consistently, and occasionally dramatically, improved the inter-rater reliability of the raters, resulting in the frequent occurrence where the phi coefficient for a single rater using augmentation was nearly equal to the phi coefficient for both raters using integers.

CONCLUDING REMARKS

It can be argued that these results are those that one would expect from the simple 3-fold expansion of the scale. We tend to agree with this observation. A measure based on a scale with more points, regardless of the exact unit, is likely to carry more information than a measurement using a similar scale with fewer points. Indeed, if one could define rigorous and distinct benchmarks for each point on the scale and at the same time train raters to consistently recognize exemplars of those benchmarks, then it is likely that augmentation of the shorter scale would present little improvement over the use of the longer scale.

It is exactly this "if," though, that a rating design using scoring augmentation addresses. If the development and implementation of a rubric with a 3-fold increase in the scale with well-developed exemplars at each benchmark is impossible or impractical, then it seems reasonable to expect that augmentation could provide some increase in the inter-rater reliability of the assessment procedure by virtue of the implicit 3-fold increase in scale width that the studies style of augmentation can produce.

The decisions that are based in part on the results of the assessment of student writing samples can have deep and far-reaching effects on the lives of many people, and it is not uncommon to find teachers, students, parents, and administrators who are, to a degree, uncomfortable with the application of scoring rubrics in the assessment of writing samples. Remuneration, employment, placement, promotion, and graduation are all posited, to some extent, on the accuracy of the assessment of writing samples, and it is frequently the limited precision of 4- and 6-point rubrics that people express as a concern. One might argue that the use of augmentation, and the implicit extension of that rating scale that augmentation provides, could, at least partially, address some of those concerns.

REFERENCES

Baxter, G., Shavelson, R., Goldman, S., & Pine, J. (1992). Evaluation of procedure-based scoring for hands-on science assessment. *Journal of Educational Measurement 29* (1), 1-17.

Breland, H. (1983). *The direct assessment of writing skill: A measurement review*. (Technical Report No. 83-6) Princeton, NJ: College Entrance Examination Board.

Brennan, R. L., & Kane, M. T. (1977). An index of dependability for mastery tests. *Journal of Educational Measurement, 14*, 277-289.

Cherry, R. D., & Meyer, P. R. (1993). Reliability issues in holistic assessment. In M. Williamson & B. Huot (Eds) *Validating holistic scoring for writing assessment: Theoretical and empirical foundations*. Cresskill, NJ: Hampton Press.

Coffman, W. (1971). On the reliability of ratings of essay examinations in English. *Research in the Teaching of English, 5*, 24-36.

Cooper, P. (1984). *The assessment of writing ability: A review of research*. (Technical Report No. 84-12) Princeton, NJ: Education Testing Service.

Crick, J. E., & Brennan, R. L. (1984). Manual for GENOVA: A generalized analysis of variance system. Iowa City: The American College Testing Program.

Cronbach, L., Linn, R., Brennan, R., & Haertel, E. (1995). *Generalizability analysis for educational assessments*. Los Angeles: Center for the Study of Evaluation, Standards, and Student Testing, University of California at Los Angeles.

Diedrich, P. (1974). *Measuring growth in English*. Urbana, IL: National Council of Teachers of English.

Elbow, P., & Yancey, K. B. (1994). On the nature of holistic scoring: An inquiry composed on email. *Assessing Writing, 1*, 91-107.

Ercikan, K., Schwarz, R. D., Julian, M. W., Burket, G. R., Weber, M. M., & Link, V. (1998). Calibration and scoring of tests with multiple-choice and constructed-response item types. *Journal of Educational Measurement, 35*, 137-154.

Wolfe, E. W. (1997). The relationship between Essay Reading Style and scoring proficiency in a psychometric scoring system. *Assessing Writing, 4(1)*, 83-106.

Georgia Department of Education (1993). *Georgia High School Writing Test: Assessment and Instructional Guide*. Author.

Godshalk, F., Swineford, F., and Coffman, W. (1966). *The measurement of writing ability.* Princeton: College Entrance Examination Board.

Gosling, G. (1966). *Marking English compositions*. Victoria: Australian Council for Educational Research.

Huot, B. (1990a). Reliability, validity, and holistic scoring: What we know and what we need to know. *College Composition and Communication, 41(2)*, 201-213.

Huot, B. (1990b). The literature of direct writing assessment: Major concerns and prevailing trends. *Review of Educational Research, 60(2)*, 237-263.

Huot, B. (1993). The influence of holistic scoring procedures on reading and rating student essays. In M. Williamson & B. Huot (Eds) *Validating holistic scoring for writing assessment: Theoretical and empirical foundations*. Cresskill, NJ: Hampton Press.

Johnson, R. & Bergman, T. (1996, April). The development and implementation of a family literacy portfolio system. Paper presented at the Fifth Annual Conference on Family Literacy in Louisville, Kentucky.

Longford, N. (1994). *A case for adjusting subjectively rated scores in the Advanced Placement tests*. (Technical Report No. 94-5) Princeton, NJ: Education Testing Service.

Masters, J. (1992). A study of arbitrations in Pennsylvania's writing assessment. Paper presented at the annual meeting of the National Council of Measurement in Education, San Francisco, CA.

McColly, W. & Remstad, R. (1965). Composition rating scales for general merit: An experimental evaluation. *Journal of Educational Research 59*, 55-56.

Meyers, M. (1980). *A procedure for writing assessment and holistic scoring.* Urbana, IL: National Council of Teachers of English.

Pula, J. J., & Huot, B. A. (1993). A model of background influences on holistic raters. In M. Williamson & B.

Huot (Eds) *Validating holistic scoring for writing assessment: Theoretical and empirical foundations*. Cresskill, NJ: Hampton Press.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability Theory*. Newbury Park, CA: Sage.

Smith, W. (1993). Assessing the reliability and adequacy of using holistic scoring of essays as a college composition placement technique. In M. Williamson & B. Huot (Eds) *Validating holistic scoring for writing assessment: Theoretical and empirical foundations*. Cresskill, NJ: Hampton Press.

Wolfe, E. W. (1997). The relationship between essay reading style and scoring proficiency in a psychometric scoring system. *Assessing Writing, 4(1)*, 83-106.

CONTACT INFORMATION

Jim Penny
Center for Creative Leadership
One Leadership Place
Greensboro, NC 27438-6300
Work: 336-286-4442
Fax: 336-286-4434
Email: pennyj@leaders.ccl.org

TABLE 1A

Distribution of Augmented Holistic Scores by Raters

Score	Count	Percent
1-	0	0
1	0	0
1+	0	0
2-	0	0
2	2	1
2+	1	0
3-	7	3
3	30	13
3+	22	9
4-	25	10
4	47	20
4+	24	10
5-	22	9
5	41	17
5+	8	3
6-	5	2
6	6	3
6+	0	0

TABLE 1B

Distribution of Augmented Holistic Scores by Experts

Score	Count	Percent
1-	0	0
1	0	0
1+	0	0
2-	0	0
2	0	0
2+	0	0
3-	0	0
3	13	13
3+	8	8
4-	8	8
4	29	30
4+	10	10
5-	9	9
5	11	11
5+	7	7
6-	2	2
6	1	1
6+	0	0

TABLE 2A

Mean, SD, and Correlation of Rater Scores

	All papers n = 120		Papers on which the raters disagreed n = 49	
	Correlation	Mean (SD)	Correlation	Mean (SD)
Integer Scores	.72	4.1 (0.9)	.31	4.2 (0.9)
Augmented Scores	.75	4.1 (0.8)	.43	4.1 (0.8)

TABLE 2B

Mean, SD, and Correlation of Expert Scores

	Correlation	Mean (SD)
Integer Scores	0.62	4.1 (0.8)
Augmented Scores	0.74	4.2 (0.7)
n	49	

TABLE 3A

Degree of Disagreement between Raters using Integer Scores

Difference in Score	Count	Percent
0	71	59
1	48	40
2	1	1

Note. 41 percent disagree by 1 or more points.

TABLE 3B

Degree of Disagreement between Experts using Integer Scores

Difference In Score	Count	Percent
0	31	63
1	16	33
2	2	4

Note. 37 percent disagree by 1 or more points.

TABLE 4A

Degree of Disagreement between Raters using Augmented Scores

Difference In Score	Count	Percent
0.00	30	25
0.33	44	37
0.66	27	23
1.00	13	11
1.33	5	4
1.66	1	1

TABLE 4B

Degree of Disagreement between Experts using Augmented Scores

Difference in Score	Count	Percent
0.00	17	35
0.33	16	33
0.66	9	18
1.00	4	8
1.33	3	6

Note. 14 percent disagree by 1 or more point.

TABLE 5

Phi Coefficient Estimates of Inter-rater Reliability Attained for Each Score Resolution Method

			Method 1	Method 2	Method 3		Method 4
	Scores	Raters	Rater 1 Rater 2	Expert 1 Expert 2	Rater 1 Rater 2 Expert 1	Rater 1 Rater 2 Expert 2	Table 1 Table 2
Integer	49	1	.31	.60	.47	.52	.67
		2	.47	.75	.64	.69	
		3			.73	.77	
Augmented	49	1	.43	.72	.54	.63	.78
		2	.60	.84	.70	.78	
		3			.78	.84	

Note. Mean improvement in phi introduced by augmented scores is .09 with sd=.027 (t=11.29, df=10, p=.0000)

Method 1: The average of two raters.
Method 2: The average of the two experts.
Method 3: The average of the two experts and each rater.
Method 4: The table score is the average produced using the rating of one expert with the rating of the "closest" rater. There were two experts, hence two tables.