

**An Examination of Variation in Rater Severity Over Time:
A Study in Rater Drift**

Mark Wilson & Harry Case
Berkeley Evaluation and Assessment Research (BEAR) Center
University of California, Berkeley
October, 1997

Address: Mark Wilson, Education, UC Berkeley, Berkeley, CA 94720

Running Head: Rater drift

Keywords: rater consistency, rater severity, item response modeling

Acknowledgments: This research was carried out by the Berkeley Evaluation and Assessment Research (BEAR) Project, with funding from Los Angeles County Office of Education, for the California State Department of Education. A version of this paper was presented at the National Conference of the Council of Chief State School Officer's Association, Phoenix, Arizona, June, 1996 (Wilson & Case, 1996). The opinions expressed, are, of course, solely those of the authors. We would like to thank staff of the California State Department of Education, especially Dale Carlson, Sue Bennett, Gerry Shelton, and Ellen Lee. We would also like to thank the staff of the Sacramento County Office of Education who organized the scoring sessions, especially Linda Murai and the raters at the two sites we visited, and the staff of CTB-McGraw Hill who provided

mark-sense services at the sites. Programming for the project was carried out by Wen-chung Wang.

Abstract

Ratings on performance tasks from two scoring sessions of an eighth grade mathematics examination, developed by the California State Department of Education, were used (a) to study the feasibility of estimating IRT rater severity information within a scoring session, (b) to investigate the variation in rater severity within rating sessions (which we called *rater drift*), and (c) to examine the relationship (or lack thereof) of the IRT rater severity information to traditional information based on expert re-ratings. We found that it was indeed feasible to give feedback on half-day intervals, that that information was interpretable by the raters, that there was considerable variation in rater severity within the scoring session, and that the IRT information was not redundant with the traditional information. We also investigated the impact of the estimated rater severities using an IRT approach, and showed that they accounted for about half of the residual misfit. In addition, we found that table leaders (expert raters assigned to lead small groups of raters) exhibited more variation in rater severity than the regular raters. This observation may have as much to do with the somewhat odd sample of student work that they examine as it does to any systematic pattern of ratings on their part. We investigated the effect of table leader's severity on the raters they were leading, and found that there was no systematic relationship.

An Examination of Variation in Rater Severity Over Time: A Study of Rater Drift

In recent years, there have been a number of attempts to incorporate essay questions and performance tasks into large-scale standardized examinations. For example, the *Scholastic Aptitude Test* (SAT) from the Educational Testing Service now includes an essay component, and *Terra Nova* from CTB-McGraw Hill includes a variety of performance-style assessments. There are several reasons why this change has come about (see Khattri & Sweet (1996) for a recent survey). Some advocates want to influence what is taught in schools; some want to remove the perceived negative influence of multiple choice standardized tests. To the extent that assessment can drive instruction, some see that it is useful to include essay questions and performance tasks on assessments and required examinations. Another motivation is the belief that essay questions and performance tasks can measure cognitive skills that are, at best, indirectly measured (Resnick & Resnick, 1996) and, at worst, unmeasurable by multiple-choice or true and false questions. The intent is that a test yield a more complete assessment of the student's abilities by offering a broader range of item types (Hogan, 1981; Mislevy, 1991).

Essay questions and performance tasks are not without their drawbacks. They can be expensive and time consuming to take and to score, and introduce a certain amount of subjectivity into the scoring process. The subjectivity arises because raters need to make judgments about the student papers. Guidelines for such judgments cannot contain all possible contingencies and therefore the rater must make a judgment. It is this judgment that allows subjectivity to enter the scoring process. Contrariwise, it is this judgment that offers the possibility of a broader and deeper interpretation of test scores.

It is the discretion raters must use in assigning a score that raises the possibility that there are inconsistencies between raters. In particular, parents may be concerned that their child was scored by a rater who was consistently assigning lower scores and, in fact, earlier work (e.g., Lunz, Wright & Linacre, 1990; Wilson & Wang, 1995) has shown that there are instances where raters differ significantly (both statistically and substantively) in their severity. One of the goals

of this paper then is to investigate ways to identify these inconsistencies between raters over the period of time while they are actively involved in the scoring process--so-called *rater drift*-- so that timely corrective measures can be taken.

Background

Failure to assign appropriate scores may be due to a variety of factors related to the rater, such as: fatigue, failure to understand the intentions of those who created the scoring guidelines, distractions due to matters such as poor handwriting on the student's part, and distractions due to a prior student's responses. The types of errors that raters make have traditionally been classified into four different categories (Saal, Downey, and Lahey, 1980). They are (a) severity or leniency, (b) halo, (c) central tendency and (d) restriction of range.

Rater *severity* or *leniency* is a consistent tendency on the part of the rater to give a score that is higher or lower than is appropriate, which is usually interpreted to mean higher or lower than the average of the other raters.

The *halo* effect may manifest itself in three distinct ways, all of which involve the rating of one response affecting the rating of another. The first may occur when a single response is scored on a number of subscales. Here the rater scores the subscales based on some overall impression rather than looking at each subscale as a unique and independent domain. The second type of halo effect is between different responses from the same person. Here the overall impression comes not from a single response but from several responses. Each response is therefore not scored on its own merit. The third manifestation of the halo effect is between persons. The response by a prior person may influence the scoring of a later response by a different person.

Central tendency and *restriction of range* are similar in that in both cases the rater is not making full use of the scoring range. However, in the case of central tendency the rater rarely awards scores at the extremes (i.e., the scores are (almost) always in the middle of the range), while in the case of restriction of range, the narrow band of scores awarded may be in any part of

the range. Thus, central tendency is a special case of restriction of range. The causes of the problems may differ as well. A rater may adopt a rating strategy that looks like central tendency because staying in the middle of the range reduces the possibility of grossly mis-scoring any student (which means that discrepancy models used to check up on raters will not be very sensitive). However, a restriction of range, for example at the low end, may be due to a failure on the part of the rater to see distinctions between the different levels.

There are several actions that can be taken to promote more appropriate ratings. One action is to provide extensive training so that the rater more fully understands the intentions of the test designers and is aware of influences that can bias his or her judgment. This training may include instructions on how to interpret the scoring guidelines and opportunities to score sample items, and may be repeated during the scoring session.

A second action is to have double or triple readings of the student papers. Differences between scores by different raters can then be resolved in a number of ways. For example, they can be arbitrated by discussion and mutual agreement, by taking the average of the scores, by deferring to the more expert of the raters, or referring them to another “expert.” A variant on this strategy would have a more expert rater re-rate a sample from each rater’s work. This strategy, termed “read-behinds”, was used in the context described below. Of course, the lighter the sampling, the less reliable this method will be in detecting discrepant raters.

A third approach uses information from one source of information about the rater to check for consistency with other pieces of information, such as other ratings. In the context to be described below, for example, information about a rater’s ratings of students, along with the students’ responses to multiple choice items, was combined to estimate a rater’s severity. In this case, *all* ratings could be used, not just a sample. A rater would be considered severe if he or she tended to give scores that were lower than would be expected from other sources of information. If the rater tended to give higher scores then that would be considered leniency. Once detected, biased scores due to severity or leniency can be used as a basis for rater advisement, and/or the scores can then be adjusted to correct for the effect of the bias.

It is important to recognize that the use of a statistical model to check consistency has its limitations as a methodology. These models show their weakness with respect to the individual student. This is because, in general, mathematical models are premised on the belief that we expect raters, students and items to act in a consistent fashion. However, raters may randomly assign a few scores that are too severe or too lenient even though, overall, they are doing a good job. Fit indices could be used to detect the most inconsistent patterns. Similarly, students who do well on most of an examination may make a mistake on a relatively easy question. In other words, we expect these models to be useful in identifying uniform severity problems and less useful with non-uniform severity problems. Consequently any adjustment to student scores may therefore improve the overall reliability and hence, improve the consistency of individual scores, but at the same time reduce the correctness of some (fewer) individual scores. Likewise, the use of mathematical models to help identify biased raters may be quite successful, but may not be so successful in helping us find inappropriate individual scores for student responses.

Nonetheless, there have been several studies that have demonstrated that mathematical models can be useful methods for analyzing rater performance. Several approaches have been used to calculate rater severity/leniency: ANOVA (Guilford, 1954), Ordinary Least Squares (OLS) or Weighted Least Squares (WLS) regression (Raymond & Viswesvaran, 1993; Braun, 1988); and IRT (Engelhard, 1994; Lunz, Wright & Linacre, 1990; Wilson & Wang 1995; Wang & Wilson, 1996).

A variety of ANOVA approaches exist, such as the investigation of rater X dimension designs and rater X ratee X dimension designs. All of these approaches attempt to establish the existence of a significant rater main effect. The OLS and WLS regression approaches use the following model.

$$Y_{ijk} = \alpha_{ik} + \beta_{jk} + \epsilon_{ijk} \quad (1)$$

where:

Y_{ijk} is the rating given to candidate i by rater j on item k ,

α_{ik} is the true rating for candidate i on item k ,

β_{jk} is the leniency index for rater j on item k , and

ϵ_{ijk} is the normally distributed random error.

This model assumes that the error terms are normally distributed with an expected value of zero and that the variance of the errors across raters is equal. The OLS procedure provides an unbiased estimate of the vector of true ratings. If, however, the reliability of scoring varies from rater to rater then the usual regression assumption of equal error variances across all raters and ratees will be violated and the use of WLS will be more appropriate.

Another approach to estimating rater severity utilizes item response theory (IRT). A basic IRT model starts with the assumption that there are measurable latent characteristics called student ability and item difficulty and that these characteristics can be expressed in terms comparable to each other. The model then goes on to assume that the greater the student's ability relative to the item difficulty, the more likely the student is to answer the item correctly, or conversely, the lower the student's ability relative to the item difficulty, the more likely the student is to answer the item incorrectly. Note that a probabilistic model is used, i.e., the student always has some probability of getting the answer right or wrong regardless of the ability of the student or the difficulty of the item. In the case of rated or judged items, one would think of rater severity as being the adjustment of item difficulty for a specific rater. For polytomous items, a further source of complication is the step from one category to the next, which we will refer to as a threshold.

The particular model that we will be using in this investigation is a polytomous Rasch-type model (Rasch, 1960/1980) with a linear model on the difficulties to allow for the effect of raters. Within this formulation, the log-odds of being in one score category, compared to the adjacent lower score category, is modeled as a linear function of different effects, in this case, student ability (θ_n), item difficulty (δ_i), rater severity (λ_j) and threshold (τ_k): If,

Φ_{nij} is the log odds that student n with a "true" proficiency of θ_n will receive from rater j a

rating in category k as opposed to receiving a rating in category $k-1$ for item i ,

P_{nij} is the probability of student n being rated k on item i by rater j , and

$P_{nij k-1}$ is the probability of student n being rated $k-1$ on item i by rater j ,

then the log-linear model can be written:

$$\varphi_{nij k} = \ln[P_{nij k} / P_{nij k-1}] = \theta_n - \delta_i - \lambda_j - \tau_k, \quad (2)$$

We estimated the parameters using the random coefficients multinomial logit model (RCML; Adams & Wilson, 1996), and an adaptation of the *ConQuest* software (Wu, Adams & Wilson, 1998) which uses a marginal maximum likelihood estimation procedure.

The IRT approach offers several advantages over regression analysis. First, the IRT model treats the student responses as ordered categorical responses, while the regression approaches assume that the item scores are measured on a linear scale. Second, the regression models estimate the effects for each item separately and ignores the fact that the item responses are actually repeated measures, each taken by one student. The IRT approach estimates the item parameters simultaneously and it directly models that the students answer the whole set of items. Third, the regression model assumes the measured student's ability is a single fixed quantity and attributes all wrong answers on relatively easy questions as error. The IRT model, in contrast, views each student as capable of manifesting a range of responses according to the model, each with its own probability, and hence, incorrect responses to easy items are not necessarily a problem (although when there are many such, it may be considered misfit). Because of these advantages, we will use the IRT approach in this study.

We will use the estimates of rater severity from the mathematical model as an aid in detecting severe or lenient raters. Previous work (Engelhard, 1994; Wilson & Wang, 1995) has shown that such effects can be detected by pooling data across rating sessions. In particular, the Wilson and Wang (1995) paper shows that strong rater effects can persist even in the presence of standard rater quality control efforts such as rater training and a 10% read-behind protocol (as is used in the context below). In this paper, we will also examine rater performance *within* scoring sessions. Scores from raters can be periodically analyzed while they are working and then a determination can be made as to how good a job the raters are doing. If this information is found to be useful, it could serve as part of a feedback mechanism by which poor raters are identified

for counseling and, perhaps, retraining. The long-term goal of this research program is to investigate the usefulness of implementing a feedback mechanism to improve rater consistency.

There are several limitations to the following study that must be borne in mind when interpreting results. First, the follow-up interventions described below were not under any sort of systematic control by us. Follow-up was left entirely to the discretion of the site administrators. In particular, we have documentation of only one follow up based on the IRT information (Case Study 10 below). Thus, the results must be seen as descriptive only. We are mainly concerned (a) to see if it is feasible to generate severity information within a scoring session, (b) to investigate the variation in rater severity within rating sessions, and (c) to examine the relationship (or lack thereof) of those IRT severity patterns to the expert re-ratings. We do have records of interventions based on these “read-behinds”, but no comparable actions were taken based on the IRT information. We did have an opportunity to investigate some ancillary issues, and these will also be reported and discussed.

Method

Procedures at the Site.

This study was conducted using the 8th grade mathematics examination administered by the California Department of Education (CDE) for the Spring of 1994. CDE administered, statewide, a battery of tests designed to assess student ability in three grades: 4th, 8th, and 12th. In 1994 the exam included 8 multiple-choice items, 8 constructed response items (students are required to show their work even though the item is scored correct/incorrect), and 2 short performance tasks (of which only one was scored). There were 6 different forms, including common items and unique items. The exact format varies from year to year, for example, in 1993, the 8th grade mathematics examination included 7 multiple-choice items, and one performance task. The procedures were followed at two different scoring sites, which we will call sites A and B. At each scoring site, we had access to the multiple choice data, but not the

constructed response data, so we will ignore the constructed response data in the remainder of the analysis.

The procedures for the on-site study were as follows. Raters assembled in a large room for a one-week scoring session during which they would score two different performance tasks (three days for one and two days for the other). The raters were divided among seven tables of six under the supervision of a “table leader.” The table leader was an experienced rater who had succeeded as a rater before. The whole scoring session was managed by a “chief reader.” The raters first read the scoring rubric and then were shown examples of exemplar scores (answers that exemplified a particular scoring category within the rubric), followed by examples of borderline scores (answers, which would fall between scoring categories but for which a decision had to be made). After looking at the examples and discussing them with other raters and the table leaders, the raters were given 10 sample responses to score as a “calibration.” The “correct” scores for these sample responses had been agreed upon beforehand by a team of experts including the chief reader. In order to qualify as a rater it was necessary that the score given by the rater matched the expert panel’s score at least 8 out of 10 times. Raters who did not calibrate were retrained. If they failed to calibrate after a number of retrainings, they were sent home.

After the calibration test was passed, packets containing the responses from 20 students each were distributed to the raters who had calibrated. The responses within a packet were randomly assembled and the distribution of packets was also random. The rater would read the response and then assign a score between 1 and 4, by applying their understanding of the scoring rubric. After they had finished scoring the entire packet they handed it to the table leader and selected another packet. The table leader would then select two responses from the scored packet (10% of the responses) and score them again. In theory, this would be done without looking at the initial score, but in practice, this did not seem to be an entirely correct assumption. These were the so-called “read-behinds.” The table leaders maintained their own records of the results of the read-behinds and used CDE’s pre-established performance standards to determine whether a rater needed additional attention. If the rater's score differed by more than one from the table

leader's score on any performance task, the rater was immediately notified and the discrepancy was discussed and resolved. If, for any 10 consecutively sampled responses (across packets or within), the rater and table leader did not agree at least 80% of the time, retraining was administered. Retraining consisted of a discussion with the table leader about the general scoring rules and specific student responses that the table leader and rater had scored differently. Next, a calibration test was readministered. If the rater failed to achieve an 80% accuracy rate on the calibration test, then the retraining process was repeated. Depending on the extent of the problem, the rater might eventually have been dismissed.

After the table leader was finished with a packet, support staff would collect it and perform some basic quality control, such as checking to see if ID numbers matched. Then they would bring it to a scanning machine where the scores were scanned into a computer file. (This step and those that follow were not routine--they were carried out specifically for this study.) After the scores were scanned, they were combined with previously scored and stored multiple choice responses, using the student's ID as the key for the merge. After a sufficient amount of data about raters had been collected, analyses to estimate the IRT model were conducted. The data for these analyses were not cumulative, i.e., the analysis for the second period did not include the data from the first period. This decision was made because the goal was to identify intervals when the raters may have had changes in their performance. If the analyses had been cumulative, then results would be less sensitive to the most recent round of data available..

Analyses

One of our concerns was whether we would be able to produce results in a timely fashion. To that end, we tried to minimize the time needed for the analysis that was to be done by the computer. One of the best ways to speed up an IRT analysis is to reduce the number of parameters that need to be estimated and one way of doing this is to “anchor” some of the parameters. Anchoring means to fix parameters at a predetermined value so that they are not estimated during the analysis. We were able to calculate IRT parameter estimates for the multiple

choice item difficulties before going to the site because we were able to obtain a large sample of student responses in advance. This was possible because the multiple choice items were machine scored and therefore the results were available well before the performance tasks were scored. The results of the IRT analysis using only the multiple choice responses were then used as the anchoring values. Note that this means that the information available to link the data together consists of (a) the multiple choice data, (b) the data from the rater's ratings, (c) the data from the table leader's re-ratings, and (d) the data from the rater's (successful) calibration sets.

During the course of our preliminary discussions with the organizers of the scoring site, a variety of different ways to present our results were put forward. Although it is standard in the research literature to present IRT results in the logit metric, it became clear that it would be more useful to present our results in the metric of the item, i.e., the 1-4 scoring range use for mathematics performance tasks. This allowed the results to be compared directly to pre-existing standards of rater quality. The actual results provided to the organizers of the scoring sessions were the effect a given rater had on an average student's score, relative to what the student would have received from the average rater.

The conversion process involved calculating the difference between the expected score for an average student with an average rater and the expected score when that average student is rated by the rater in question. This difference, D_j , is the severity effect in score units of the rater.

The formula for the expected score for an average rater (E) is as follows.

$$E = \sum_{k=1}^4 kP_{ik} \quad (3)$$

where P_{ik} is the probability of a student at the mean being rated k on item i by an average rater (calculated using the RCML model given in (2))

The formula for the expected score for rater j (E_j) is

$$E_j = \sum_{k=1}^4 kP_{ijk} \quad (4)$$

where P_{ijk} is the probability of a student at the mean being rated k on item i by rater j .

Thus, we can calculate $D_j = E - E_j$.

We had two opposing goals in trying to determine how often we should run analyses. On one hand, we wanted to produce analyses as often as possible, which would minimize the amount of data available. On the other hand we wanted to have accurate estimates, which means having as much data as possible. The issue here is the size of the standard errors. Without a sufficient amount of data, the standard errors of our estimates would be so large that it would be almost impossible to identify any level of poor performance.

Based on past experience and advice from the session organizers we estimated that a typical rater would be able to score about 60 responses per hour. We then estimated standard errors in the score metric, based on different numbers of responses, and used these as a basis for choosing approximately 180 responses as a reasonable target. This also is approximately the number of student papers scored by a typical rater in a half-day session. As a half-day is an administratively convenient unit of time, we chose that as the period of time analysis.

Results

IRT Information on Rater Performance

Rater severities were calculated for each rater within each time period, along with a standard error for each severity. The size of the standard error is dependent on the number of responses scored by the rater and the amount of unexpected variation in the scores. To give an estimate of the confidence interval for the rater severity, we used a range of 1.96 standard errors on either side of the rater severity, which, under an assumption of normally-distributed errors, corresponds to an approximate 95 percent confidence interval. If this range did not include zero, then we concluded that the rater had a statistically significant bias, and the rater was flagged.

For each scoring period, we produced a chart that compared all the raters at the table and displayed a mean severity estimate and the approximate 95% confidence interval around the mean. Informal and anecdotal experience with charts that displayed the severities led us to believe that this was not the best way to communicate with the raters. Hence, we produced a

second chart designed to make the information more interpretable. This chart displays the severity estimates in terms of expected counts in each score category. We counted the scores in each category for each rater, and compared it to the counts expected for the average rater when scoring the same set of students. We found that the expected counts for an average rater were extremely close to the observed average across all students and raters, in fact, it could not be distinguished when displayed on a graph. This happens because the student responses were randomly assigned to raters, so the distribution of students assigned to any one rater should be similar, over the long run, to that of all the students assigned to all the raters. If the number of papers scored by a single rater is large enough this is a reasonable assumption. Thus, instead of calculating the expected counts, we displayed the “room average” counts, which was a saving in computational time (and also easier to explain to the raters). Hence, the second chart shows the distribution of scores assigned by an individual rater and compares that to the distribution of scores assigned by all raters. Figure 1 shows the distribution of scores for Rater 65 in comparison to the distribution of all the raters combined. The severity of rater 65 is reflected in the increased proportion of scores of 1, and the decrease in the proportions of scores 2, 3 and 4. The raters found this type of chart relatively easier to interpret--their interpretation was typically that Rater 65 had set the cut-off between a 1 and a 2 too high (which had also influenced the cut-off for the scores above). If this had been a feedback study, then we could have investigated the impact of such information. However, due to the exploratory nature of the study, we cannot report on systematic effects from such information (but, see Case Study 10 below).

Insert Figure about 1 here

Read-behind Information on Rater Performance

The read-behind approach provides the most direct method for checking rater performance but it assumes (a) that the table leader is assigning scores correctly, and (b) that a 10% sample was adequate to detect inconsistencies or inadequacies. Although it would have been better to have *all* the responses scored twice, the figure of 10% was standard in CDE sessions because of time and cost constraints. We produced figures that contained the read-behind results for the table that the rater was working at, as well as the individual rater. If a rater differed with the table leader more than 20% of the time, he or she was flagged. As an example, Figure 2 shows the performance of Rater 65 in comparison to Table 6 during period 1. During this period Rater 65 matched the read-behind score of the table leader for every paper. Looking at Figure 2, we can see that during the same period, the table average for matches was about 85%. For approximately 4% of the read-behinds the raters at the table gave a score that was one point higher than the table leader (rater score - table leader score = 1). For approximately 4% of the read-behinds the table leader gave a score that was one point higher than the raters (rater score - table leader score = -1). For approximately 2% of the read-behinds the table leader gave a score that was one point lower than the raters (rater score - table leader score = 2). In this case, the results from the read-behinds for rater 65 are quite contradictory to those from the IRT analysis. This we attribute to the light sampling employed in the read-behinds.

Insert Figure 2 about here

Case Studies of Rater Performance over Time

Severity estimates were calculated separately in the five periods for each of the raters. When examining the results of individual raters over time, a very wide range of scoring patterns can be observed. Some of these patterns will be described in the case studies listed below, which are illustrated in Figures 3 through 9. In these figures, at any time that a rater was flagged, the explanatory note is displayed below the X-axis. In addition, to the extent that we have documentation from table leader records that an intervention took place, (e.g., a discussion was

held with the rater), a note is included above the flag notes. Note that in these figures, the different severities are for the same rater over different periods.

The first two case studies are examples of raters who showed considerable drift in their severity.

Case Study 1. Rater 32 from Site A (Figure 3) was flagged in two consecutive periods but the IRT estimates were on opposite sides of the X-axis. This indicates that the rater had a major fluctuation in how s/he was scoring papers. In the first period the rater is scoring student responses at about 1 score above what an average rater would give, for about 3 student responses out of 10. In the second it had become about 2 out of 10 *below*. It is interesting to note that the IRT estimates were very consistent from the second to the fifth periods although they were somewhat severe. In addition, the rater picked up a read-behind flag in the fifth period.

Insert Figure about 3 here

Case Study 2. Rater 54 from Site B (Figure 4) started out with an IRT flag in Period 1. The IRT estimate was below the X-axis. By Period 3, the rater was flagged again and this time the estimate was above the X-axis. This rater displays a steady drift upwards in severity from Period 1 to Period 3 and then a steady drift downwards in severity from Period 3 to Period 5. In the first period the rater is scoring student responses at about 1 score below what an average rater would give on about 5 papers out of 10. In the second period this has been reduced to less than 1 out of 10. In the third period the rater switched to scoring papers at about 1 score above what an average rater would give on about 5 papers out of 10. Although there was no read-behind flag in either Period 1 or 3, there was a read-behind flag in Period 2. Again, it is an instance of a rater for whom the IRT flags and the read-behind flags are not in agreement.

Insert Figure 4 about here

In the next two cases, we show raters who showed very little drift in their IRT severity

Case Study 3. Rater 14 from Site A (Figure 5) represents a rater whose severity does not drift above 0, although it is never significant in a statistical sense. The difference in scoring is,

on average, about .1, which translates into a score of 1 more than an average rater would give on 1 out of 10 student responses. The rater receives three read-behind flags, although at no point was there an IRT flag. This is another indication that the two criteria report different types of information about rater performance, both of which may be important.

Insert Figure 5 about here

Case Study 4. Rater 14 from Site B (Figure 6) is someone whose severity never drifts below zero although it never is statistically significant in its difference from zero. The difference in scoring is, on average, about .1, which translates into a score of 1 more than an average rater would give on 1 out of 10 student responses. This rater had multiple read-behind flags but at no time triggers an IRT flag. In addition, there were records of interventions but no indication that it had any effect on the severity of the rater's performance.

Insert Figure 6 about here

In the next case study, we show a rater for whom interventions seemed not to have had the desired effect on rater drift.

Case Study 5. Rater 65 from Site B (Figure 7) received a Read-behind flag in Period 1 and an IRT flag in Period 3 but no intervention occurred until Period 4. In that period the rater has both an IRT flag and a Read-behind flag. However, after the intervention the rater's performance continues to drift down and the worst IRT estimate occurs in Period 5 along with a third Read-behind flag. Across the five periods, the difference in scoring is, on average, about -.2, which translates into a score of 1 less than an average rater would give, on 2 out of 10 student responses.

Insert Figure 7 about here

In the next case study, we show a rater for whom at least some of the interventions seemed to have had a desirable effect on rater drift.

Case Study 6. Rater 52 from Site B (Figure 8) received interventions from the table leader in all of the first four periods but the IRT estimate only seems out of line in the first period according to the IRT drift. There may have been some other type of indication that caused the table leader interventions, such as the rater talking too much at the table. In Periods 1 and 4 the difference in scoring is, on average, about .2 of a score, or 1 score more than an average rater would give for 2 out of 10 responses. In Periods 2 and 5 the difference in scoring is, on average, about -.1 of a score, or 1 score less than an average rater would give for 1 out of 10 student responses.

Insert Figure 8 about here

In the final case study, we show a rater for whom the intervention, which was based on the IRT information, seems to have succeeded.

Case Study 7. Rater 65 from Site A (Figure 9) received an IRT flag in Period 1 although the read-behind results showed all perfect matches (as is mentioned above). During the conference after Period 1, it was decided that the chief reader herself would check the scoring on 20 of the student responses. The results of these special read-behinds were that out of 10 scores assigned a 1 by the rater, 50% of them should have been a 2. This information was conveyed to the rater, and also led to an increase in the read-behind rate by the table leader. The rater's severity drifted up, and for the next 4 periods the rater's performance did not generate any more flags.

Insert Figure 9 about here

Comparison of IRT information with Read-behind information.

The number of instances when the two criteria were in agreement for the regular raters (i.e., not the table leaders) is detailed in Table 1. As might be inferred from the evidence provided by the case studies shown above, the two criteria are in agreement that there is an effect only rather rarely. But they are agreeing that there is no effect quite often. We cannot say which

criterion is correct from this study, as there was no absolute criterion available to determine which flaggings were correct. What we can say is that the IRT information is definitely not redundant with respect to the traditional read-behind information.

Insert Table 1 about here

Effect Size of Rater Severities

The case studies above and Table 1 give an impression that there is a discernible impact of rater severity. In order to get a measure of the overall impact of rater severity on the scores of students, we conducted a series of IRT analyses with the complete set of data from one of the sites. In these analyses, we calculated the effect sizes of (a) ignoring rater severities altogether, (b) considering rater severities as constant across periods, and (c) allowing rater severities to vary between periods. We did so by examining the total absolute residuals from each of these IRT analyses. The results are shown in Table 2: In the first column, we see that typically, if we ignore rater severity altogether, then that corresponds to an inaccuracy of between 9 and 14 score points in every 100 ratings. By adding in a constant rater effect, this is reduced to between 7 to 9 score points per 100. And allowing the rater severities to vary between periods reduces this to between 3 and 5 score points per 100 scorings. There is no absolute standard available here, but a reduction of 1/2 to 1/3 in the residuals is a clear indicator of the relative strength of the rater effects.

Insert Table 2 about here

Influence of Table Leaders and Tables

At Site A, the table leaders received IRT flags 20% of all possible times while the regular raters were flagged at a rate of 10%. At Site B, the table leaders received IRT flags 20% of all possible times while the regular raters were flagged at a rate of 14%. That is, the table leaders, who had previously been identified as expert raters, were found to be less consistent over time

than the regular raters. In interpreting this result, it is important to recall that the table leaders did not receive a random sample of the student work, as did the regular raters. Instead, their selection was (increasingly) determined by the performance of “poorer” raters, and often, their attention was drawn to specific score categories. Given this, it is not clear that their apparently less consistent behavior should be interpreted as indicating that they are not indeed good raters. However, given their status as leaders, it is reasonable to ask whether their rating severity or leniency was transmitted to the members of their tables.

An analysis was done to see if there was an effect from the raters being trained together and working together with a specific table leader. The analysis with the data from all periods combined was conducted with an additional parameter estimated for each table, but leaving out the data relating to table leaders. One possible explanation for a table effect would be that the raters absorbed the rating criteria of their table leader. If this was the case then we would expect the correlation between the table leader severity (from the full data set) and the table effect (from this analysis) to be high. The results are given in Table 3. There are indeed statistically significant table-leader effects, and statistically significant table effects, but these do not seem to be systematically related. At Site A there was a weak and statistically non-significant correlation ($r=0.24$, $p=0.60$) between the table leader and the table, while at Site B there was a larger negative (but still non-significant) correlation (-0.51 , $p=0.24$). Thus, we cannot report a consistent pattern of relationship between table effects and table leader effects.

Insert Table 3 about here

Conclusion

Procedures were implemented at two scoring sites and analyses were conducted on-site to determine IRT estimates of rater performance, and to examine patterns of rater drift in severity. It was found that we could obtain useful estimates from the ratings that accumulated over a half-day period. This information was communicated in ways designed to allow interpretation by raters and their table leaders. An array of case-studies was displayed to show the patterns of drift

in rater performance that was observed. From the different case studies we can see that: (a) rater performance drifts significantly from period to period in both a statistical and substantive sense, and (b) the effect of interventions seems to vary from rater to rater. Examination of the case studies, and tabulation of the results across the entire study shows that the IRT and read-behind criteria are contributing different information about the raters.

Additional analyses were conducted to investigate two issues. First, the overall impact of the estimated rater severities were calculated using residuals from a series of IRT analyses. These showed that use of rater severities that varied between periods could reduce the amount of error considerably. Second, we observed that table leaders exhibited surprisingly large amounts of rater severity and leniency, but that these did not seem to be associated with tendencies by the raters they supervised to drift in a consistent way.

The analyses that we carried out at the sites, and those we conducted later, and our reflections on these results lead us to a number of conclusions.

- (a) The use of IRT models for on-site feedback is both feasible and potentially valuable.
- (b) The use of training to eliminate rater severity effects is leaving many raters still with large severities.
- (c) The use of expert re-ratings of samples of student work (“read-behinds”) does not necessarily capture all the significant rater variation in severity.
- (d) Raters’ severities may drift significantly between scoring sessions.
- (e) Adjusting student scores for rater severities can be used to reduce the effects of rater severity variation.
- (f) The degree and direction of table leader severity does not necessarily determine the way that the severities of raters on that table will drift.

Moreover, our experiences at the sites also led us to a number of other conclusions that, although not directly supported by our data, are, we feel important observations. First, the gap between ratings and feedback needs to be minimized in order to maximize the usefulness of the feedback. In the case of the CDE context, having direct entry of rater’s ratings would make the

feedback process both simpler and more effective. Second, having a simple mechanism to allow the identification and re-presentation of student scripts would also make the feedback much more useful. This is simply too difficult to organize in a paper-driven system--the CDE would need to move to a scanned image system to accomplish this. Third, although we have concentrated on rater severity effects in this study, we also noted interesting cases that looked more like central tendency effects. It would be a logical next step to investigate the usefulness of IRT models of central tendency.

References

- Adams, R.J. & Wilson, M. (1996). Formulating the Rasch model as a mixed coefficients multinomial logit. In, G. Engelhard & M. Wilson (Eds.), *Objective Measurement III: Theory Into Practice*. Norwood, NJ: Ablex
- Braun, H.I. (1988). Understanding scoring reliability: Experiments in calibrating essay readers. *Journal of Educational Statistics*, 13, 1-18.
- Engelhard, G. Jr., (1994) Examining rater errors in the assessment of written composition with a many-faceted Rasch Model. *Journal of Educational Measurement*, 31(2), 93-112.
- Engelhard, G. Jr., (1996) *Models of Judgment and Rasch Measurement Theory* Presented at the American Educational Research Association, New York, April 1996.
- Guilford, J.P. *Psychometric methods*. New York: McGraw-Hill 1954.
- Hogan, T.P. (1981) *Relationship between free-response and choice-type tests of achievement: A review of the literature*. Green Bay, WI: University of Wisconsin. (Eric Document NO. ED 224 81).
- Khattri, N., & Sweet, D. (1996). Assessment reform: Promises and challenges. In, M. Kane & R. Mitchell, *Implementing performance assessment: Promises, problems and challenges*. Mahwah, NJ: Erlbaum.
- Lunz, M.E., Wright, B.D., & Linacre, J.M. (1990). Measuring the impact of judge severity on examination scores. *Applied Measurement in Education*, 3(4), 331-345.
- Mislevy, R.J. (1991) A Framework for Studying Differences Between Multiple-choice and Free-response Test Items in R.E. Bennet & W.C. Ward (Eds.), *Construction vs. Choice in Cognitive Measurement* Hillsdale NJ. Erlbaum
- Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Denmark's Paedagogistic Institut. (Reprinted 1980, University of Chicago Press).
- Raymond, M.R., & Viswesvaran, C. (1993). Least-Squares Models to Correct for Rater Effects in Performance Assessment. *Journal of Education Measurement*, 30 , 253-268.
- Resnick, D. & Resnick. L. (1996). Performance assessment and the multiple functions of educational measurement. In, M. Kane & R. Mitchell, *Impelementing performance assessment: Promises, problems and challenges*. Mahwah, NJ: Erlbaum.
- Saal, F.E., Downey, R.G., & Lahey, M.A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin*, 88(2), 413-428.
- Wang, W., & Wilson, M. (1996). Comparing open-ended items and performance-based items using item response modeling. In, G. Engelhard & M. Wilson (Eds.), *Objective measurement III: Theory into practice*. Norwood, NJ: Ablex.

Wilson, M., & Case, H. (1996, June). An investigation of the feasibility and potential effects of rater feedback on rater errors. Paper presented at the Council of Chief State School Officers National Conference, Phoenix, AZ.

Wilson, M., & Wang, W. (1995). Complex composites: Issues that arise in combining different modes of assessment. *Applied Psychological Measurement*, 19(1), 51-72.

Wu, M., Adams, R.J., & Wilson, M. (1998). *ConQuest* [computer program]. Melbourne Australia: ACER Press.

Table 1
Number of Flaggings Per Site Per Criterion

Site	Agreement		Disagreement	
	No Flags	Both Flags	Only IRT Flag	Only R-B Flag
A	149	3	18	35
B	147	4	26	33

Table 2
Effect Sizes of Rater Severities at Site B

Period	Ignore Rater	Constant Rater	Rater within Cycle	Number of Students
1	.10	.07	.03	4375
2	.12	.09	.04	3783
3	.14	.09	.05	3418
4	.09	.07	.03	4540
5	.12	.08	.04	4962

Table 3

Table Leader Severities and Table Effects

Table	<u>Table Leader Severity</u>		<u>Table Effect</u>	
	Site A	Site B	Site A	Site B
1	-0.455	-0.183	-0.098	-0.019
2	-0.258	0.021	0.088	-0.205
3	0.042	0.017	0.035	0.050
4	-0.263	0.158	0.039	-0.052
5	-0.244	-0.109	0.025	0.009
6	-0.375	0.162	0.016	0.157
7	-0.153	-0.166	-0.105	0.060

Figure Captions

Figure 1. Expected and actual score distributions for rater 65 at Site A.

Figure 2. Read-behinds for rater 65 at Site A.

Figure 3. Case study 1.

Figure 4. Case study 2.

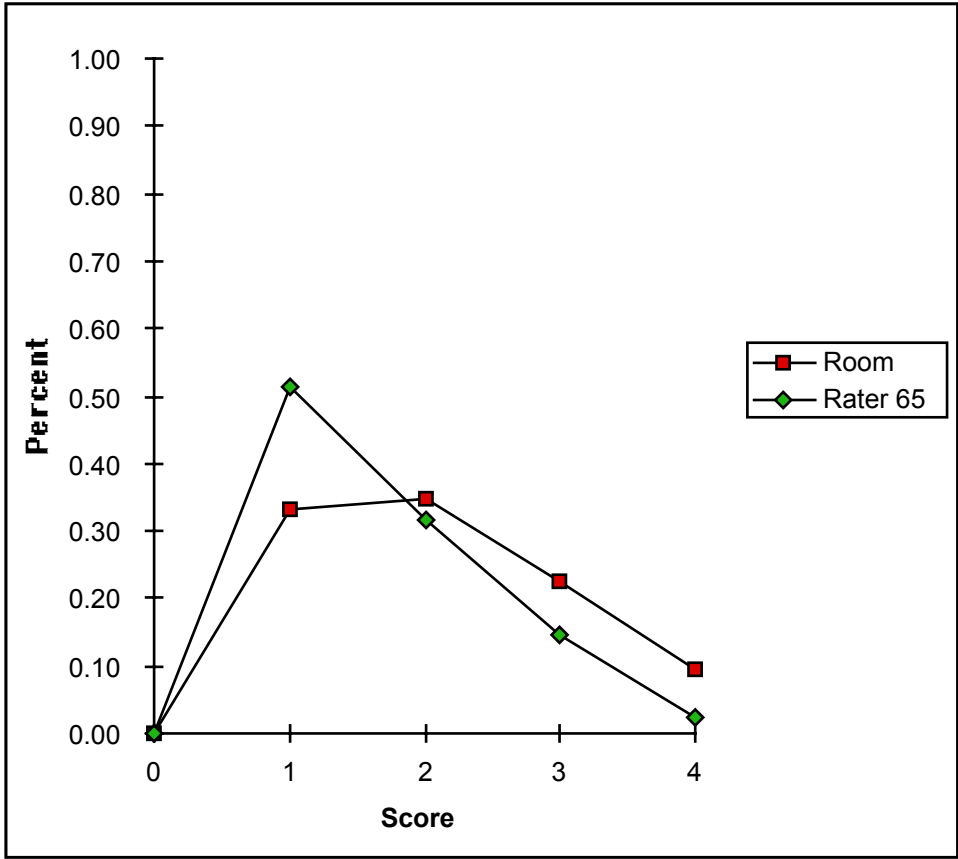
Figure 5. Case study 3.

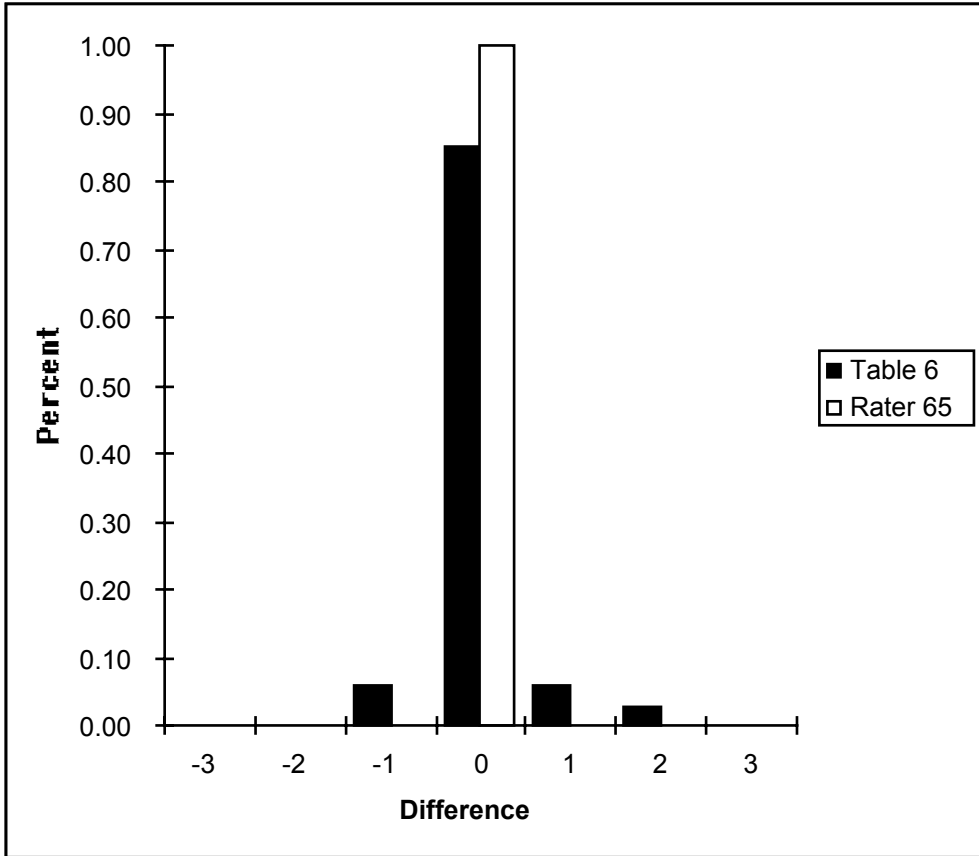
Figure 6. Case study 4.

Figure 7. Case study 5.

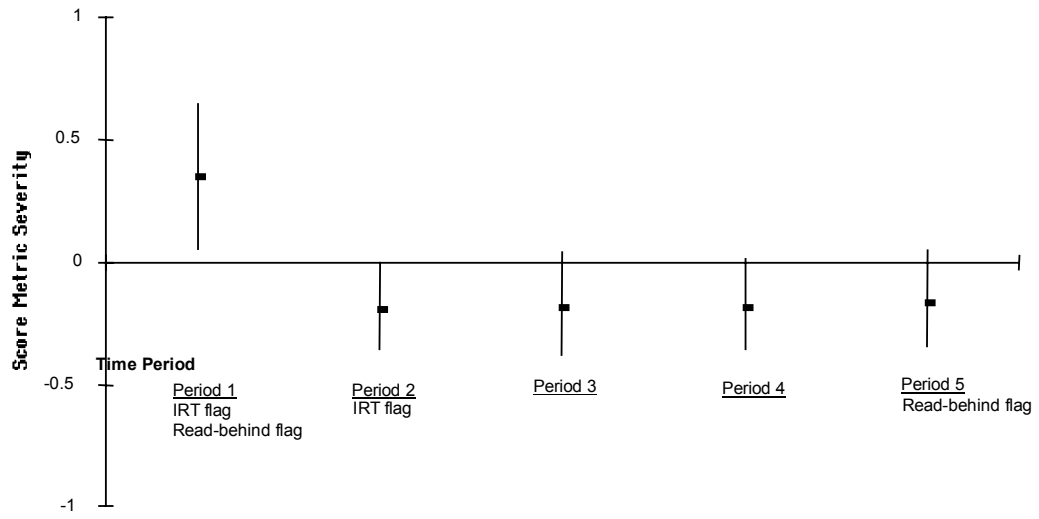
Figure 8. Case study 6.

Figure 9. Case study 7.

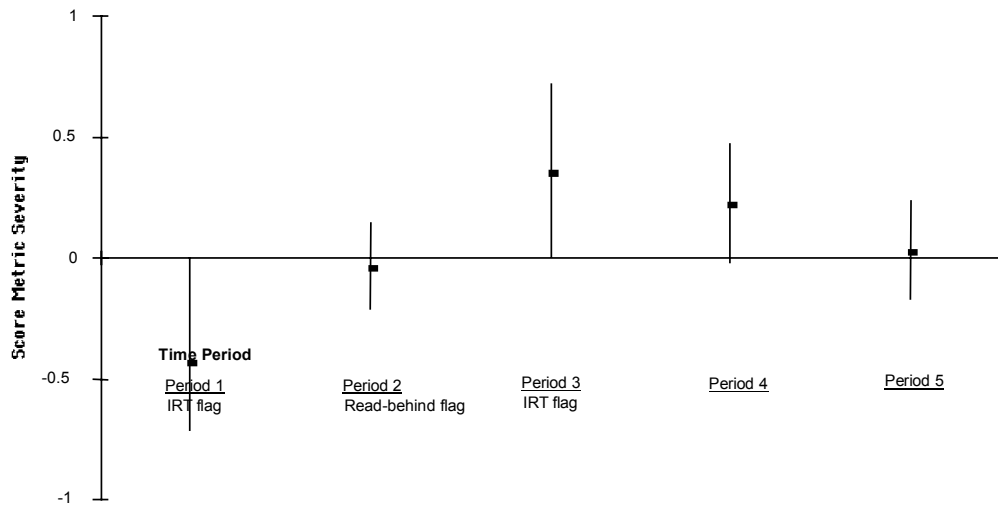




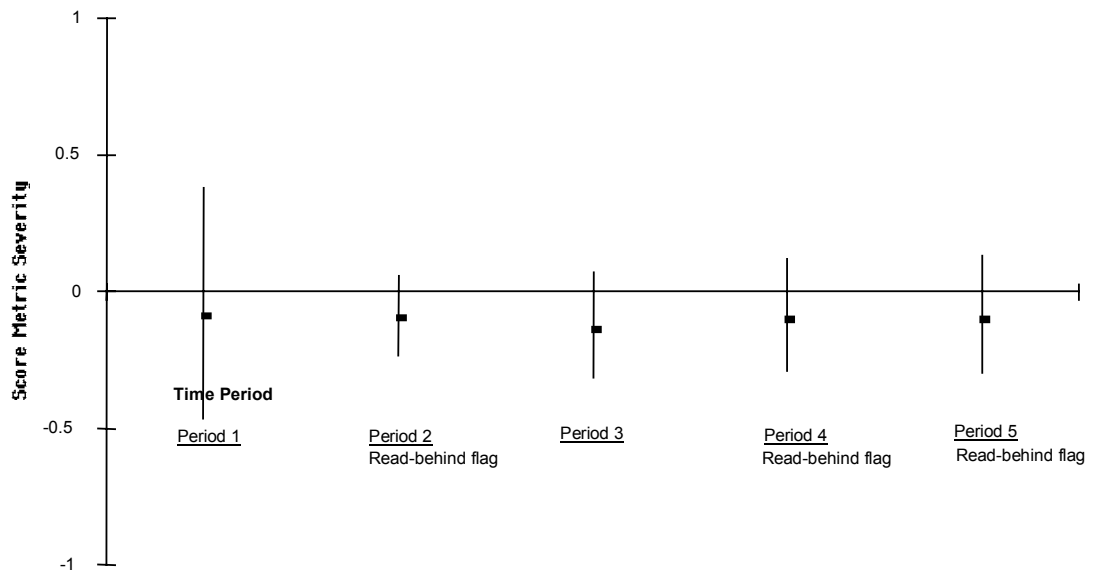
Case Study 1 – Rater 32



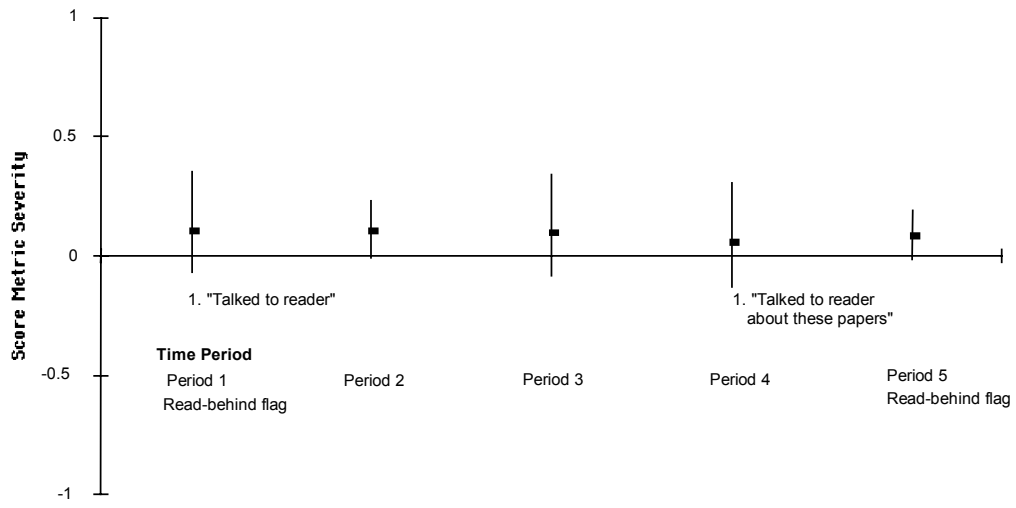
Case Study 2 – Rater 54



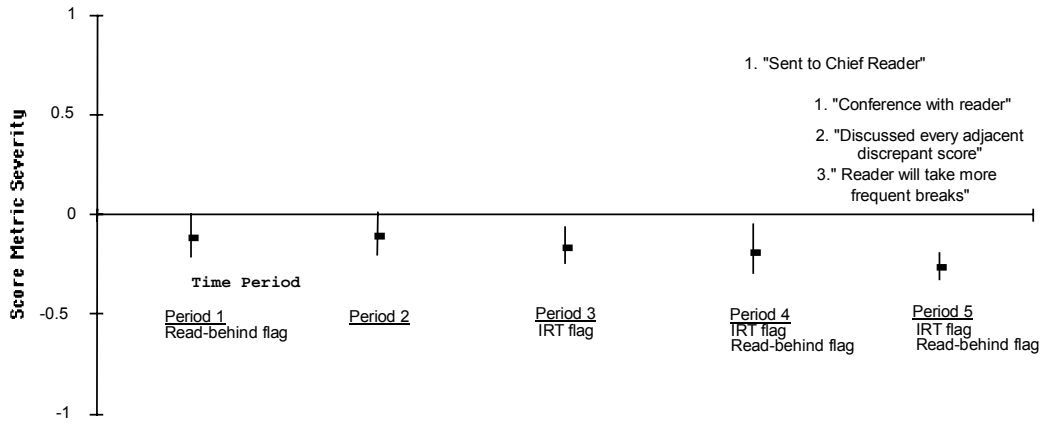
Case Study 3 -- Rater 14



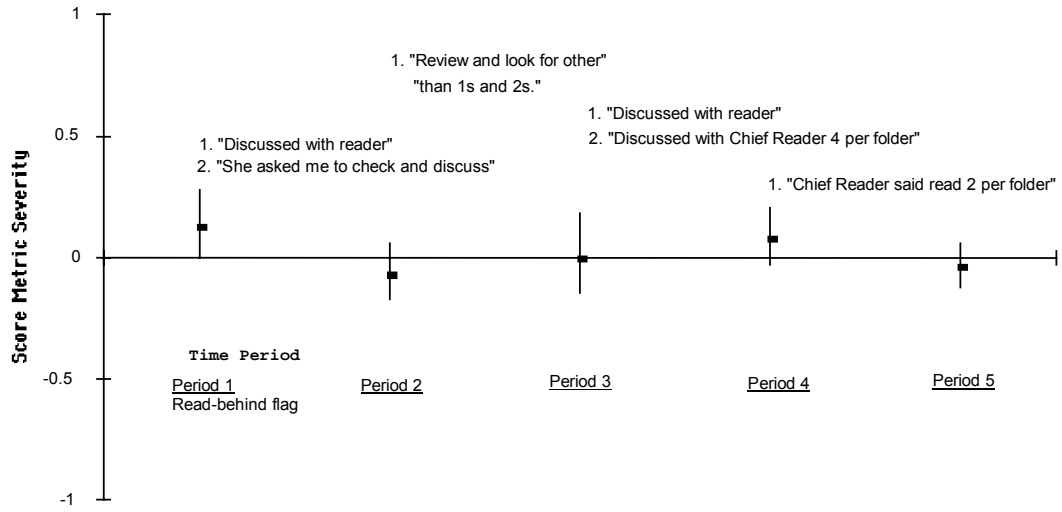
Case Study 4 -- Rater 14



Case Study 5 -- Rater 65



Case Study 6 -- Rater 52



Case Study 7 -- Rater 65

