DOCUMENT RESUME

ED 411 303 TM 027 338

AUTHOR McQueen, Joy; Congdon, Peter J.
TITLE Rater Severity in Large-Scale Assessment: Is It Invariant?
PUB DATE 1997-03-00
NOTE 41p.; Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, IL, March 24-28, 1997).
PUB TYPE Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Educational Assessment; Elementary Education; Elementary School Students; Foreign Countries; *Interrater Reliability; *Scoring; *Test Interpretation; *Writing Tests
IDENTIFIERS *Rasch Model; Rater Reliability; *Rater Stringency Error

ABSTRACT

A study was conducted to investigate the stability of rater severity over an extended rating period. Multifaceted Rasch analysis was applied to ratings of writing performances of 8,285 primary school (elementary) students. Each performance was rated on two performance dimensions by two trained raters over a period of 7 rating days. Performances rated on the first day were rerated at the end of the rating period. Statistically significant differences between raters were found within each day and in all days combined. Daily measures of the relative severity of individual rates were found not to be invariant when compared to single, on-average measures for the whole rating period. For more than half of the raters, severity measures on the last day were significantly different from measures on the first day. These findings must cast doubt on the practice of using a single calibration of rater severity as the basis for adjustment of person measures. (Contains 8 figures, 12 tables, and 31 references.) (Author/SLD)

Rater Severity in Large-Scale Assessment: Is it Invariant?

Joy McQueen and Peter J Congdon

Australian Council for Educational Research

Abstract

The purpose of this study was to investigate the stability of rater severity over an extended rating period. Multifaceted Rasch analysis was applied to ratings of writing performances of 8285 primary school students. Each performance was rated on two performance dimensions by two trained raters over a period of seven rating days. Performances rated on the first day were re-rated at the end of the rating period. Statistically significant differences between raters were found within each day and in all days combined. Daily measures of the relative severity of individual raters were found not to be invariant when compared to single, on-average measures for the whole rating period. For more than half of the raters, severity measures on the last day were significantly different from measures on the first day. These findings must cast doubt on the practice of using a single calibration of rater severity as the basis for adjustment of person measures.

Introduction

Performance assessment has been enthusiastically espoused for its directness and for its potential for positive washback. However, as Messick (1994, p.13) reminds us, it is necessarily subject to the same validity criteria as other forms of assessment. As with other forms, all care must be taken to minimise the distorting effect of construct-irrelevant variance.

A number of sources of error have been discussed in the literature on performance assessment (see, for example, Cantor & Hoover, 1986; Engelhard, 1992; Engelhard, Gordon & Gabrielson, 1991; Gabrielson, Gordon & Engelhard, 1995; McNamara, 1996; Ruth & Murphy, 1988). Prominent among these sources is the variance associated with raters. This is a reflection of the concern that, no matter how carefully constructed, the reliability of a rating scale is critically dependent on the raters who operate it (Overall & Magee, 1992). As Dunbar (1991, p.291) puts it, "fallible raters can wreak havoc on the trustworthiness of scores and add a term to the reliability equations that does not exist in tests that can be scored objectively". This paper is concerned with identification of some rater effects, and particularly with changes in rater behaviour from one occasion to the next.

Inter-Rater agreement

The subject of inter-rater agreement extends far back in the measurement literature (see for example Diederich, French & Carlton, 1961; Guilford, 1954; Huddleston, 1954; Thurstone, 1927). Dissatisfaction has been particularly evident in relation to the use of rating scales which extend over multiple qualitative categories, such as those scales used for the direct assessment of writing. Considerable evidence of poor rater agreement exists (e.g. Coffman & Kurfman, 1968; Engelhard, 1992; Lumley, Lynch & McNamara, 1994) and where adequate agreement is reported, it is often on the basis of correlations alone. As Lunz, Stahl and Wright (1994) demonstrate, even a perfect correlation may ignore systematic differences between raters.

The rater training which is a common feature of rating programs is presumably in part intended to maximise inter-rater agreement. However, it has been shown that even extensive training has little effect on the standards maintained by raters (Engelhard, 1992; Linacre, 1991; Lunz & Stahl, 1990; Weigle, 1994). In practice, the main benefit of training appears to be orientation of the rater to the rating scale, and improvement of raters' self-consistency (Wigglesworth, 1994).

Not only is rater agreement difficult to attain, but the value of such agreement is far from axiomatic. It does not, of itself, guarantee rating quality (Buckner, 1959; Saal, Downey & Lahey, 1980, citing Freeberg, 1969). A number of writers (Barritt, Stock & Clark, 1986; Hake, 1986; Lumley & McNamara, 1995; Weigle, 1994) have warned of the dangers of forced agreement, and have highlighted individual self-consistency as a more worthy aim of training programs.

In the absence of rater agreement, raters are not equally likely to award the same score to the same performance. Severity, the relative likelihood of raters to award lower scores, and leniency, its counterpart for higher scores, are phenomena which can turn an assessment into a lottery. Since few assessment programs can afford the time and expense of having every piece of work assessed by every rater, steps must be taken to ensure that no candidate is disadvantaged by the chance allocation of his or her work to a particular rater, however self-consistent that rater may be. As Webb (1990, p.16) puts it, whenever an incomplete rating design is used, investigation of rater effects is an ethical obligation. Any discovery of significant rater effects should inform quality control procedures during the rating period, and indicate where adjustment of candidates' scores is needed to compensate for these effects.

Effect of time

The Rasch model and extensions of it produce measures which are separable and invariant when the data fit the model. In using these models to produce measures that are comparable across different groups of persons or occasions, the measures of the instrument are required to remain invariant. "Only if the item calibrations are invariant from group to group and from time to time can meaningful comparisons of person measures be made" (Wright & Masters, 1982, p.114). With the inclusion of raters in multifaceted Rasch models the same requirements applied to items also apply to raters. It is acknowledged that raters are not items, nonetheless raters do have the potential to be influenced by a greater array of variables than items have, making them more susceptible to performances that are not invariant from one occasion to the next.

> *Judges may worry about being "too easy" or "too severe" ... and so may under- or overcompensate ... Judges' grades may be affected by personal factors such as hunger, fatigue, illness, or disagreement with fellow judges ... by lights too bright or too dim and by rooms too hot or too cold.* (Lunz et al., 1994, 914-915)

The effect of factors such as these will vary from one occasion to the next. It is reasonable then to suppose that rater severity will also vary with time.

In an early study that looked at rater severity over time, Coffman (1968), using analysis of variance, found that raters of history papers rated more harshly on the second day than the first.

Webb (1990) used an ordinary least squares regression approach to investigate the ratings from administrations of an oral certificate examination in the health profession over a three-year period. Her findings indicate high stability within years, but, for some raters, a high degree of change between years.

Lunz and Stahl (1990) used multifaceted Rasch (MFR) analysis in an investigation of the stability of rater severity over grading periods (from this point we use the term *severity* to refer to the severity/leniency continuum, and individuals' positions on it). Analysing the data from three different examinations (an English literature essay examination, a clinical examination and a health profession oral examination) they found that raters demonstrated significant instability in two (essay and clinical) of the three, over grading periods ranging from one to four days.

Myford (1991) also used MFR, this time to analyse the ratings of dramatic performances, and found that three groups of judges with varying levels of expertise (buffs, experts and novices) all showed significant changes in severity over a period of one month.

In a further MFR study, Lumley and McNamara (1995) investigated three sets of ratings given over a 20 month period for a test of spoken English. Examining both main effects and rater-time interactions, they found significant changes in rater severity.

The purpose of this study is to examine the severity of individual raters at several points within a rating program. Each of these points is compared with a single, on-average measure (gross) severity for each rater. In addition, raters rescore the same performances at the beginning and end of the program, and rater severities are compared. The intention is to examine the validity of using a single, on-average estimate of rater severity to adjust person measures.

Method

The Test Background

The writing test discussed here is one component of a larger, state-wide program of literacy and numeracy testing. The purpose of the program was to provide teachers and parents with information on individual student performances.

In 1996 almost 47 000 students participated. They were allowed 30 minutes to write up to two pages in response to a single prompt designed to elicit a newspaper report of a recent event.

Scoring followed a criterion-referenced rating scale which had been derived from the curriculum framework in use. There were six described levels for each of the two performance dimensions ("Overall Performance" and "Textual Features") within the scale. Each paper was scored by two raters.

Because of time constraints, papers were marked approximately in the order in which they arrived. They were handled in bundles of about 15.

A half-day training session was conducted in the week prior to the start of rating. Following the training, the raters scored a set of papers, and multifaceted Rasch (MFR) analysis was carried out so that rater severity and fit could be estimated. Raters with unacceptable fit (i.e. those who were grading inconsistently) were excluded from the rest of the program. Raters who were not using the whole scale, or who were significantly more lenient or more severe than the other raters, were counselled.

Raters worked at designated tables of about eight, each with its table leader. The role of the table leader was to monitor the quality of scores given by raters at her table, and to counsel raters whose ratings differed by more than one grade from those given by her. She also acted as a point of reference and offered guidance in dealing with problem scripts.

The study

This study was designed to compare some of the rating characteristics of a subset of 16 of the operational raters. The design of the study linked the 16 raters over a period of seven working days (Monday to Friday, Monday and Tuesday). For the remainder of this paper the rating days are referred to as Days 1, 2, 3, 4, 5, 8 and 9, respectively. On Days 6 and 7 (Saturday and Sunday), no ratings took place. Day 9 was dedicated to re-rating the performances scored on the first day.

The 16 raters were randomly selected from the pool of raters who were prepared to work in both the morning and the afternoon sessions each day. Twelve of the 16 had previously participated in at least one rating program of this kind.

During the rating period the 16 raters graded 8285 papers. Each day bundles of papers to be rated by these raters were taken randomly from the bundles available for that day. Two groups of eight raters worked at separate tables. Each table had its own leader, who was not one of the eight raters, and whose ratings are not considered in this study.

Movement of bundles followed two separate regimes each day, one to establish links between raters at the same table (Regime 1) and one to establish links between tables (Regime 2). Raters followed Regime 1 in the morning session on Days 1, 3 and 5, and in the afternoon session on Days 2, 4 and 8.

Under Regime 1 the two tables operated independently, i.e. there was no cross-over between tables. Each paper was rated by two raters at the same table, and each person rated six bundles of about 15 papers each. Figure 1 represents the pattern of linking within each table. Each line in the figure represents one bundle of about 15 papers.

FIGURE 1 ABOUT HERE.

In order to maintain a smooth flow of performances to be scored, it was necessary to specify the order in which bundles were to be processed. Table 1 shows the allocation of bundles (A to X, listed vertically on the left-hand side of the table) to raters (1 to 8, listed in the top row), and the order in which they were to be read ("1st", "2nd", etc, in the body of the table). So, for example, Rater 1 was to rate bundles A, I, Q, H, O and V, in that order. In this way, each rater was directly linked with six other raters at the same table, and indirectly with the remaining one. When raters had completed their six bundles, they devoted the rest of the session to bundles from the general pool of scripts. These additional ratings are not considered in this study.

Links between the two tables were established in the remaining session each day (Regime 2). In the first half of the session, papers were rated for the first time. In the second half, papers which had been given their first ratings at Table 1 were rated a second time at Table 2, and vice-versa. For the purposes of this study, each rater averaged 173 ratings per day on each performance dimension.

The raters were linked within days but not across days except for the last day, which consisted of a blind rescoring of the performances scored in the morning session of Day 1.

TABLE 1 ABOUT HERE.

Measurement Model

The model used to analyse the data from this study was a multifaceted version of the Rasch model, ConQuest (Wu, Adams & Wilson, 1996). The ConQuest software produces measures of all terms and elements in a common metric (logits) and a variety of commonly used fit statistics. The label *term* is used to describe a group of components, e.g. raters. The label *element* is used to identify components of the term, e.g. Rater 1, Rater 2, Rater 3 etc.

The terms that were modelled in the analysis were person ability, rater severity and a rating scale step structure. It was intended that the structure of the rating scale remain constant across these terms. Separate analyses were made on each of the performance dimensions, Overall Performance and Textual Features (OP and TF). Initial analysis attempted to calibrate OP and TF together. However, the fit statistics were low for both performance dimension elements, and high for all raters. This suggested that the amount of agreement between these elements was far greater than the amount of agreement between the raters, and that raters were not treating these dimensions as independent variables. A requirement of the measurement model used is that the elements and terms be independent. The two performance dimensions were therefore analysed separately.

The model used was:

$$\ln [P_{nij} / P_{nij\text{-}1}] = B_n - (R_i + S_j)$$

where:

P_{nij} = Probability of person *n* being rated *j* by rater *i*

$P_{nij\text{-}1}$ = Probability of person *n* being rated *j*-1 by rater *i*

B_n = Writing ability of person *n*

R_i = Severity of rater *i*

S_j = Difficulty of scoring step *j* relative to step *j*-1.

This model was first applied to all of the data collected over the whole program to produce a gross measure of rater performance. The same model was then applied separately to the data from each of the six operational scoring days (Days 1 to 8), to give daily measures of rater performance.

To preserve the comparability of rater performance, person ability and score step difficulty were constrained to a mean of zero logits for each of these calibrations. These calibrations resulted in seven unconstrained measures of rater severity and rater fit for each performance dimension. Because the rater measures were unconstrained, differences in the mean value of rater severity from any single calibration could have represented differences in the mean ability of the group of person performances. Subsequent analyses involving relative rater severity have removed this effect from each calibration by re-centring the relative rater severity measures to a mean of zero logits. The constraining of terms to a set mean value uses information from more than one element within the term to determine the measure of any single element. This procedure can influence the fit statistics and cause underestimation of the standard errors of all elements within the term. Nominating *rater* as the only unconstrained term in the model has produced measures that are the most appropriate for comparing rater performance with this software.

Rater Severity Comparisons (common raters)

Each calibration estimates the relative severity of the raters. Variations between raters are tested for significance. Significant variation implies that person measures are rater-dependent unless adjusted for rater severity. This method of analysis was used to determine the level of difference in rater severity in any single calibration, giving us a test of within-occasion variation across raters.

The rater severity measures were further analysed for within-rater variation across occasions. The rater measures used here are not directly comparable in the absolute sense, as there were no common performances scored across occasion. However, it remains valid to look at the stability of the relative rater severity estimates from each calibration. These measures will not take account of any changes in the group as a whole. The method used to compare relative rater severity estimates was to calculate the standardised difference between the gross rater severity (i.e. rater severity over the rating period as a whole) and the daily rater severity.

Rater Severity Comparisons (Common Performances and Common Raters)

To produce rater measures that were directly comparable between Day 1 and Day 9, only those performances which were scored on both occasions were included in the data set. The same raters scored the same performances on both occasions.

The terms modelled in this analysis were person ability, rater severity, day difficulty, rater by day interaction and a rating scale step structure. It was intended that the structure of the rating scale remain constant across these terms. Separate analyses were made on each of the performance dimensions, OP and TF.

The model used was:

$$\ln [P_{nikj} / P_{nikj\text{-}1}] = B_n - (R_i + D_k + C_{ik} + S_j)$$

where:

P_{nikj} = Probability of person *n* being rated *j* by rater *i* on day *k*

$P_{nikj\text{-}1}$ = Probability of person *n* being rated *j*-1 by rater *i* on day *k*

B_n = Writing ability of person *n*

R_i = Severity of rater *i*

D_k = Difficulty of day *k*

C_{ik} = Rater *i* by day *k* interaction

S_j= Difficulty of scoring step *j* relative to step *j*-1.

This model tests for the effect of a general shift in rater severity across occasions by using the *day* term as a main effect. The model also tests for individual rater differences in the magnitude and direction of severity changes between occasions by using the interaction term "*rater* by *day*".

The *rater* term was left unconstrained. No adjustment of rater severity measures was required for comparative purposes as only data that were scored on both days were used in this calibration.

Results

Data Fit to the Model

The weighted mean square residual was selected as a measure to monitor the adequacy of the data fit to the model. This statistic was developed for marginal maximum likelihood estimation procedures for generalised item response models (Wu, 1997) and provides a measure of the relative consistency of each rater's performance. Values greater than 1.0 indicate that there is more variation in the observed responses than expected and, conversely, values less than 1.0 indicate less variation than expected. The level at which these values become problematic is arbitrary and perhaps dependent on the impact that person measures will have on individuals. Acceptable ranges suggested by other researchers include 0.5–1.5 (Lunz, Stahl & Wright, 1996) and 0.8–1.2 Linacre (1989). Generally speaking, in the data presented, only Rater 4 showed an unacceptable level of misfit.

From the results shown in Tables 2 and 3, it would appear that, for OP, Rater 4 was the least consistent rater, and Rater 10 the most consistent. For TF, Rater 2 was the most consistent, and Rater 6 the least.

TABLE 2 ABOUT HERE.

TABLE 3 ABOUT HERE.

Reliability

The ConQuest software used here produces a separation reliability value (Wright & Masters, 1982, pp.91-94). This value is the proportion of the observed variance that is not due to measurement error. When applied to the rater term, it describes how well the elements (raters) within the term are separated in order to define rater severity reliably. Rater separation reliability results are shown in Tables 4 and 5. These values suggest that there were meaningful differences in rater severity levels. The slightly lower value for Day 5 on the OP dimension could reflect the smaller amount of data contributed by that day and or less variation in between-rater severity.

TABLE 4 ABOUT HERE.

TABLE 5 ABOUT HERE.

Rater Severity Comparisons (Common Raters)

The results from each of the fourteen calibrations (Days 1, 2, 3, 4, 5, 8 and "gross"), for both performance dimensions, produced a significant ($p<0.001$) difference in the variation of rater severity as a main effect, indicating that person measures, if left unadjusted, would be rater-dependent.

To demonstrate the impact of the differences found in rater severity, the rater severity measures from the gross calibration on the OP dimension are shown in Table 6. Together with each rater's severity measure is their expected use of the score categories and expected score given to a person of average ability. Rater 2 was the most severe rater and Rater 5 the least severe. The modelled probabilities of each score category suggest that Rater 2 would give a score of 3 to an average performance only 18 times out of 100, where Rater 5 would give a 3 to an average performance 58 times out of 100.

The rater severity measures from the gross calibration on TF are shown in Table 7. As for OP, Rater 2 was the most severe rater and Rater 5 the least severe. The modelled probabilities predict that Rater 2 would give a score of 3 to an average performance only 12 times out of 100, where Rater 5 would do so 52 times out of 100.

TABLE 6 ABOUT HERE.

TABLE 7 ABOUT HERE.

The rater severity results from the daily and gross calibrations are shown in Table 8 for Overall Performance, and in Table 9 for Textual Features.

For each rater the range in these measures is shown. The range column shows by how much relative rater severity changed across occasions. The average of these ranges was 0.98 logits for OP, and 1.02 for TF. To measure the impact of the differences in relative rater severity found across the occasions on the person measures, each rater's range was divided by the standard deviation (1.89 logits for OP, 2.17 for TF) of the person measures produced from the gross calibrations. These values, shown in the last column of the tables, represent the extent to which person measures could change, depending on which day's calibration was used when adjusting person measures for rater severity.

For OP, the average value across raters was 52% of one standard deviation; for TF it was 47%. In the case of Rater 4, if individual person measures for OP were adjusted using the rater severity estimates from Days 3 and 8, they could change by more than one standard deviation. For Rater 5, on TF using the rater severity estimates from Days 2 and 5, individual person measures directly connected to this rater could change by nearly one standard deviation.

It can be seen from these results that these raters' severity measures were not invariant over occasion and the impact on the person measures would not have been trivial.

A comparison of the ranges for each rater between performance dimensions showed that the differences found in relative rater severities within one dimension did not necessarily materialise in the other. Seven of the sixteen raters' ranges changed by 0.4 logits or more between performance dimensions. This change is equivalent to one third of the average range found across raters.

The standard deviations of each calibration show that the variance in rater severity started relatively low on Day 1 and increased up to Day 3 for both performance dimensions. Within the OP dimension the variance in relative severity decreased from Day 3 to Day 5 then increased again on Day 8. Within the TF dimension the standard deviation in relative severities appeared to plateau from Day 3 onwards. The relatively low standard deviation in the gross measure of rater severity for both performance dimensions would indicate that these raters were not consistent in their relative severities across the days. If raters had maintained the same magnitude and direction of severity across the days, the standard deviation for the gross measure would have been larger than those reported here. That is, if a rater is relatively severe on one day and lenient on another the sum effect on the gross standard deviation will tend towards zero.

TABLE 8 ABOUT HERE.

TABLE 9 ABOUT HERE.

To test further the stability of relative rater severity measures over occasion, the gross measure of rater severity was compared to the daily measure for each rater.

The measure of gross rater severity is in part made up of the same data that is used for the daily measure of rater severity. As Table 10 shows, the contribution of performances from the individual days to the performances used in the gross calibration was lowest from Day 5. Because of this, the difference between the rater severity measures on Day 5 and the gross measures may be overestimated relative to the other days. All of the standardised differences reported here are undersestimated, as the gross measure is in part made up of the daily measure that it is being compared with. This covariance between measures was not accounted for in the calculation of standardised differences.

TABLE 10 ABOUT HERE.

The results from the analysis of standardised differences are shown in Table 11 and Table 12. From the chi-squared values there were three features that stood out. Firstly, Day 5 (OP) was the only day where the chi-squared value was not significantly ($p<0.05$) different from the gross calibration. However, within that day there still remained two raters with significantly different measures of rater severity.

Secondly, for both dimensions, the chi-square value for each occasion tended to become smaller as the working week progressed, and indicated that this group of raters may have required at least three days of operational scoring before producing relative severity measures that were stable.

Finally, for both performance dimensions, on the first day of the next working week (Day 8) there was an increase in the overall chi-squared value for that day compared to the preceding two days. This pattern, which we have labelled the "weekend effect" is best seen in Figures 2 and 3.

On the OP dimension, the number of raters with daily severity measures which were significantly ($p<0.05$) different from their gross severity measure was 8, 11, 7, 2, 2 and 6 from Days 1 to 8 respectively. Once again we see a trend towards relative stability as the working week progressed, and a reversal of the trend on the first day of the new working week. For TF, where the number of raters showing significant differences was 7, 12, 7, 5, 4 and 4 on Days 1 to 8 respectively, the trend of increasing stability towards the end of the week was evident but there was no apparent reversal on Day 8.

TABLE 11 ABOUT HERE.

TABLE 12 ABOUT HERE.

FIGURE 2 ABOUT HERE.

FIGURE 3 ABOUT HERE.

Rater Severity Comparisons (Common Performances and Common Raters)

It will be recalled that Day 9 was devoted to a re-rating of scripts from Day 1. This produced directly comparable measures of rater severity. On both performance dimensions the calibrations reported a significant (p<0.001) *rater* effect, a significant (p<0.001) *day* effect and a significant (p<0.001) *rater* by *day* interaction effect. The magnitude of the main effect for *day* was 0.48 logits for OP and 0.84 logits for TF. That is to say, between Day 1 and Day 9, raters as a group became more severe by the amounts quoted. Additional to these main effects is the interaction effect of *rater* by *day*. After taking account of this effect, the difference in average rater severity between Days 1 and 9 was 0.45 for OP and 0.76 TF respectively. This represents differences of 0.14 and 0.20 of a score point, or 24% and 35% of a standard deviation, for a person of average ability. These differences are comparable with the differences in gross severity between the most lenient and the most severe raters. The maximum difference found for any single rater was 0.69 of a score point for a person of average ability.

A significant *rater* by *day* interaction effect indicates that the change in rater severity was not the same for all raters in direction and/or magnitude. To illustrate all of these effects, Figures 4 and 5 show, for each rater, the expected average score for persons of average ability across the two occasions.

FIGURE 4 ABOUT HERE.

FIGURE 5 ABOUT HERE.

Tests of significant differences between the rater severity measures from Day 1 and Day 9 on OP show that nine raters became significantly more severe and one rater became significantly more lenient between these occasions (Table 13). These results have been plotted in figure 6, where a scatterplot of the two severity measures is presented with 95% confidence intervals.

TABLE 13 ABOUT HERE.

FIGURE 6 ABOUT HERE.

On the TF dimension, ten raters became significantly more severe and one rater became significantly more lenient between Day 1 and Day 9 (Table 14, Figure 7).

TABLE 14 ABOUT HERE.

FIGURE 7 ABOUT HERE.

Nine of the ten raters whose severity changed significantly ($p<0.05$) from Day 1 to Day 9 on the OP dimension, also changed significantly ($p<0.05$) on the TF dimension. Raters 5 and 16 showed significant change in severity only on the TF dimension. The direction of these changes in rater severity was not the same across dimensions for all raters. Only Rater 4 became significantly more lenient on the OP dimension. However, on the Textual Features dimension Rater 4 became significantly more severe. Rater 15 was the only rater to become significantly more lenient on the TF dimension, however, on the Overall Performance dimension Rater 15 became significantly more severe. Four raters (1, 8, 10 and 12) showed no significant change between Day 1 and Day 9 on either performance dimension.

To examine the question of whether raters maintained their relative severity across performance dimensions the gross rater severity estimates from the OP and TF calibrations were plotted against each other with their 95 per cent confidence intervals (Figure 12). Those raters whose values fall in between the confidence intervals are considered as maintaining their same level of relative rater severity. Fewer than half of these raters were able to maintain the same level across performance dimensions.

FIGURE 12 ABOUT HERE.

Discussion

The results above have shown that a significant (p<0.001) effect of rater severity existed within each day, and in all days combined, for both performance dimensions.

Daily measures of relative rater severity were not invariant when compared with gross measures of relative rater severity. The days toward the start of the rating program, and the first day after the weekend, tended to show the most disagreement with gross measures.

When these raters rescored the same performances eight days later, there was a significant difference in their level of severity, for both performance dimensions, for more than half of the raters measured. The direction of this change across performance dimensions was not the same for all raters.

These raters, in general, did not maintain the same level of relative severity across performance dimensions.

These results clearly indicate that rater severity is not invariant over the term day. It has been shown that changes in rater severity can produce differences of more than half a score point on a six-point scale. The impact of these differences should be considered together with the fact that approximately 80 percent of the data were scored at the two middle score categories for both performance dimensions. Under these circumstances a small change in rater severity can have a major effect on an individual's relative position in the distribution of person abilities.

These findings have particular importance for large-scale rating programs such as the one described, which typically extend over a week or more. They suggest that constant monitoring of rater stability is desirable, and, in the case of high-stakes assessment, critical. While table leaders play an important role in helping to monitor ratings, their input is clearly insufficient to ensure stability between one rater and the next, and between one occasion and the next with the score criteria used in this assessment program. This conclusion may have been different if a finer-grained rating scale had been used. Monitoring via multifaceted Rasch analysis allows rapid feedback to raters. Adjustments can be made for instability that persists despite feedback to raters, and for degrees of instability which would be too small for raters to correct, but large enough to have an impact on candidates' futures.

This study confirms the warning given by Lumley and McNamara (1995) about the danger of certifying raters on the basis of a once-only calibration. Furthermore, adjustment of ratings on the basis of rater measures from precalibrated rating banks, or any single on-average measure, may be as inaccurate as unadjusted measures.

REFERENCES

Barritt, L., Stock, P. L., & Clark, F. (1986). Researching practice: evaluating student essays. College Composition and Communication, 37, 315-327.

Buckner, D. N. (1959). The predictability of ratings as a function of interrater agreement. Journal of Applied Psychology, 43(1), 60-64.

Cantor, N. K., & Hoover, H. D. (1986, April). The reliability and validity of writing assessment: an investigation of rater, prompt within mode, and prompt between mode sources of error. Paper presented at the paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Coffman, W. E., & Kurfman, D. (1968). A comparison of two methods of reading essay examinations. American Education Research Journal, 5(1), 101-120.

Diederich, P. B., French, J. W., & Carlton, S. T. (1961). Factors in judgments of writing ability. (Research Bulletin 61-15) . Princeton, NJ: Educational Testing Service.

Dunbar, S. B., Koretz, D. M., & Hoover, H. D. (1991). Quality control in the development and use of performance assessments. Applied Measurement in Education, 4(4), 289-303.

Engelhard, G., Jr. (1992). The measurement of writing ability with a many-faceted Rasch model. Applied Measurement in Education, 5(3), 171-191.

Engelhard, G., Jr, Gordon, B., & Gabrielson, S. (1991). The influences of mode of discourse, experiential demand, and gender on the quality of student writing. Research into the Teaching of English, 26(3), 315-336.

Gabrielson, S., Gordon, B., & Engelhard, G., Jr. (1995). The effects of task choice on the quality of writing obtained in a statewide assessment. Applied Measurement in Education, 8(4), 273-290.

Guilford, J. P. (1954). Psychometric Methods. (2nd ed.): McGraw-Hill.

Hake, R. (1986). How do we judge what they write? In K. L. Greenberg, H. S. Wiener, & R. A. Donovan (Eds.), Writing Assessment: Issues, Strengths and Strategies, (pp. 153-167). New York: Longman.

Huddleston, E. M. (1954). Measurement of writing ability at the college-entrance level: objective vs subjective testing techniques. Journal of Experimental Education, 22(3), 165-213.

Linacre, J. M. (1991, April). Constructing measurement with a many-facet Rasch model. Paper presented at the Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Lumley, T., Lynch, B. K., & McNamara, T. F. (1994). A new approach to standard-setting in language assessment. Language Testing, 3(2), 19-40.

Lumley, T., & McNamara, T. F. (1995). Rater characteristics and rater bias: implications for training. Language Testing, 12(1), 54-71.

Lunz, M. E., & Stahl, J. A. (1990). Judge consistency and severity across grading periods. Evaluation and the Health Professions, 13(4), 425-444.

Lunz, M. E., Stahl, J. A., & Wright, B. D. (1994). Interjudge reliability and decision reproducibility. Educational and Psychological Measurement, 54(4), 913-925.

Lunz, M. E., Stahl, J. A., & Wright, B. D. (1996). The invariance of judge severity calibrations. In M. R. Wilson & G. E. Jr (Eds.), Objective Measurement Theory into Practice, (Vol. 3, pp. 99-112). Norwood, NJ: Ablex.

McNamara, T. F. (1996). Measuring Second Language Performance. New York: Addison Wesley Longman Ltd.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. Educational Researcher, 23(2), 13-23.

Myford, C. M. (1991, April). Judging acting ability: the transition from novice to expert. Paper presented at the Paper presented at the American Educational Research Association, Chicago IL.

Overall, J. E., & Magee, K. N. (1992). Estimating individual rater reliabilities. Applied Psychological Measurement, 16(1), 77-85.

Ruth, L., & Murphy, S. (1988). Designing Writing Tasks for the Assessment of Writing. Norwood, NJ: Ablex Publishing.

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.

Thurstone, L. L. (1927). A law of comparative judgment. Psychological Review, 34, 273-286.

Webb, L. C., Raymond, M. R., & Houston, W. M. (1990, April). Rater stringency and consistency in performance assessment. Paper presented at the annual meeting of the American Educational Research Association and the National Council on Measurement in Education, Boston CA.

Weigle, S. C. (1994, February). Using FACETS to model rater training effects. Draft. Paper presented at the Language Testing Research Colloquium, Washington DC.

Wigglesworth, G. (1994). Patterns of oral behaviour in the assessment of an oral interaction test. Australian Review of Applied Linguistics, 17(2), 77-103.

Wright, B. D., & Masters, G. N. (1982). Rating Scale Analysis. Chicago IL: MESA Press.

Wu, M., Adams, R. J., & Wilson, M. (1996). ConQuest: Generalised Item Response Modelling Software. Melbourne: Australian Council for Educational Research.

Wu, M (1997). The Development and Application of a Fit Test for Use with Marginal Maximum Likelihood Estimation and Generalised Item Response Models. Unpublished Master's thesis. University of Melbourne, Australia.

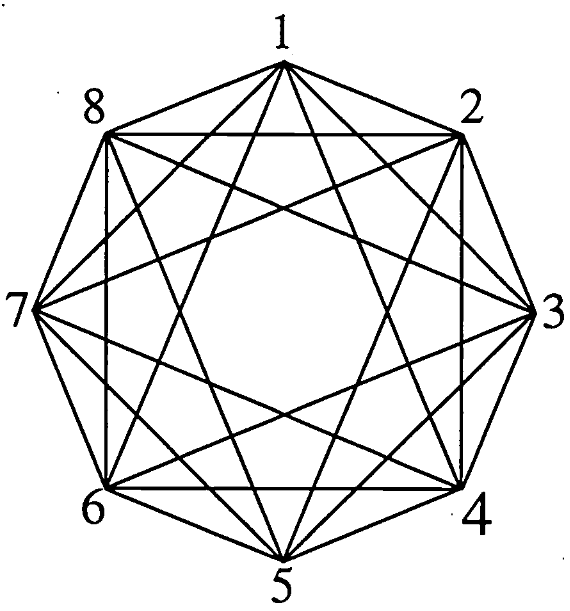

Figure 1: Design of rater linkage

Table 1. Rotation plan.

Bundle	Rater							
	1	2	3	4	5	6	7	8
A	1st	4th						
B		1st	4th					
C			1st	4th				
D				1st	4th			
E					1st	4th		
F						1st	4th	
G							1st	4th
H	4th							1st
I	2nd		5th					
J		2nd		5th				
K			2nd		5th			
L				2nd		5th		
M					2nd		5th	
N						2nd		5th
O	5th						2nd	
P		5th						2nd
Q	3rd			6th				
R		3rd			6th			
S			3rd			6th		
T				3rd			6th	
U					3rd			6th
V	6th					3rd		
W		6th					3rd	
X			6th					3rd

Table 2. Rater Weighted Mean Square Residuals, Overall Performance Dimension (minimum and maximum values within each calibration underlined).

Rater	Day 1	Day 2	Day 3	Day 4	Day 5	Day 8	Gross
1	0.91	1.12	0.80	0.80	0.81	0.69	0.89
2	0.82	0.91	0.81	0.97	0.69	0.85	0.89
3	1.62	0.93	0.90	0.79	0.87	1.00	0.96
4	1.28	1.45	1.54	1.83	1.90	1.68	1.44
5	0.92	1.12	1.11	0.95	1.07	0.96	1.05
6	1.17	1.17	1.05	1.21	1.18	1.26	1.17
7	0.94	0.83	0.97	0.80	0.73	0.97	0.85
8	0.97	1.14	1.10	1.01	1.07	0.97	0.99
9	0.94	1.05	0.82	0.84	1.10	1.32	1.02
10	0.75	0.93	0.81	0.82	0.77	0.95	0.81
11	1.08	1.24	1.46	0.92	1.07	0.82	1.15
12	0.98	0.68	1.03	0.89	0.69	0.87	0.89
13	0.90	0.83	0.84	0.85	0.71	0.80	0.87
14	0.98	0.70	0.96	1.22	0.78	1.05	0.96
15	0.76	0.87	0.68	1.05	0.94	0.90	0.91
16	0.93	0.86	1.12	0.90	----	0.73	0.94
Average	1.00	0.99	1.00	0.99	0.96	0.99	0.99

Table 3. Rater Weighted Mean Square Residuals, Textual Features Dimension (minimum and maximum values within each calibration underlined).

Rater	Day 1	Day 2	Day 3	Day 4	Day 5	Day 8	Gross
1	0.79	0.75	0.84	0.75	0.88	0.73	0.84
2	0.80	0.79	0.76	0.80	0.80	0.85	0.74
3	1.26	1.13	0.85	0.82	0.91	1.05	1.05
4	1.26	1.56	1.37	0.97	1.29	0.93	1.20
5	0.85	1.01	1.06	1.04	0.86	1.03	0.92
6	1.11	1.19	1.28	1.12	1.35	1.03	1.21
7	1.21	0.76	0.85	1.03	0.98	0.82	0.90
8	0.86	0.96	0.93	0.70	0.86	1.03	0.85
9	0.88	0.83	0.97	0.85	0.83	0.98	0.87
10	0.80	1.04	0.81	0.94	0.74	0.80	0.80
11	0.89	1.12	1.33	0.74	0.84	0.64	0.95
12	0.96	0.82	0.82	0.90	0.72	0.74	0.93
13	1.01	0.95	0.88	0.95	0.83	0.96	0.85
14	1.14	0.84	0.96	1.37	0.79	0.94	0.99
15	0.83	0.96	0.73	0.89	0.86	1.20	0.91
16	1.07	1.16	1.26	0.82	----	0.80	0.97
Average	0.98	0.99	0.98	0.92	0.90	0.91	0.94

Table 4. Rater Separation Reliability, Overall Performance Dimension.

Day 1	Day 2	Day 3	Day 4	Day 5	Day 8	Gross
0.925	0.938	0.973	0.954	0.827	0.960	0.987

Table 5. Rater Separation Reliability, Textual Features Dimension.

Day 1	Day 2	Day 3	Day 4	Day 5	Day 8	Gross
0.934	0.948	0.967	0.969	0.949	0.963	0.989

Table 6. Probability of Score for an Average Performance, OP Dimension.

		Probability of Score Category						
Rater	Severity (logits)	1	2	3	4	5	Expected score	Most probable score
1	0.28	0.07	0.66	0.26	0.01	0.00	2.21	2
2	0.74	0.11	0.71	0.18	0.00	0.00	2.07	2
3	0.07	0.05	0.63	0.31	0.01	0.00	2.28	2
4	-0.04	0.04	0.61	0.33	0.02	0.00	2.32	2
5	-1.27	0.01	0.31	0.58	0.10	0.00	2.78	3
6	0.71	0.11	0.70	0.18	0.00	0.00	2.08	2
7	0.20	0.06	0.65	0.28	0.01	0.00	2.24	2
8	0.44	0.08	0.68	0.23	0.01	0.00	2.16	2
9	-0.08	0.04	0.60	0.34	0.02	0.00	2.33	2
10	-0.24	0.03	0.57	0.38	0.02	0.00	2.39	2
11	-0.13	0.04	0.59	0.35	0.02	0.00	2.35	2
12	-0.32	0.03	0.55	0.40	0.03	0.00	2.42	2
13	0.11	0.05	0.64	0.30	0.01	0.00	2.27	2
14	-0.20	0.04	0.57	0.37	0.02	0.00	2.37	2
15	0.02	0.05	0.62	0.32	0.01	0.00	2.30	2
16	-0.28	0.03	0.56	0.38	0.02	0.00	2.40	2
Average	0.00	0.06	0.64	0.29	0.01	0.00	2.26	2

Table 7. Probability of Score for an Average Performance, TF Dimension.

		Probability of Score Category						
Rater	Severity	1	2	3	4	5	Expected	Most probable
	(logits)						score	score
1	0.01	0.04	0.66	0.29	0.01	0.00	2.27	2
2	1.01	0.13	0.75	0.12	0.00	0.00	1.99	2
3	0.12	0.05	0.63	0.31	0.01	0.00	2.28	2
4	-0.12	0.04	0.63	0.31	0.01	0.00	2.31	2
5	-1.04	0.01	0.41	0.52	0.06	0.00	2.64	3
6	0.33	0.06	0.71	0.22	0.01	0.00	2.18	2
7	-0.41	0.02	0.57	0.38	0.02	0.00	2.41	2
8	0.45	0.07	0.72	0.20	0.01	0.00	2.14	2
9	0.26	0.06	0.70	0.24	0.01	0.00	2.20	2
10	0.28	0.06	0.70	0.23	0.01	0.00	2.19	2
11	-0.25	0.03	0.61	0.35	0.02	0.00	2.35	2
12	-0.74	0.01	0.49	0.46	0.04	0.00	2.52	2
13	-0.52	0.02	0.55	0.40	0.03	0.00	2.44	2
14	0.24	0.06	0.70	0.24	0.01	0.00	2.20	2
15	0.77	0.10	0.74	0.15	0.00	0.00	2.06	2
16	-0.40	0.02	0.57	0.38	0.02	0.00	2.40	2
Average	0.00	0.06	0.64	0.29	0.01	0.00	2.26	2

Table 8. Relative Rater Severity Measures, Overall Performance Dimension.

	Rater Severity Measures (logits)								Percent of
Rater	Day 1	Day 2	Day 3	Day 4	Day 5	Day 8	Gross	Range	Person S.D.
1	0.40	-0.20	0.35	0.27	0.22	0.62	0.28	0.82	43%
2	0.43	1.09	0.57	0.82	0.74	0.91	0.74	0.66	35%
3	0.03	0.19	0.13	-0.07	0.08	-0.23	0.07	0.41	22%
4	-0.15	0.02	0.99	0.00	0.20	-1.02	-0.04	2.01	106%
5	-0.63	-0.78	-2.22	-1.80	-1.12	-1.35	-1.27	1.59	84%
6	0.24	0.77	1.08	0.90	0.38	1.01	0.71	0.84	44%
7	-0.17	0.55	0.48	0.08	0.04	0.22	0.20	0.72	38%
8	0.80	-0.27	0.77	0.80	0.11	0.38	0.44	1.07	56%
9	-0.02	-0.48	-0.40	-0.09	-0.22	0.65	-0.09	1.12	59%
10	0.06	-0.30	-0.73	-0.14	-0.43	-0.18	-0.24	0.79	41%
11	0.46	0.19	-0.40	-0.13	-0.31	-0.70	-0.13	1.16	61%
12	-0.59	0.02	-0.55	-0.39	0.02	-0.36	-0.32	0.61	32%
13	-0.04	-0.53	0.22	0.08	0.17	0.84	0.11	1.36	72%
14	-0.87	0.43	0.01	-0.35	-0.18	-0.16	-0.20	1.29	68%
15	0.25	-0.27	0.08	0.34	0.28	-0.55	0.02	0.89	47%
16	-0.21	-0.43	-0.40	-0.33	----	-0.06	-0.28	0.38	20%
Average	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.98	52%
SD	0.44	0.51	0.80	0.63	0.43	0.70	0.47	0.43	23%

Minimum and maximum values for each rater have been underlined.

Table 9. Relative Rater Severity Measures, Textual Features Dimension.

	Rater Severity Measures (logits)								Percent of
Rater	Day 1	Day 2	Day 3	Day 4	Day 5	Day 8	Overall	Range	Person S.D.
1	0.23	-0.61	-0.14	0.30	0.15	0.11	0.01	0.91	42%
2	0.64	1.28	1.10	0.96	0.83	1.16	1.01	0.64	29%
3	0.08	0.06	0.25	-0.01	0.40	-0.05	0.12	0.45	21%
4	-0.04	-0.70	0.27	-0.23	0.18	-0.30	-0.12	0.97	45%
5	-0.28	-0.14	-1.55	-1.70	-2.10	-1.27	-1.04	1.96	90%
6	-0.09	-0.30	-0.11	0.89	0.76	1.15	0.33	1.44	66%
7	-0.35	-0.34	-0.13	-0.75	-0.36	-0.49	-0.41	0.62	28%
8	0.51	0.25	0.53	0.58	0.32	0.31	0.45	0.33	15%
9	-0.20	-0.37	0.11	0.62	0.44	1.13	0.26	1.50	69%
10	0.81	0.90	-0.05	0.16	-0.29	0.09	0.28	1.19	55%
11	0.52	0.12	-0.59	-0.08	-0.52	-1.03	-0.25	1.56	72%
12	-0.86	-0.16	-0.95	-1.03	-0.37	-0.97	-0.74	0.87	40%
13	-0.40	-0.82	-0.54	-0.52	-0.86	-0.16	-0.52	0.71	32%
14	-0.55	0.82	0.56	0.28	0.36	0.26	0.24	1.37	63%
15	0.57	0.17	1.57	1.09	1.06	0.45	0.77	1.40	64%
16	-0.59	-0.16	-0.34	-0.56	----	-0.39	-0.40	0.42	20%
Average	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.02	47%
SD	0.50	0.59	0.75	0.77	0.79	0.75	0.54	0.48	22%

Minimum and maximum values for each rater have been underlined.

Table 10. Performances Used in each Calibration.

Occasion	Day 1	Day 2	Day 3	Day 4	Day 5	Day 8	Gross
Number	1293	1491	1560	1446	1044	1451	8285
Percent	15.6%	17.9%	18.8%	17.4%	12.6%	17.5%	100%

Table 11. Standardised Differences between Gross and Daily Measures of Relative Rater Severity (OP).

Rater	Day 1	Day 2	Day 3	Day 4	Day 5	Day 8
1	0.95	-2.97	0.54	-0.05	-0.28	2.23
2	-2.34	2.73	-1.29	0.53	0.13	1.14
3	-0.23	0.87	0.41	-0.93	0.23	-1.93
4	-0.78	0.33	6.31	0.28	1.03	-6.50
5	5.18	3.90	-7.67	-4.05	0.88	-0.58
6	-3.63	0.44	2.61	1.34	-1.75	1.94
7	-2.65	2.39	1.77	-0.70	-0.61	0.10
8	2.68	-5.43	2.33	2.38	-2.15	-0.45
9	0.55	-2.98	-2.02	0.00	-0.80	4.68
10	2.43	-0.37	-3.83	0.69	-1.17	0.42
11	4.86	2.87	-2.25	-0.01	-1.24	-4.43
12	-1.95	2.51	-1.79	-0.49	2.04	-0.23
13	-1.19	-5.27	0.80	-0.25	0.45	5.28
14	-5.79	5.33	1.56	-1.09	0.27	0.35
15	1.52	-2.01	0.36	1.92	1.07	-3.30
16	0.54	-1.27	-0.74	-0.41	----	1.54
chi-square	132.5	153.1	147.0	31.4	19.1	139.7
p value	0.000	0.000	0.000	0.008	0.160	0.000

Standardised difference values greater than ±2 indicate significance at the 0.05 level.

Table 12. Standardised Differences Between Gross and Daily Measures of Relative Rater Severity (TF).

Rater	Day 1	Day 2	Day 3	Day 4	Day 5	Day 8
1	1.60	-3.65	-1.05	1.90	1.07	0.65
2	-2.67	2.01	0.64	-0.32	-0.58	1.03
3	-0.26	-0.36	0.82	-0.87	1.95	-1.02
4	0.54	-2.93	2.35	-0.71	1.28	-1.16
5	5.62	6.64	-3.88	-4.68	-5.65	-1.49
6	-3.10	-4.43	-3.00	3.82	2.61	5.23
7	0.37	0.40	1.68	-2.12	0.35	-0.52
8	0.42	-1.43	0.56	0.88	-0.75	-0.91
9	-3.38	-4.49	-0.94	2.38	1.31	5.46
10	3.95	4.11	-2.51	-0.81	-3.75	-1.24
11	6.02	3.08	-2.63	1.17	-1.80	-5.65
12	-0.86	3.99	-1.64	-2.26	2.26	-1.46
13	0.89	-2.32	-0.06	0.03	-1.81	2.64
14	-6.43	4.63	2.21	0.28	0.90	0.19
15	-1.23	-3.78	4.92	1.97	1.22	-1.74
16	-1.42	1.84	0.41	-1.30	----	0.02
chi-square	161.3	199.0	80.9	65.2	76.0	109.9
p value	0.000	0.000	0.000	0.000	0.000	0.000

Standardised difference values greater than ±2 indicate significance at the 0.05 level.

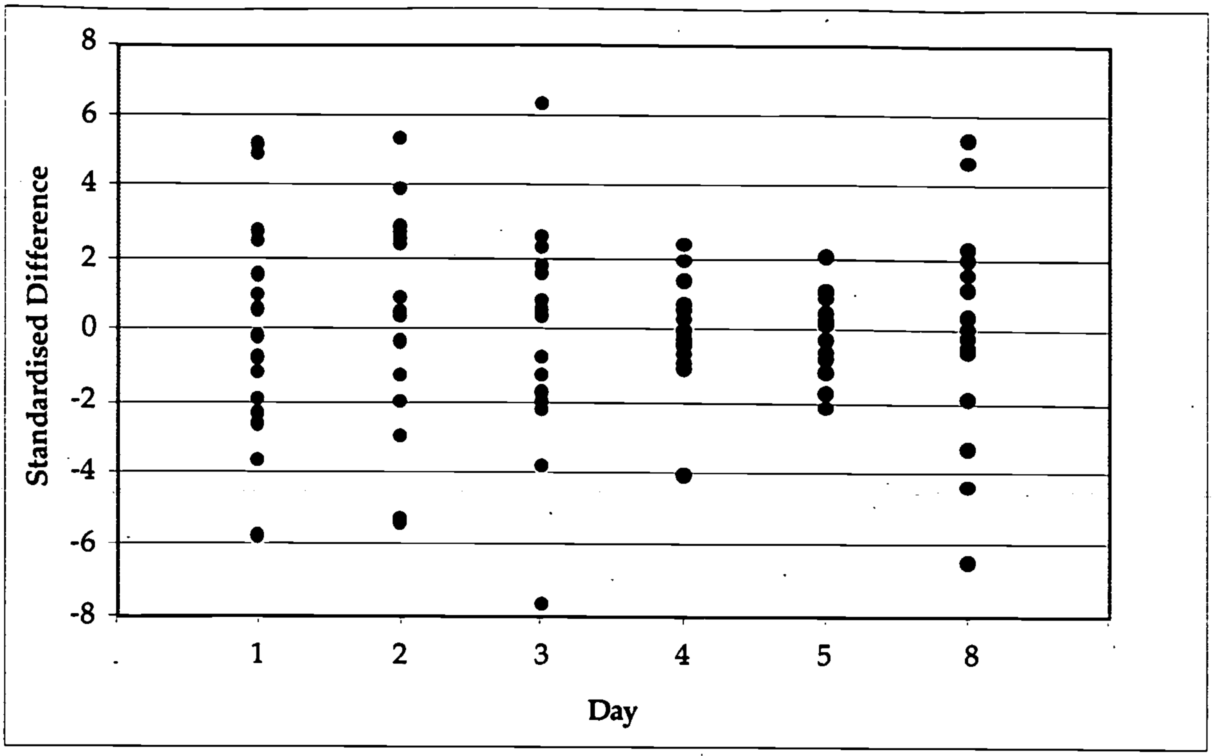

Figure 2: Differences in relative rater severity across occasions, Overall Performance dimension

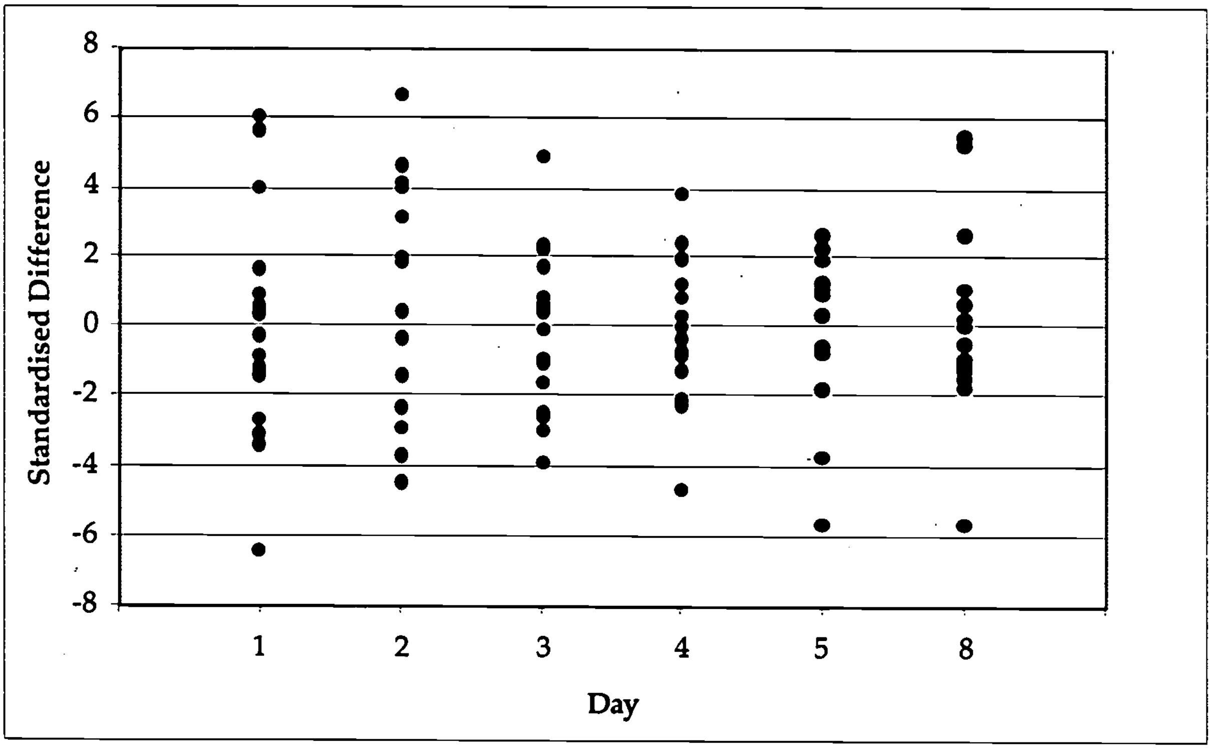

Figure 3: Differences in relative rater severity across occasions, Textual Features dimension

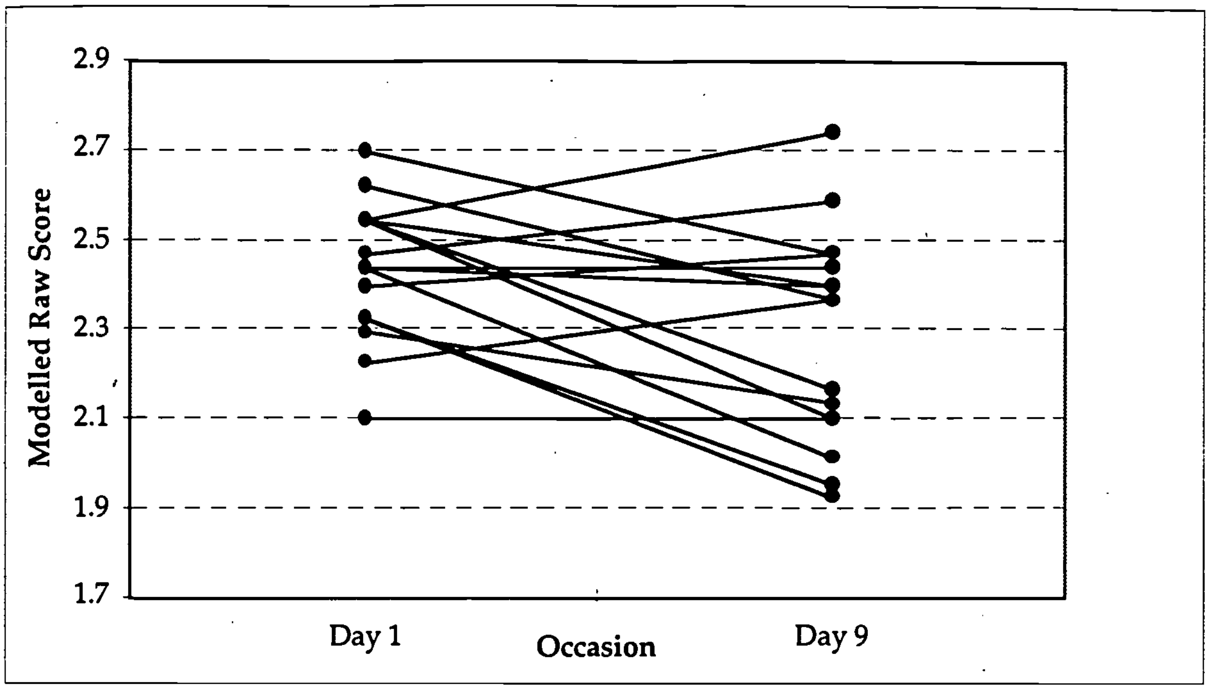

Figure 4: Expected average score for persons of average ability, Day 1 and Day 9, OP dimension

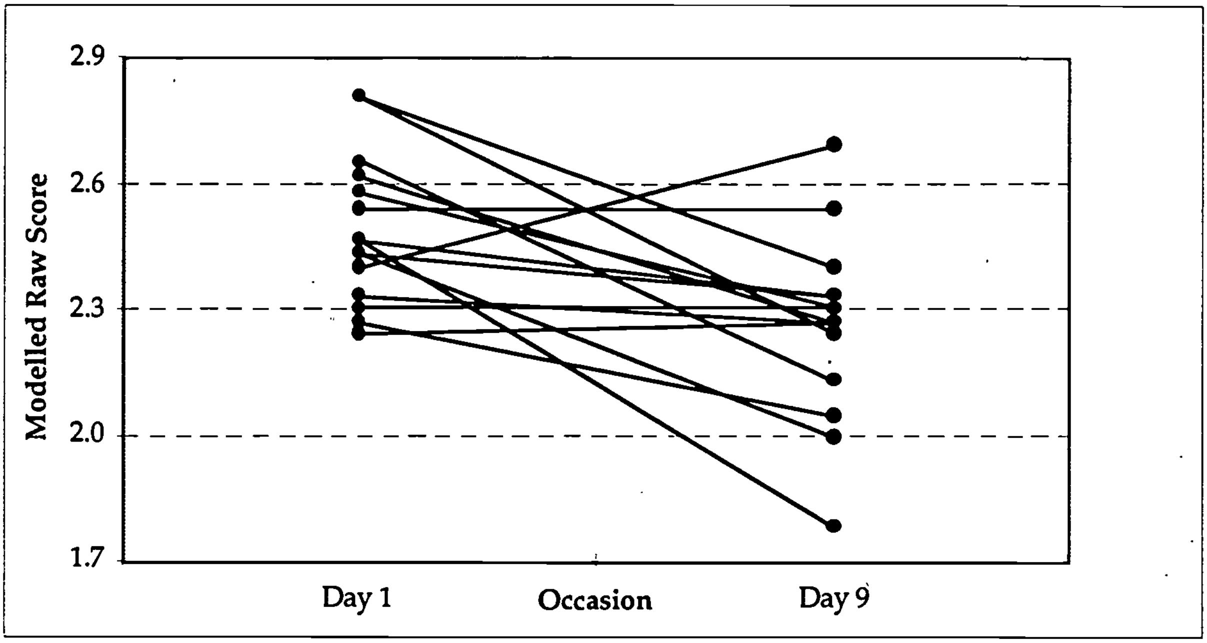

Figure 5: Expected average score for persons of average ability, Day 1 and Day 9, TF dimension

Table 13. Overall Performance, Day 1 and Day 9.

Rater	Severity (logits)		Standard Error		Difference	Standardised	chi-square	p value
	Day 1	Day 9	Day 1	Day 9		Difference		
1	0.45	0.45	0.20	0.20	0.00	0.00	0.00	1.00
2	-0.30	0.93	0.13	0.13	-1.23	6.77	45.87	0.00
3	-0.84	0.30	0.12	0.12	-1.14	6.49	42.11	0.00
4	-0.86	-1.37	0.13	0.13	0.51	-2.86	8.19	0.00
5	-0.64	-0.92	0.13	0.13	0.28	-1.56	2.44	0.12
6	-0.27	1.01	0.14	0.14	-1.28	6.34	40.19	0.00
7	-0.86	0.41	0.13	0.13	-1.27	7.18	51.61	0.00
8	-0.40	-0.68	0.14	0.14	0.28	-1.39	1.94	0.16
9	-0.57	0.78	0.14	0.14	-1.35	6.65	44.21	0.00
10	0.02	-0.35	0.17	0.17	0.37	-1.53	2.34	0.13
11	-0.16	0.34	0.14	0.14	-0.51	2.65	7.02	0.01
12	-0.56	-0.56	0.14	0.14	0.00	0.00	0.00	1.00
13	-0.89	-0.46	0.12	0.12	-0.43	2.49	6.20	0.01
14	-1.04	-0.33	0.13	0.13	-0.71	3.95	15.63	0.00
15	-1.23	-0.61	0.14	0.14	-0.62	3.14	9.88	0.00
16	-0.53	-0.49	0.13	0.13	-0.03	0.19	0.04	0.85
Average	-0.54	-0.10						
Difference	-0.45							

Rater	Severity (logits)		Standard Error		Difference	Standardised	chi-square	p value
	Day 1	Day 9	Day 1	Day 9		Difference		
1	0.45	0.45	0.20	0.20	0.00	0.00	0.00	1.00
2	-0.30	0.93	0.13	0.13	-1.23	6.77	45.87	0.00
3	-0.84	0.30	0.12	0.12	-1.14	6.49	42.11	0.00
4	-0.86	-1.37	0.13	0.13	0.51	-2.86	8.19	0.00
5	-0.64	-0.92	0.13	0.13	0.28	-1.56	2.44	0.12
6	-0.27	1.01	0.14	0.14	-1.28	6.34	40.19	0.00
7	-0.86	0.41	0.13	0.13	-1.27	7.18	51.61	0.00
8	-0.40	-0.68	0.14	0.14	0.28	-1.39	1.94	0.16
9	-0.57	0.78	0.14	0.14	-1.35	6.65	44.21	0.00
10	0.02	-0.35	0.17	0.17	0.37	-1.53	2.34	0.13
11	-0.16	0.34	0.14	0.14	-0.51	2.65	7.02	0.01
12	-0.56	-0.56	0.14	0.14	0.00	0.00	0.00	1.00
13	-0.89	-0.46	0.12	0.12	-0.43	2.49	6.20	0.01
14	-1.04	-0.33	0.13	0.13	-0.71	3.95	15.63	0.00
15	-1.23	-0.61	0.14	0.14	-0.62	3.14	9.88	0.00
16	-0.53	-0.49	0.13	0.13	-0.03	0.19	0.04	0.85
Average	-0.54	-0.10						
Difference	-0.45							

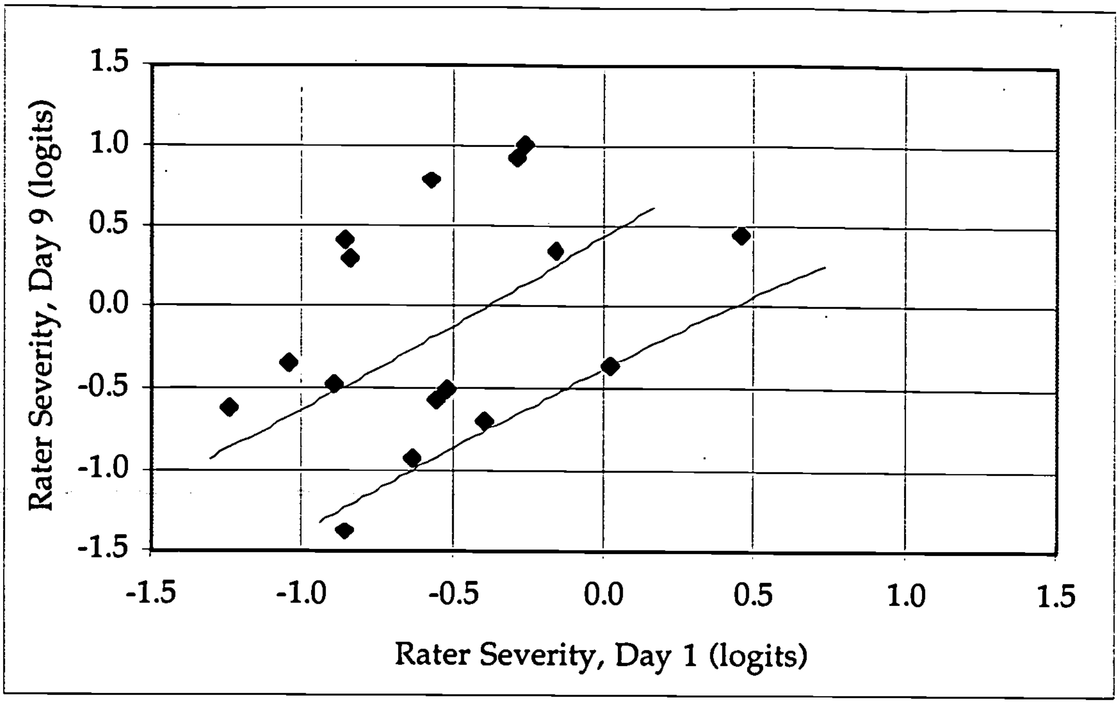

Figure 6: Change in rater severity for OP, Day 1 to Day 9

Table 14. Textual Features, Day 1 and Day 9.

Rater	Severity (logits)		Standard Error		Difference	Standardised	chi-square	p value
	Day 1	Day 9	Day 1	Day 9		Difference		
1	-0.20	-0.20	0.20	0.20	0.00	0.00	0.00	1.00
2	-0.01	0.73	0.13	0.13	-0.74	3.96	15.71	0.00
3	-0.65	-0.23	0.13	0.13	-0.42	2.30	5.27	0.02
4	-1.60	-0.47	0.13	0.13	-1.13	6.14	37.64	0.00
5	-0.66	1.76	0.13	0.13	-2.42	12.97	168.33	0.00
6	-1.12	0.49	0.14	0.14	-1.61	8.08	65.35	0.00
7	-1.01	-0.08	0.13	0.13	-0.93	5.13	26.28	0.00
8	-0.27	-0.09	0.15	0.15	-0.18	0.86	0.74	0.39
9	-0.50	0.95	0.15	0.15	-1.46	6.82	46.49	0.00
10	0.08	-0.05	0.17	0.17	0.13	-0.53	0.28	0.60
11	-0.56	-0.23	0.14	0.14	-0.33	1.66	2.75	0.10
12	-0.88	-0.88	0.14	0.14	0.00	0.00	0.00	1.00
13	-1.13	0.50	0.12	0.12	-1.63	9.36	87.59	0.00
14	-0.93	-0.18	0.13	0.13	-0.75	4.10	16.81	0.00
15	-0.41	-1.27	0.14	0.14	0.86	-4.36	19.04	0.00
16	-1.59	0.00	0.13	0.13	-1.60	8.82	77.73	0.00
Average	-0.72	0.05						
Difference	-0.76							

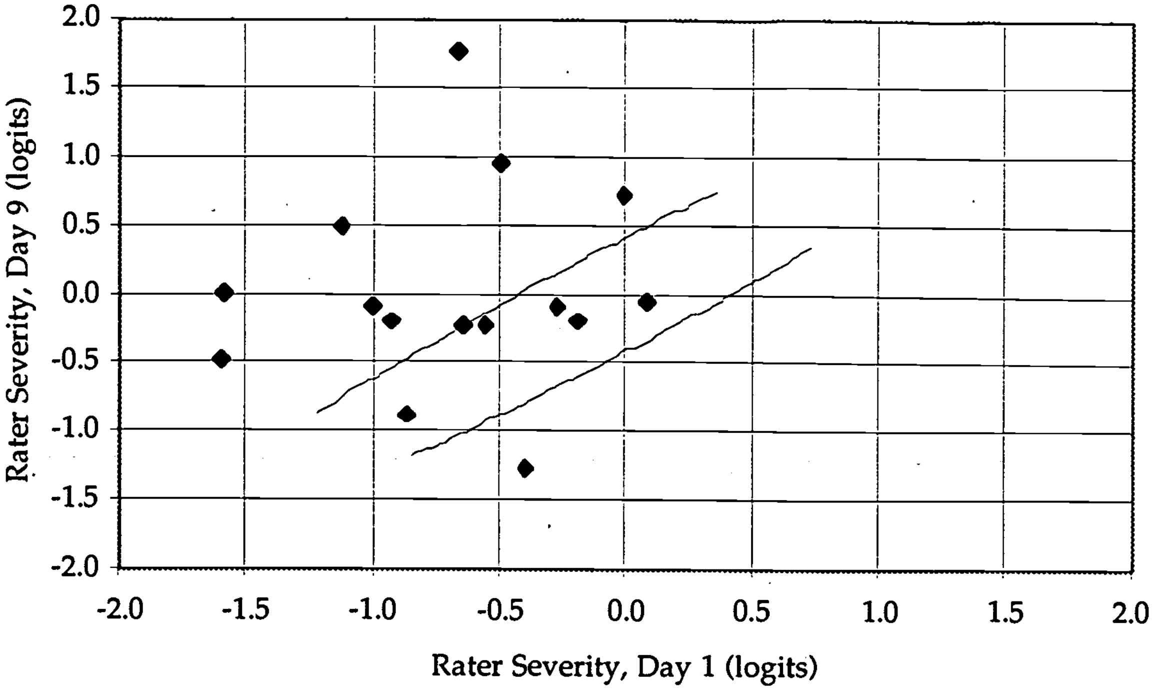

Figure 7: Change in rater severity for TF, Day 1 to Day 9

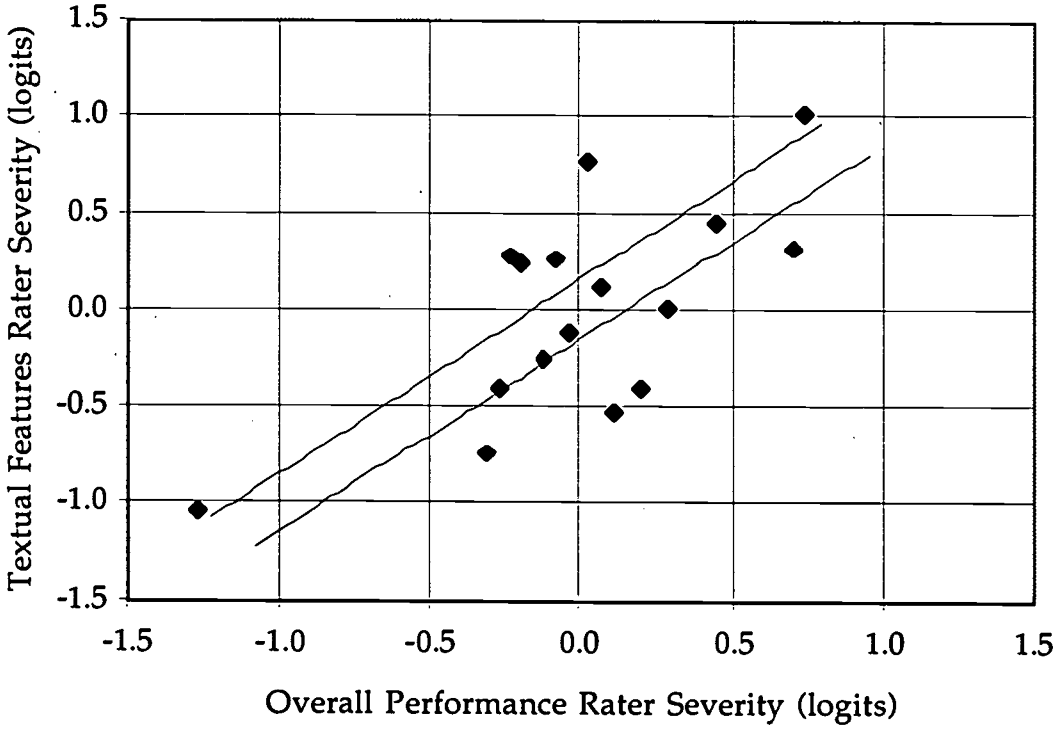

Figure 8: Relative gross rater severity across performance dimensions.

AREA 1997

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement (OERI)
Educational Resources Information Center (ERIC)

REPRODUCTION RELEASE

(Specific Document)

ERIC®

TM027338

I. DOCUMENT IDENTIFICATION:

Title: Rater Severity in Large-Scale Assessment: Is it Invariant?	
Author(s): MCQUEEN, JOY AND CONGDON, PETER J	
Corporate Source:	Publication Date:

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic/optical media, and sold through the ERIC Document Reproduction Service (EDRS) or other ERIC vendors. Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce the identified document, please CHECK ONE of the following options and sign the release below.

☑ **Sample sticker to be affixed to document**

Check here

Permitting microfiche (4"x 6" film), paper copy, electronic, and optical media reproduction

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Level 1

Sample sticker to be affixed to document □

or here

Permitting reproduction in other than paper copy.

"PERMISSION TO REPRODUCE THIS MATERIAL IN OTHER THAN PAPER COPY HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Level 2

Sign Here, Please

Documents will be processed as indicated provided reproduction quality permits. If permission to reproduce is granted, but neither box is checked, documents will be processed at Level 1.

"I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce this document as indicated above. Reproduction from the ERIC microfiche or electronic/optical media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries."

Signature:	Position: RESEARCH FELLOW
Printed Name: JOY MCQUEEN	Organization: AUSTRALIAN COUNCIL FOR EDUCATIONAL RESEARCH
Address: AUSTRALIAN COUNCIL FOR EDUCATIONAL RESEARCH. LOCKED BAG 55 CAMBERWELL 3124, AUSTRALIA	Telephone Number: (613) 9277 5582
	Date: 13/5/97

OVER

CUA

THE CATHOLIC UNIVERSITY OF AMERICA

Department of Education, O'Boyle Hall
Washington, DC 20064
202 319-5120

February 21, 1997

Dear AERA Presenter,

Congratulations on being a presenter at AERA[1]. The ERIC Clearinghouse on Assessment and Evaluation invites you to contribute to the ERIC database by providing us with a printed copy of your presentation.

Abstracts of papers accepted by ERIC appear in *Resources in Education (RIE)* and are announced to over 5,000 organizations. The inclusion of your work makes it readily available to other researchers, provides a permanent archive, and enhances the quality of *RIE*. Abstracts of your contribution will be accessible through the printed and electronic versions of *RIE*. The paper will be available through the microfiche collections that are housed at libraries around the world and through the ERIC Document Reproduction Service.

We are gathering all the papers from the AERA Conference. We will route your paper to the appropriate clearinghouse. You will be notified if your paper meets ERIC's criteria for inclusion in *RIE*: contribution to education, timeliness, relevance, methodology, effectiveness of presentation, and reproduction quality. You can track our processing of your paper at http://ericae2.educ.cua.edu.

Please sign the Reproduction Release Form on the back of this letter and include it with **two** copies of your paper. The Release Form gives ERIC permission to make and distribute copies of your paper. It does not preclude you from publishing your work. You can drop off the copies of your paper and Reproduction Release Form at the **ERIC booth (523)** or mail to our attention at the address below. Please feel free to copy the form for future or additional submissions.

Mail to:

AERA 1997/ERIC Acquisitions
The Catholic University of America
O'Boyle Hall, Room 210
Washington, DC 20064

This year ERIC/AE is making a **Searchable Conference Program** available on the AERA web page (http://aera.net). Check it out!

Sincerely,

Lawrence M. Rudner, Ph.D.
Director, ERIC/AE

[1] If you are an AERA chair or discussant, please save this form for future use.