ABSTRACT
                Two factors which may affect the ratings assigned to
an essay test are investigated: (1) context effects; and (2) score
level effects. Context effects exist in essay scoring if an essay is
rated higher when preceded by poor quality essays than when preceded
by high quality essays. A score level effect is defined as a change
in the score (value) assigned to the second reading of an essay when
compared to the first reading, where the change is a function of the
range in which a score may increase or decrease. The potential dual
effect of both these factors is theorized in a formula called the
essay score change. Examples of the possible utility of this index
are outlined. Tentative hypotheses for investigating and interpreting
possible essay score change in light of potential dual effects of
context and score level are discussed. (Author/GDC)

ED280852

RESEARCH REPORT

THE POTENTIAL DUAL EFFECT OF CONTEXT EFFECTS AND SCORE LEVEL EFFECTS ON THE ASSIGNMENT OF SCORES TO ESSAYS

Patricia A. Paden

Educational Testing Service
Princeton, New Jersey
October 1986

THE POTENTIAL DUAL EFFECT OF CONTEXT EFFECTS AND SCORE LEVEL EFFECTS
ON THE ASSIGNMENT OF SCORES TO ESSAYS

Patricia A. Paden

Educational Testing Service

3

4

THE POTENTIAL DUAL EFFECT OF CONTEXT EFFECTS AND SCORE LEVEL EFFECTS ON THE ASSIGNMENT OF SCALES TO ESSAYS

ABSTRACT

This paper represents a systematic treatment of the potential dual effect of the context in which an essay is reread and the previously assigned score (value) of that essay on the subsequently assigned essay score. This effect is theorized in a formula referred to as "essay score change" (ESC). Examples of the possible utility of the ESC index are outlined. Tentative hypotheses for investigating and interpreting possible essay score change in light of potential dual effects of context and score level are discussed.

THE POTENTIAL DUAL EFFECT OF CONTEXT EFFECTS AND SCORE LEVEL EFFECTS ON THE ASSIGNMENT OF SCORES TO ESSAYS

Patricia A. Paden[1,2]

INTRODUCTION

Obtaining a reliable measure of a student's ability to write is a very important factor in the assessment of essay writing skills. The score level (within a score scale) serves as a measure of the judged quality of a student's ability to perform a writing task. In this measurement process, many technical problems arise. The purpose of this paper is to show that certain of these problems may be due to the relationship between the context in which an essay is read and the characteristics of the score levels within a score scale. This relationship will be illustrated by defining and elaborating upon context effects and score level effects.

One approach to assessing the reliability of scores awarded to essays is to conduct double-readings. In this process a first-reading essay score is compared to a second-reading essay score. It is often found that essays are not awarded the same score on both readings. Research findings indicate that some differential awarding of scores to (reread) essays may be attributed to context (or contrast) effects (Hales and Tokar, 1975; Hughes, Keeling and Tuck, 1980a; 1980b). These effects exist in essay scoring if essays are rated higher when preceded by poor quality essays than when preceded by high quality essays. Context effects are a potential source of

6

reader inconsistency, since the context for any two independent readers is likely to be different.

In addition to context effects, a score level effect could exist for scores obtained from reread essays. In this paper, a score level effect is defined as a change in the score (value) assigned to the second reading of an essay when compared to the first reading, where the change is a function of the range in which a score may increase or decrease. The amount that a reread essay score can change (decrease or increase) is related to its relative position within a score scale and its possible range of increase or decrease. This is a well known consequence of floor and ceiling effects associated with a given score level within a score scale. However, there is little or no available systematic treatment of the extent to which context effects can be related to score level effects which operate on score scales. A systematic treatment will be attempted by (i) reviewing and summarizing research findings on context effects in essay scoring and (ii) relating those findings to possible score level effects within a score scale.

REVIEW OF THE LITERATURE

In reviewing the literature on context effects, one encounters research efforts that attempt to find ways to reduce or eliminate these effects in essay scoring. Hughes, Keeling and Tuck (1980b) predicted that the concentrated effort needed to define the several judgments to be made in analytic scoring should reduce the effects of context in the reading. Upon investigating this prediction, they found that analytic score procedures, in which scorers were provided with guidelines regarding the weighting to be awarded for particular essay features such as writing style, originality of ideas,

7

grammar and so on, were as susceptible to context effects as holistic scoring procedures.

Daly and Dickson-Markman (1982) reviewed the Hughes et al (1980b) study and found the conclusions drawn to be limited by the absence of adequate comparison control groups. In the study by Hughes and associates, the only rating of the criterion essay was obtained after subjects had read an experimental block of either good or bad essays. Daly and Dickson-Markman (1982) contend that finding a difference between these two experimental conditions does not demonstrate a meaningful effect if considered in the absence of two critical control conditions. The first essential control is the rating of the criterion essay by itself, unaffected by other papers. The second necessary control is a rating of the criterion essay following a block of papers of variable quality. Daly and Dickson-Markman further state:

> The first control provides an index of the value of the essay judged without comparison to other essays. The second control provides an index of the essay's value as it is comparatively judged in light of other papers but where order and quality is not intentionally biased in a positive or negative fashion. This control approximates a normal judgment situation. The two experimental conditions (criterion paper preceded by a block of good or bad papers) must not only be significantly different from one another but also significantly different from the two control conditions to clearly demonstrate a contrast effect... Finding a difference in rating for the criterion essay in the two experimental groups will replicate earlier findings. Finding differences between the experimental groups and conditions will clarify the nature of the contrast effect in essay evaluation. (p.310).

Conducting an experiment using the above controls, Daly and Dickson-Markman (1982) found that the results for their experimental groups replicated earlier findings of a significant difference between ratings of a

8

cri erion essay as a function of previously read papers. That is, when a
middle-ranked essay is read after a series of high quality essays, it is
rated lower than when it is preceded by a group of low quality ones. Daly
and Dickson-Markman place this occurrence within Helson's adaptation level
theory (see Daly and Dickson-Markman, 1982, p.313). According to Daly and
Dickson-Markman: "The theory suggests that people form standards or norms
for judging stimuli on the basis of their experience with whatever stimuli
of the type they have been exposed to. When a person encounters a
particular stimulus significantly different from the established norm he or
she adjusts or contrasts the new stimulus to a more extreme position than is
warranted by the object's true value."

Daly and Dickson-Markman (1982) also compared the two experimental
groups to the two control groups in their study. They found that there was
virtually no difference between the score value of the essay when rated by
itself and when rated after four high quality pieces. This was interpreted
to mean that, when evaluating an essay preceded by a series of high quality
essays, judges (teachers) did not evaluate the essay less positively than
they did when it was presented without any prior essays. On the other hand,
criterion papers were rated higher when judges read a random series of
varied quality papers prior to reading the criterion piece than when the
criterion piece was their first rating task. Daly and Dickson-Markman
(1982) note that the mean value for the criterion essay in this condition
(second control) was closer to the hypothetical midpoint of the rating
scales than ratings given under other conditions. Thus, in the random
condition, there was a tendency for judges to move toward an average (or
neutral evaluation). We note that this may be especially true with a scale

9

with an odd number of score levels. The sample sizes for the experimental
reading: ranged from 36 to 47 essays.

Hughes et al (1983) and Hughes and Keeling (1984) made additional
attempts to reduce context effects. Hughes et al (1983) sought to eliminate
context effects by giving scorers explicit warning about their influence and
also by requiring scorers explicit warning about their influence and also by
requiring scorers initially to sort essays into a few qualitative categories
before rereading them and awarding final grades. The results of these
procedures were compared with those obtained by scorers who were merely
warned of the existence of context effects and those obtained by scorers who
were given no information about the influence of context. Results showed
that all three groups were influenced by context and to about the same
degree.

Hughes and Keeling (1984) investigated the effectiveness of providing
scorers with model essays to reduce the influence of context effects.
Context effects persisted despite the use of model essays during scoring.
Hughes and Keeling (1984) conclude that "we may be forced to accept context
effects as an unavoidable concomitant of essay scoring" (p.281).

DISCUSSION: CONTEXT EFFECTS AND SCORE LEVEL EFFECTS

There is a common observation underlying the studies of context effects
on individual essays that have been reread by different readers. This
observation is the change in score level that can occur for an essay whose
first-reading score is a middle level on a score scale. The research
indicates that the general occurrence is such that a middle-ranked
(criterion) essay is perceived to be of higher score value on the score

scale when reread after a block of essays judged to be of poor quality than

when reread after a block of essays judged to be of good quality. Since the

score value of a middle-ranked criterion essay tends to change with a given

context, there is a relationship between the judgment of the quality of an

essay in a given context and the (potential) change in score points of an

average essay rated on a given scale score. In other words, the context

affects the grading behavior and grading behavior, in turn, determines the

score level assigned to the essay.

This phenomenon is described in the following analysis. Consider the

seven-point score scale such as that used by Daly and Dickson-Markman. The

possible relative range of increase (R_i+) and the possible relative range of

decrease (R_i-) for each score level are defined as follows:

```
Score
Scale    |---+---+---+---+---+---+---|
          1   2   3   4   5   6   7

R_i+    .86  .71  .57  .43  .29  .14  .00
R_i-    .00  .15  .29  .43  .57  .72  .86
```

where R_i+ is calculated as $\frac{7-1}{7}$, $\frac{7-2}{7}$, . . ., $\frac{7-7}{7}$ and R_i- is $(.86-R_i+)$.

Observe that a score of 7 has a .00 possible relative range of increase

while a score of 1 has a .86 possible relative range of increase. This same

type of comparison can be made at the other s re levels. The midpoint of

the above score scale is 4. At this midpoint, R_i+ and R_i- are identical.

Some characteristics of context effects and score level effects within a

score scale can be illustrated by employing the score scale model defined

above. For example, an essay that is assigned a middle rank (score at the

midpoint) has a 1:1 chance to be assigned a lower or higher score when

reread. Such is the case with Daly and Dickson-Markman's (1982) use of the middle-ranked essay as the criterion in the study of context effects. This is tantamount to creating a condition in which there are no score level effects. That is, the mid-score is not differentially restricted in range in either direction on the score scale due to the floor and ceiling effects of the scale. Thus, this is the only condition where we can investigate context effects as a factor of change independent of a potential score level effect. It will be shown that this condition is necessary in the development of a model that represents the relationship between context effects and score level effects. Here we note that the effec$^{..}$ scorers tending toward the middle score would not be an issue in this analysis because that effect would be virtually synonymous to using the middle-ranked essay as the criterion. The following analyses establish a procedures to interpret context effects and score level effects in terms of the metric (R_i+) established in the score scale model above.

If we let the possible relative range of increase or decrease $(R_i+$ or $R_i-)$ at the midpoint serve as reference point $(R_i+ = R_i- = .43)$, then we can consider departures from this point as measures of possible change in score level for reread essays. The implication for possible changes in score level can be summarized by relating differences in the essay means obtained from the following rereading conditions reported in Daly and Dickson-Markman's (1982) study:

1. A criterion essay preceded by four high quality essays -

    $\overline{X} = 3.46$

2. A criterion essay preceded by four low quality essays -

    $\overline{X} = 4.74$

3. A criterion essay read first or alone   -

    $\overline{X} = 3.47$

4. A criterion essay read after a random pattern   -

    $\overline{X} = 4.14$

We can calculate R_i+'s for Daly and Dickson-Markman (1982) contextual

means by using the R_i+ previously defined and interpolating. This is shown

in Table 1, where

> HHHHC = the rereading of the criterion middle-ranked essay (C) after
>   high ranked essays
>
> LLLLC = the rereading of th. criterion essay after low ranked essays
>
> C First = the rereading of the criterion essay first
>
> Random C = the rereading of the criterion essay after a selection of essays
>   of varied ranks.

13

Table 1. Contextual Score Level Changes (R_1+) for Reread Criterion Essays

Context Condition j	Score Level	R_1+	Absolute Percentage Change in R_1+ for Context j	Direction of Change in C
1. HHHHC	4.00 3.46 3.00	.43 X = .51 .57	$\frac{.08}{.43}$ = .186	↓ (Decrease)
2. LLLLC	5.00 4.74 4.00	.29 X = .33 .43	$\frac{.10}{.43}$ = .233	↑ (Increase)
3. C First	4.00 3.47 3.00	.43 X = .50 .57	$\frac{.07}{.43}$ = .163	↓ (Decrease)
4. Random C	5.00 4.14 4.00	.29 X = .41 .43	$\frac{.02}{.43}$ = .047	↑ (Increase)

The location of Daly and Dickson-Markman (1982) contextual essay means can be interpreted in light of the range of increase (or decrease) for the criterion essay employed in the study. The middle-ranked criterion essay is assumed to have a score value of 4 since Daly and Dickson-Markman's hypothesized scale had 7 score levels. However, in the rereading of this criterion, we find that the score value of 4 is approximately obtained only under the random reading condition (as shown in Table 1). This tends to suggest that context effects were present in the initial reading of the criterion. When the criterion is reread before the rereading of other essays, we tend to get an "absolute value." In this case, that value is 3.47. Thus the middle-ranked essays used in this study are not exactly on average (equal to 4) at the midpoint.

The possible tendency for a middle-ranked essay to decrease or increase in score level (points) when reread in a given context is shown in Table 1. The greatest absolute percentage change in the score level ($R_i+ = .233$) occurs when the criterion is read after a series of essays that received lower than middle scores on the score scale (LLLLC). In this condition, scores for the criterion essays increased along the score scale. The least percentage change ($R_i+ = .047$) in score level occurred when the criterion was read after a random selection of essays that included scores of various ranks along the score scale. In this condition (Random C), there was a slight increase in scores on the criterion essays. When the criterion was read after a series of essays that received scores greater than the middle scores on the score scale (HHHHC) and when read first (C First), there was a decrease in score points along the score scale. Now that we have obtained possible measures of the effect of context alone, we can proceed to construct a theoretical model that relates context effects to score level effects.

We can use R_i+'s in Table 1 to calculate the possible difference between an initial and reread essay score for a criterion essay read under various conditions (j). For example, the difference between the initial R_i+ and the reread R_i+ for essay scores in condition 1 (HHHHC) is .43-.51=-.08. We can employ the absolute difference in R_i+'s that result from the assignment of different scores to the reread essays and refer to this difference as an index of possible essay score change (ESC). Here we note that in order to substantiate the usefulness of the R_i's with regard to ESC, we would need to examine the relative direction and magnitude of ESC data from at least two scales of radically different lengths.

15

Suppose we let ESC be defined in terms of context effects and score level effects. It is hypothesized that there is a relationship between the possible relative range of increase (R_i+) or decrease (R_i-) in score points for an essay that is to be reread and the context j in which it might be read. We can develop a mathematical model to represent this relationship. That is, essay score change can be expressed as a mathematical statement of the hypothesized relationship between a context variable and a relative range of increase or decrease variable. Thus, the general formula for observing potential essay score change at each score level can be defined as:

$$ESC_{ij} = \left| [(R_j) \text{ Conefct}_j - (R_i+) \text{ Conefct}_j] \right| \qquad (1)$$

where Conefct_j = the context effect or the absolute percentage change in the score value of a criterion essay associated with rereading under particular reading condition j,

R_i+ = the relative range of increase for each score level i within the score scale, and

R_j = the highest value of R_i-, if the criterion essay (C) <u>decreases</u> under a given condition j, or the lowest value of R_i+, if C <u>increases</u> under a given condition j.

Equation (1) represents an interaction model. It has a context effect term and an interaction term. In the interaction term, the relative range of increase or decrease interacts with the context effect to produce a score level effect. This interaction will differ for each score level.

For example, suppose we assume that the score scale model presented in this paper is valid, then the potential ECS for an essay that is reread

16

under the HHHHC condition (condition number 1 in Table 1) and has an
original score of 5 is

$$ESC_{51} = \begin{vmatrix} (.86)(.186) - (.29)(.186) \end{vmatrix}$$
$$= \begin{vmatrix} .160 - .054 \end{vmatrix}$$
$$= .106$$

where $Conefct_1$ = .186 or the absolute percentage change in score value of a

        criterion essay associated wtih rereading under the HHHHC

        condition,

    R_1^+ = .29 or the relative range of increase for score level 5

        within the score scale, and

    R_j = .86 or the highest value of R_1^- since the criterion

        decreased under the HHHHC condition.

Since ESC_{51} = .106, this means that .106 could be translated into a possible
predicted score change for that essay paper in that HHHHC condition.

    Essays reread under the HHHHC condition which have original scores of 1
and 4 (a middle-ranked essay) would have potential ECS values of:

$$ESC_{11} = \begin{vmatrix} (.86)(.186) - (.86)(.186) \end{vmatrix}$$
$$= .000$$

$$ESC_{41} = \begin{vmatrix} (.86)(.186) - (.43)(.186) \end{vmatrix}$$
$$= .160 - .080$$
$$= .080$$

    The examples above show that it is very important to examine the
location of a score on a score scale when assessing possible score change
due to context effects. In a situation where context effects could cause
decreases in scores (i.e., HHHHC), the score level mitigates the extent to
which a score might change (decrease). For a most extreme score (i.e.,

score level 1), there would be a floor effect such that ESC_{11} = .000. This same analogy can be made when context effects could cause increases. There would be a comparable ceiling effect at the upper end of the score scale.

The calculation of ESC_{41} highlights the possible utility of the score scale model. That is, when the model is applied to the criterion score under condition Number 1 (HHHHC), ESC_{ij} is equal to .08 which is the absolute value of the difference (.43-.51 = .08) between the initial and reread score of the criterion essay (as shown in Table 1). This indicates that the model could retrieve the actual value for essay score change of a criterion essay. Confidence in the accuracy of the ESC_{ij}'s for the other score levels within the score scale is based on the assumption that the R_i+ or R_i- associated with a particular context (Conefctj) would be the same throughout the score scale. This assumption appears plausible since the R_i+ and R_i- associated with the particular context are derived from the condition when an essay, previously assigned a middle rank, was reread. That is, we are employing the assumption that measurements based on central data points are reliable for examining a phenomenon provided those points have been measured in a reliable manner. This assumption does not entail the possible score change among various contexts. The results from the modeling in this study suggests that context effects do operate differentially among the score levels. The interaction term in the ESC_{ij} model, $(R_i)(Conefct_j)$, represents this possible differentiation.

The plots of the ESC_{ij} for the four context conditions show that ESC_{ij} is a monotone function (as seen in Figures 1a-1d). There is evidence from actual essay score data, observed at every point along the score scale, that essay score changes for a first-reading versus a second-reading do result in

monotonic increases or decreases in second-reading scores (Paden, 1984; 1985). In Paden (1984), the general finding is that the increases or decreases in essay scores that changed from a first-reading to a second-reading mirror the monotonic increases or decreases in the relative ranges of increase (R_i+) or decrease (R_i-) along the score scale. Thus, we have two sources of evidence that give support to the formulation of the hypothesized essay score change model (1) presented in this study. That is, the ESC_{ij} model retrieves the actual data for the criterion (C) essay and it models the monotonic behavior of increases or decreases as found in actual score changes along a score scale. Nonetheless, the validity of the ESC model as a predictive tool needs to be cross validated with a fresh sample of essay score data.

RE 1a. ESSAY SCORE CHANGES (ESCi1)
CONTEXT1 -HHHHC

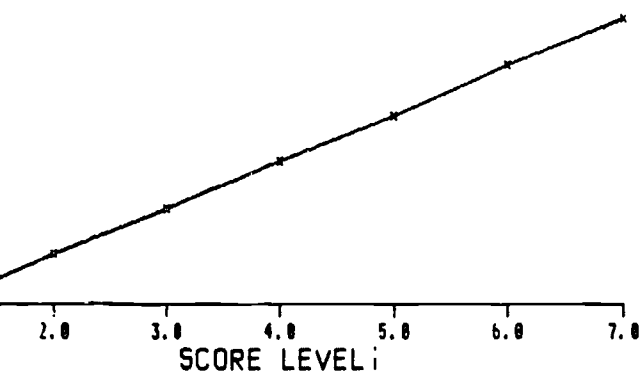SCORE LEVELi

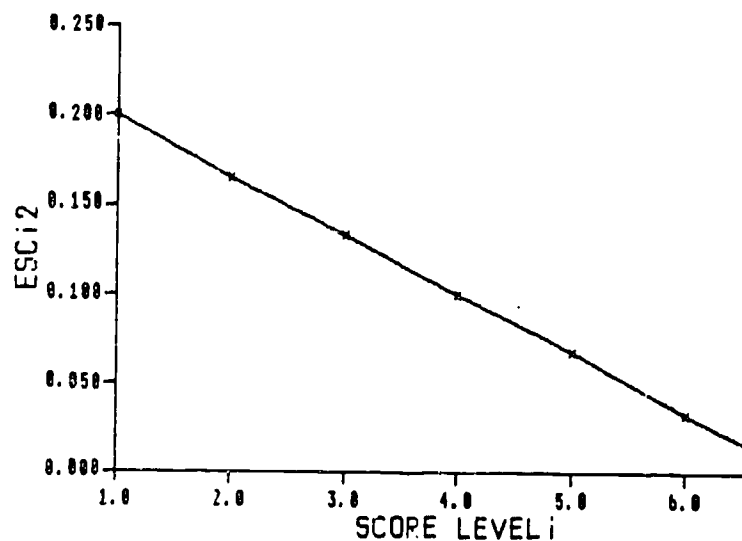FIGURE 2a. ESSAY SCORE CHANGES (ES
CONTEXT2 -LLLLC

SCORE LEVELi

RE 3a. ESSAY SCORE CHANGES (ESCi3)
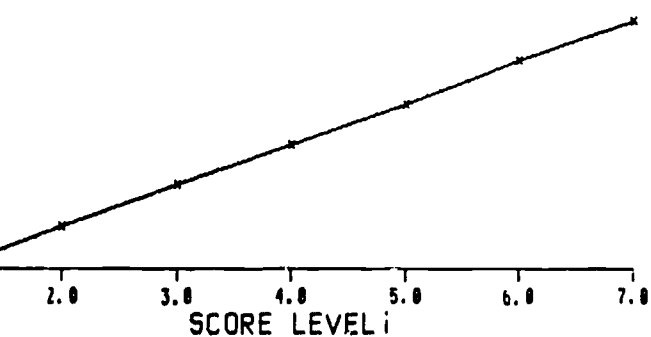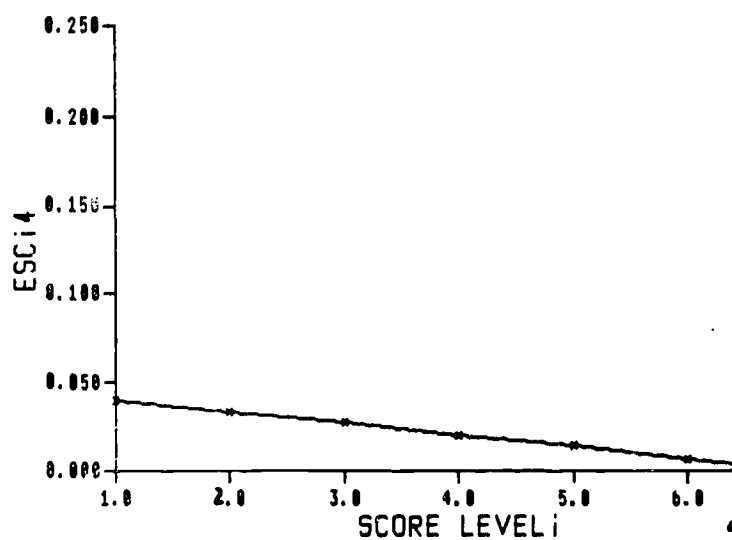CONTEXT3 - CFIRST

SCORE LEVELi

FIGURE 4a. ESSAY SCORE CHANGES (E
CONTEXT4 - RANDOMC

SCORE LEVELi

CONCLUSION

This study illustrates the possible relationship between context effects and score level effects. It is shown that context effects should be viewed within the range of possible score levels that could be assigned to a given written essay. The analyses above were made possible by drawing upon research that demonstrates that there is a differential awarding of scores (values) to a middle-ranked criterion essay depending on context conditions.

In essence, this study has provided a framework for further investigating the nature of the score scale. We are already familiar with regression toward the mean, where duplicate measurements regress toward the center of the score scale. A study of context effects could indicate some influential changes for <u>scores</u> whose original measurements were <u>at the</u> <u>midpoint</u>. That is, score changes for middle-ranked criterion essays might be determined by the nature of the reading process, which would entail contrasts in the perceived quality of a written sample in a given context.

This leads us to some suggestions for theorizing about the outcome of a given rereading of essays. If many essays receiving middle scores for the first reading (MES) tend to receive lower scores for the second reading, this could indicate that a large number of other previously reread essays were perceived to be high on the score scale (i.e., a prototype of HHHHC condition). This condition would tend to produce a lower mean score for the first reading in comparison to the second reading. If MES's tend to receive higher scores for the second reading, this could indicate that a large number of other previously reread essays were perceived to be low on the score scale (i.e., a prototype of LLLLC condition). This condition would

22

tend to produce a higher mean score for the first reading in comparison to the second reading. If MES' tend to remain average or near the middle of the score scale, this gives some indication that scores were (approximately) normally distributed about the midpoint of the score scale (i.e., a prototype of C after random condition). In this condition the means for the first and second readings would tend to be about the same. Thus we have another perspective from which we could judge the outcome of the essay reading and scoring process. That is, the study of essay grades resulting from context effects and score level effects has helped us to consider another aspect of the reading process with regard to the rereading of middle-ranked essays.

If we validate the hypothesized models outlined in this study and link them to our knowledge of the regression toward the mean phenomenon, we may offer a more unified explanation for the behavior of changes in scores assigned to original and reread essays across the entire score scale. That is, we could examine the extent to which regression effects and context effects contribute to a change in score levels for reread essays. This could be accomplished by explaining the nature of the direction and magnitude of score change present after regression effects have been measured and partialed out of discrepant scores for reread essays. Such an explanation is needed in our quest to assign reliable scores to essays regardless of the score level. The operational consequences of this research could provide us with a tool to analyze, interpret and, perhaps, monitor the extent to which context effects and score level effects influence the reliability of essay scores.

REFERENCES

Daly, J. A. & Dickson-Markman, F. (1982). Contrast effects in evaluating essays. Journal of Educational Measurement, 19, 309-316.

Hales, L. W. & Tokar, E. (1975). The effect of the quality of preceding responses on grades assigned to subsequent responses to an essay question. Journal of Educational Measurement, 12, 115-117.

Hughes, D. C. & Keeling, B. (1984). The use of model essays to reduce context effects in essay scoring. Journal of Educational Measurement, 21, 277-281.

Hughes, D. C., Keeling, B., & Tuck, B. F. (1980). Essay marking and the context problem. Educational Research, 22, 147-148. (a)

Hughes, D. C., Keeling, B., & Tuck, B. F. (1980). The influence of content position and scoring method on essay scoring. Journal of Educational Measurement, 17, 131-135. (b)

Hughes, D. C., Keeling, B., & Tuck, B. F. (1983). The effects of instructions to scorers intended to reduce context effects in essay scoring. Educational and Psychological Measurement, 43, 1047-1050.

Paden, P. A. (1984). Preliminary study of decreases in essay scores for score-split versus random AP Literature and Composition candidate sample. Unpublished memorandum. Princeton, NJ: Educational Testing Service.

Paden, P. A. (1985). English Literature and Composition score-split study. (Statistical Report SR-85-38). Princeton, NJ: Educational Testing Service.

PPRRNCMEKG226