This study examined rater effects on essay scoring in an operational monitoring system from England's 2008 national curriculum English writing test for 14-year-olds. We fitted two multilevel models and analyzed: (1) drift in rater severity effects over time; (2) rater central tendency effects; and (3) differences in rater severity and central tendency effects by raters' previous rating experience. We found no significant evidence of rater drift and, while raters with less experience appeared more severe than raters with more experience, this result also was not significant. However, we did find that there was a central tendency to raters' scoring. We also found that rater severity was significantly unstable over time. We discuss the theoretical and practical questions that our findings raise.

Biases in assessment results stemming from raters are a threat to validity because they are construct-irrelevant sources of variance (Messick, 1995). Linear and item response theory modeling offer statistical approaches to identifying and rectifying some of these rater biases. However, using these approaches does not guarantee that all rater effects will be identified and eradicated from the test scores. An important difficulty is that, while there is a sizeable literature on rater effects, we still do not know enough about the causes and corollaries of rater effects. Research has been conducted on the relationship between rater characteristics and rater effects (e.g., Lunz & Stahl, 1990; Powers & Kubota, 1998; Royal-Dawson & Baird, 2009; Shohamy, Gordon, & Kraemer, 1992) and recently there has been an active interest in whether rater effects are stable over time (Congdon & McQueen, 2000; Harik et al., 2009; Hoskens & Wilson, 2001; Lamprianou, 2006; Myford & Wolfe, 2009). Nevertheless, a great deal remains unknown about the nature of rater effects and how rater characteristics and rating conditions affect scoring.

The current study examined rater effects on essay scoring in an operational monitoring system from England’s 2008 national curriculum English writing test for 14-year-olds. We performed a secondary analysis of the operational scoring data and addressed the following three research questions:

1. Severity drift—did raters change in severity over time and was this effect stable?
2. Central tendency—did raters make full and correct use of the point scale?
3. Rater experience—did rater experience influence the accuracy of scoring?

Severity Drift

Several studies have investigated whether raters change their practices over time. These studies have presented mixed findings. Lunz and Stahl (1990) conducted research on three assessments: essay, clinical, and oral examinations. They did not find...
significant effects of time on scoring for the oral examination. For the clinical and essay examinations, they found severe ratings on the second and third half-day sessions, respectively, but thereafter ratings became more lenient on average and significantly so for the clinical examination. Pinot de Moira, Massey, Baird, and Morrissey (2002) investigated drift in rater severity over a scoring period which lasted several weeks. Their study, which applied multilevel modeling, analyzed 2,692 students’ English extended writing test performances. They found that the 78 raters involved scored significantly more severely toward the end of their scoring period than at the beginning. Congdon and McQueen’s (2000) study on the writing performances of 8,285 elementary school students, with 10 raters, was conducted over a shorter time period of 7 days. Nine of the raters became more severe over time while one rater became more lenient. Their finding suggests that, even within a testing and scoring environment, different raters may have different trends in their scoring over time. Indeed, Myford and Wolfe’s (2009) study of 101 raters and 28 check essays (representative samples of students’ work) found significant positive and negative drift in rater accuracy over time for a small proportion of their raters.

A related issue raised by the Congdon and McQueen (2000) study and by subsequent researchers (Harik et al., 2009; Hoskens & Wilson, 2001; Lamprianou, 2006) is whether rater severity effects are stable over time. Myford and Wolfe (2009) found that, when rater effects were not stable, these effects did not follow a trend. Rather, they tended to change erratically from check to check. Lamprianou (2006) investigated rater effects across academic subjects as well as over time. He found that rater severity was unstable across both dimensions. All of these studies showed significant within-rater variability of rater effects over time. This common finding raises the question: Are rater effects stable traits or unstable states?

Studies on the stability of rater effects have tended to be small in scale, involving few raters or scores (Congdon & McQueen, 2000; Hoskens & Wilson, 2001; Lamprianou, 2006; however, see Wolfe, Myford, Englhard, & Manolo, 2007 for a notable exception), or have, at least on some score scales, confounded task with rater effects (Harik et al., 2009). The current study analyzed a large data set, and confounding the task and rater effects was prevented by having all raters score the same tasks at each time point.

Central Tendency

Kingsbury (1922) categorized rater effects in terms of severity, halo effects (Thorndike, 1920), and central tendency. In this paper, we examined drift in severity and central tendency but we did not investigate halo effects. Halo effects refer to the tendency to give highly correlated ratings to individuals across a range of criteria. Central tendency is the propensity to award a restricted range of scores around the mean (or mode or median) and to avoid awarding extreme scores (Saal, Downey, & Lahey, 1980). Central tendency is a well-established phenomenon documented in various contexts including in the assessment of Advanced Placement English Literature and Composition essays (Myford & Wolfe, 2009), school writing examinations in Georgia (Engelhard, 1994), English as a second language (Knoch, Read, & von Randow, 2007), and writing and speaking in German as a foreign language (Eckes,
Knoch et al. (2007) found that rater training increased the central tendency effect in their research into the scoring of an English writing examination in a New Zealand university. They discussed the possibility that raters were more likely to exhibit central tendency when they knew that they were being monitored. Wolfe et al. (2007) referred to this as a “play-it-safe” effect because raters knew that their scoring was less likely to come under scrutiny if they did not use the extremes of the scale.

Rater Experience

Research also has attempted to examine how rater experience is related to rater effects. Weigle (1999) found that, for one of the two essay prompts included in her study, inexperienced raters initially scored more severely relative to experienced raters but then scored more leniently with time; experienced raters’ scoring remained steady throughout. While rater age is not the same as rater experience, the effects of rater age on scoring may well be similar to those of rater experience. In research looking for the criteria emphasized by different types of raters in written language performance assessments, Eckes (2008) found that older raters, regardless of their rating experience, tended to see the scoring criteria as less important than younger raters. Further, those raters who saw the scoring criteria as less important scored more severely. Two other studies have found that expert raters used a wider range of criteria when making judgments compared with inexperienced raters (Cumming, 1990; Wolfe, Kao, & Ranney, 1998). The finding that older and more experienced raters may make judgments about essay quality using factors not recognized by the scoring rubric is a serious threat to both the validity and reliability of assessment procedures.

Other studies have looked at particular aspects of rater experience, such as whether a teaching background is necessary for accurate scoring. Several of these studies have shown that teaching experience is not always a necessary criterion for the selection of raters. These include studies of a test of English as a foreign language (Shohamy, Gordon, & Kraemer, 1992), an oral test for Japanese language tour guides (Brown, 1995), an occupational English test (Lumley, Lynch, & McNamara, 1994), and an extended writing test of English for 14-year-olds (Royal-Dawson & Baird, 2009). A study of the Graduate Management Admission Test found that even subject knowledge was not a requirement for effective rating (Powers & Kubota, 1998). Royal-Dawson and Baird (2009), however, argued that teaching experience was necessary for highly curriculum-specific tests with holistic scoring.

Despite the literature referenced above, much is still unknown about the way rater experience interacts with rater effects. Indeed, Myford and Wolfe (2009) have called for more research to be conducted on the impact of rater experience on rater effects and their stability.

The current paper analyzed essay scores for a writing test with a curriculum-specific essay question. All of the raters in our analysis had curriculum-specific teaching experience. We compared rater effects for three groups of raters with differing levels of prior rating experience: team leaders, experienced raters, and new raters. Thus, we examined whether rater severity drift and rater central tendency
differed significantly by prior rater experience. With respect to central tendency, our approach was to model average central tendency for each of our three groups of raters, rather than modeling individual central tendency rater effects (see residuals-expected rating correlation in Wolfe, 2004 and Myford & Wolfe, 2009). We summarized the magnitude of these average central tendency effects by (1) calculating the correlation between the average scores assigned to essays with the “correct scores” assigned to the same essays by an expert committee (discussed in the Method section) for each type of rater, and (2) comparing the variance of these average essay scores to the variance of the expert committee scores.

Method

The English Test

Since 1988, England has operated a national curriculum, which standardizes the academic content taught across schools. There are four key stages: Key Stage 1 (ages 5–7), Key Stage 2 (ages 7–11), Key Stage 3 (ages 11–14), and Key Stage 4 (ages 14–16). A main purpose of the national curriculum is to enable individual assessments in English, mathematics, and science to be administered to all students in the country at the end of each key stage. The high numbers of students (there are approximately 600,000 students in a year group) and tests involved means that each student’s work is scored by only one rater. The data from these tests are used by the government to publish so-called “school league tables” of schools’ academic performances (http://www.education.gov.uk/performance) and are used to hold schools accountable as part of the government’s official school inspection system (http://www.ofsted.gov.uk/). These tables also are promoted to parents as guides to aid them in choosing their child’s school. Further discussion of the national curriculum tests and their high-stakes role in the publication of school league tables can be found in Leckie and Goldstein (2009, 2011).

The current study analyzed scoring data from the writing paper of the 2008 Key Stage 3 English test. (See https://orderline.qcda.gov.uk/gempdf/1847217923/1847218016.pdf to view the question paper.) The question paper required students to answer two essay questions; we focused on the scoring of the first essay question, which accounted for 60% of the total available points. Students had to imagine that they were explorers who, on their last trip, lost contact with the outside world. They then had to write the story of their experience. Students were given 45 minutes to write the essay.

The essay was scored on a 30-point scale, but to aid raters in scoring, this was divided into three subscales: (1) sentence structure and punctuation (a maximum of 8 points were available); (2) text structure and organization (8 points); and (3) composition and effect (14 points). Each subscale then was further divided into narrow bands of [typically] one, two, or three points. Raters were further aided by a detailed scoring scheme and were given comprehensive worked examples (see https://orderline.qcda.gov.uk/gempdf/1847217923/1847217982.pdf).
Procedure

Raters were recruited from the population of teachers of the Key Stage 3 curriculum. All raters underwent training before they were allowed to score live papers. Training took place in face-to-face meetings involving 100 raters allocated to teams of 10, and each team was headed by a team leader. Raters scored training essays and, when there were disagreements, they discussed with their team the reasoning for the scores they assigned. Once training was completed, raters scored 10 check papers to confirm they were scoring accurately. Check papers were selected and prescored by an expert committee and were chosen as representative and uncontroversial examples of students’ work at different points on the scoring scale. The expert committee consisted of senior raters and test developers who had experience in designing essay prompts, developing scoring guidelines, training and supervising raters, and rating essays for this test (Myford & Wolfe, 2009 refer to this as an “external frame of reference”). The scores assigned by the expert committee were consensus scores agreed upon through discussion. Unfortunately, separate scores for each expert on the panel were not present in the data and so it was not possible to explore the reliability of the panel members’ scoring. Raters who failed to score the check papers closely enough to these expert scores were mentored by their team leaders and those who failed at a second attempt were prevented from progressing to live scoring.

During live scoring, raters underwent further checks every time they scored 100 live papers. At each check, raters scored a set of six check papers taken from 10 available sets for that check. Thus, 60 papers were scored at each check. However, across all checks, 14 sets were used at some point and therefore a total of 84 check papers were scored. As with the training check papers, these check papers were selected and prescored by the expert committee and were again chosen to provide raters with “typical” essays for a range of students’ English proficiencies. It is the rater scores and expert scores assigned to these check papers that were analyzed in this study, not the scores assigned to live papers. The expert scores indicated that these essays varied considerably in quality and some essays scored the minimum or maximum scores available. Unfortunately, student background information was not present in the data and so it was not possible to relate these scores to students’ demographic characteristics. As with initial training, raters who failed to score the check papers closely enough to the expert scores were provided with mentor feedback, and those who failed at a second attempt were withdrawn from live scoring and all previously scored live papers were rescored. We therefore did not analyze failed raters as they did not assign any student’s final scores. It was the raters who assigned students’ final scores who were of real interest in this high-stakes testing context.

In total, there were 689 raters: 135 team leaders, 372 experienced raters, and 182 new raters. Experienced raters scored the same test in at least one previous year; for new raters, this was their first year. The data did not contain rater background information and so it was not possible to summarize these three groups by their demographic characteristics. All three types of raters scored live papers and therefore underwent quality checks. However, team leaders additionally acted as mentors to the experienced and new raters and communicated feedback to them about the
quality of their scoring. To fulfill this role, team leaders needed to observe the scores that the raters in their teams assigned to essays. Team leaders therefore were allocated different sets of check papers for their own quality checks to those allocated to the experienced and new raters: 6 of the 14 sets (i.e., 36 of the 84 check papers) were reserved for the team leaders; the remaining 8 sets (i.e., 48 check papers) were reserved for the experienced and new raters. Delays in the allocation of live papers to raters and differences in the speed with which raters worked led to raters scoring different total numbers of live papers. This in turn led to variation in the number of checks that each rater underwent. The number of raters at checks one, two, three, four, and five was 689, 664, 630, 524, and 186, respectively. Thus the majority of the raters (524 of 689) underwent at least four checks.

Overall, the data contained 34,920 scores that 689 raters assigned to 84 check essays. The mean score was 14.31 of 30 while the standard deviation was 7.74. In our statistical models, the response variable was the score difference, defined as the signed difference between the score assigned by the rater and the consensus score assigned by the expert committee. The score difference measured the extent to which the assigned scores deviated from the expert scores. A value of zero indicated that the rater assigned the same score to the essay as the expert score. A positive value indicated that the rater overscored the essay relative to the expert score (s/he was lenient), while a negative value indicated that the rater underscored the essay relative to the expert score (s/he was severe). The mean score difference was .28 and so raters, on average, slightly overscored the essays. While the mean was small, the associated standard deviation of 3.72 suggested that some raters assigned scores that were quite different from the expert scores. (Distributional plots of the scores and score differences are available from the authors upon request.)

Multilevel Models

We fitted two separate multilevel models (or hierarchical models, random-effects models, mixed models) to the data to address our three research questions (Goldstein, 2010; Raudenbush & Bryk, 2002). We used the first model to measure raters’ mean levels of severity and to measure whether there was a central tendency to their scoring (research question 2). We fitted the second model to establish whether raters’ levels of severity drifted over time (research question 1). In both models, we examined whether these rater effects differed across the three groups of raters (research question 3).

Multilevel models have become a standard approach in the analysis of clustered or hierarchical data. In our data, scores can be viewed as nested within raters, but they also can be viewed as nested within essays. However, raters and essays are not themselves nested within one another as each rater scored multiple check essays and each check essay was scored by multiple raters. Instead, raters and essays form a cross-classification and scores are nested within the cells of this cross-classification. We therefore fitted cross-classified multilevel models (or crossed random effects models) to our data, which explicitly recognized this complex structure (Rasbash & Goldstein, 1994; Raudenbush, 1993). We fitted all models using Iterative Generalized Least Squares (IGLS), a maximum likelihood-based procedure in the MLwiN.
statistical software package (Rasbash, Charlton, Browne, Healy, & Cameron, 2009). We ran MLwiN through the Stata statistical software package by using the user written runmlwin Stata command (Leckie & Charlton, 2011).

We introduce each multilevel model in turn below. To make our description of the models as accessible as possible, we explain each model in two stages. In the first stage, we present a simplified version of the model, which ignores rater type and assumes that rater effects are the same across the three groups of raters. In the second stage, we present the full model where we do account for rater type and so rater effects are allowed to differ across the three groups.

**Model 1.** Model 1 ignored time and therefore assumed that all rater effects were stable across the five checks.

\[
y_{ij} = \beta_0 + u_j + v_i + e_{ij}
\]

\[
 u_j \sim N(0, \sigma^2_u), \quad v_i \sim N(0, \sigma^2_v), \quad e_{ij} \sim N(0, \sigma^2_e).
\]  

The response variable \(y_{ij}\) was the score difference (i.e., the signed difference between the score assigned by the rater and the consensus score assigned by the expert committee) assigned to essay \(i (i = 1, \ldots, I)\) by rater \(j (j = 1, \ldots, J)\). The fixed part of the model contained only an intercept parameter \(\beta_0\) which gave the predicted score difference across all raters.

The random part of the model contained rater random effects \(u_j\), essay random effects \(v_i\), and a residual random effect \(e_{ij}\). These random effects partitioned the variance in score differences about \(\beta_0\) into independent sources of variation due to raters, essays, and the residual. Each random effect was assumed normally distributed with mean zero and a constant variance. In the analysis, the legitimacy of these assumptions were investigated through normal quantile-quantile (or probability) plots.

The sum \(\beta_0 + u_j\) gave the predicted score difference for rater \(j\) and was a measure of whether that rater, on average, scored their essays severely relative to the consensus scores assigned by the expert committee. The variance parameter \(\sigma^2_u\) measured the extent to which raters differed in their scoring severity from one another.

The sum \(\beta_0 + v_i\) gave the predicted score difference for essay \(i\) and was a measure of whether that essay was, on average, scored severely by the raters relative to the score assigned by the expert committee. The variance parameter \(\sigma^2_v\) measured the extent to which essays differed in the severity of scoring that was applied to them. Plotting \(\beta_0 + v_i\) against the consensus score assigned by the expert committee for each essay provided a graphical indication, albeit not a formal statistical test, of whether there was, on average, a central tendency bias to raters’ scoring. If there was a central tendency bias, the scatter plot would be expected to show a negative association between the two variables. This plot therefore allowed research question 2 to be addressed: Did raters make full and correct use of the point scale? It is worth noting that this approach is similar to a rater averaged version of the “residual-expectation” correlation approach developed by Wolfe (2004) and subsequently applied by Myford and Wolfe (2009). (In our approach, the predicted score differences are
analogous to their “residuals” while the expert consensus scores are analogous to their “model-based expected ratings.”

Model 1 assumed that each parameter $\beta_0$, $\sigma^2_u$, $\sigma^2_v$, and $\sigma^2_e$ was equal across the three rater groups. In our analysis, this assumption was relaxed by interacting the intercept parameter $\beta_0$ and the three random effect terms $u_j$, $v_i$, and $e_{ij}$ in the model equation with a set of three binary indicator variables, one for each of the three rater groups. This gave a separate set of four parameters for each rater group and therefore allowed each of the effects described above to vary across groups. In the analysis, Wald tests then were used to test whether each parameter varied significantly across the groups. These tests allowed research question 3 to be addressed: Did rater experience influence the accuracy of scoring?

**Model 2.** Model 2 allowed raters’ levels of severity to vary flexibly across checks. Specifically, model 2 extended model 1 by allowing raters’ predicted score differences to change from check to check and to be modeled as nonlinear trajectories. We specified the model such that each rater’s nonlinear trajectory was constructed from two parts: a rater-specific linear time trend plus a set of rater-specific check-to-check departures from that time trend. It is the latter which allowed each rater’s trajectory to have a flexible nonlinear shape. The advantage of this specification was that the model directly gives us an overall average linear time trend. The slope of this overall average linear time trend allowed a simple test of whether raters, on average, became significantly more (or less) severe across the five checks. This method of modeling change is essentially the same as that used in standard growth curve models that are fitted to repeated measures data (Goldstein, 2010; Hedeker & Gibbons, 2006; Raudenbush & Bryk, 2002; Singer & Willett, 2003):

$$y_{ijt} = \beta_0 + \beta_1 \text{time}_{ijt} + u_{0j} + u_{1j} \text{time}_{ijt} + c_{jt} + v_i + e_{ijt}$$

$$\begin{pmatrix} u_{0j} \\ u_{1j} \end{pmatrix} \sim N \left( \begin{pmatrix} 0 \\ 0 \end{pmatrix} , \begin{pmatrix} \sigma^2_u & \sigma_{u01} \\ \sigma_{u01} & \sigma^2_{u1} \end{pmatrix} \right) , \quad c_{jt} \sim N \left( 0, \sigma^2_c \right) ,$$

$$v_i \sim N \left( 0, \sigma^2_v \right) , \quad e_{ijt} \sim N \left( 0, \sigma^2_e \right) .$$

A $t$ subscript is added to the response variable $y_{ijt}$ to indicate the time $t (t = 1, \ldots , 5)$ at which essay $i$ was scored by rater $j$. The predictor variable $\text{time}_{ijt}$ was added to the model and took the value zero at check one and incremented by one unit with each successive check. The average linear time trend describing the relationship between $y_{ijt}$ and $\text{time}_{ijt}$ was given by $\beta_0 + \beta_1 \text{time}_{ijt}$. The intercept of this line $\beta_0$ measured the mean score difference across raters when they started off at the first check ($\text{time}_{ijt} = 0$), while the slope of this line $\beta_1$ measured the average change in the score difference for each check. If $\beta_1 > 0$ then raters, on average, became more lenient with time while if $\beta_1 < 0$ they became more severe with time. Thus, we used this average linear time trend to detect overall changes in the mean level of scoring severity across all raters as time progressed. A formal test of whether $\beta_1$ was significant allowed research question 1 to be addressed: Did raters change in severity over time and was this effect stable?
The rater random effects $u_{0j}$ and $u_{1j}$ were, respectively, the individual rater departures from the intercept and slope of the average linear time trend. Thus, the linear part of rater $j$’s flexible trajectory was given by $(β_0 + u_{0j}) + (β_1 + u_{1j}) \text{time}_{ijt}$. Positive values of $β_0 + u_{0j}$ indicated raters who were initially lenient relative to the expert committee. Positive values of $β_1 + u_{1j}$ indicated raters who became more lenient over time relative to the expert committee. The variance parameter $σ^2_u$ measured the extent to which raters initially differed in severity, while $σ^2_{u1}$ measured the extent to which raters differed in their rates of change in severity. The covariance parameter $σ_{u01}$ measured the covariance between the intercepts and slopes of raters’ linear time trends. A positive covariance implied that raters with above average initial score differences tended to have above average rates of change in score differences; a negative covariance implied that those raters tended to have below average rates of change. The random effect $c_{jt}$ played an important role in the model as it was a rater-check random effect which allowed the scoring severity of each rater at each check to depart from that predicted by their linear time trends. It was these rater-check random effects, which allowed raters’ trajectories their flexible, nonlinear shapes. The variance parameter $σ^2_c$ summarized the importance of these departures and therefore the extent to which raters’ trajectories were nonlinear. An examination of the relative magnitude of $σ^2_c$ allowed the second part of research question 1 to be addressed: Did raters change in severity over time and was this effect stable?

As with model 1, model 2 assumed that each parameter was equal across rater groups. In the analysis, this assumption was relaxed in the same way as before, by interacting each term in the model equation with a set of three binary indicator variables, one for each of the three rater groups. Wald tests for whether each parameter varied significantly across groups were again used to address research question 3: Did rater experience influence the accuracy of scoring?

### Results

#### Model 1: Rater Severity, Experience, and Central Tendency

Table 1 presents parameter estimates for model 1 fitted to the essay score differences (i.e., the difference between the scores assigned by the raters and the consensus scores assigned by the expert committee) where we fitted separate parameters for each rater group. A likelihood ratio test comparing this model to one where we restricted all parameters to be equal across rater groups indicated that, at the .05 level of significance, there were statistically significant differences across the rater groups ($χ^2_8 = 497. p < .001$).

In the fixed part of the model, the estimated intercept parameters gave the predicted score difference for each rater group. Team leaders had a score difference of .49 points and the $p$-value of .014 indicated that this effect was statistically significant. Thus, team leaders, on average, overscored their essays by half a point on the 30-point scoring scale compared to the scores assigned to the same essays by the expert committee; team leaders scored leniently. Experienced and new raters overscored by .30 and .07 points, respectively, but the corresponding $p$-values of .290 and .787 indicated that neither of these effects was statistically significant. The estimated intercept parameters also suggested that raters overscored by smaller amounts...
Table 1
Parameter Estimates for Model 1 Fitted to the Essay Score Differences

<table>
<thead>
<tr>
<th></th>
<th>Team Leaders</th>
<th>Experienced Raters</th>
<th>New Raters</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Est.</td>
<td>SE</td>
<td>p</td>
</tr>
<tr>
<td>Fixed part</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept ((\beta_0))</td>
<td>.49</td>
<td>.20</td>
<td>.014</td>
</tr>
<tr>
<td>Random part</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Between-rater variance ((\sigma^2_u))</td>
<td>.74</td>
<td>.13</td>
<td></td>
</tr>
<tr>
<td>Between-essay variance ((\sigma^2_v))</td>
<td>.95</td>
<td>.30</td>
<td></td>
</tr>
<tr>
<td>Residual variance ((\sigma^2_e))</td>
<td>4.96</td>
<td>.14</td>
<td></td>
</tr>
</tbody>
</table>

Note. \(-2 \log L = 89633\). Est. = parameter estimate, SE = standard error, and \(p = p\)-value. \(P\)-values are not provided for the variance terms as the Wald test for these parameters is only approximate. Normal quantile-quantile plots suggested that the normality assumptions made for the random effects in this model were reasonable. (These plots are available from the authors upon request.)

the less previous scoring experience they had: the team leaders overscored the most, experienced raters overscored to a lesser degree and new raters overscored the least. However, a Wald test did not reject the null hypothesis that all three intercepts were equal and so these differences were not statistically significant\(\chi^2 = 7.23, p = .065\).

In the random part of the model, the estimated between-rater variances for team leaders, experienced raters and new raters were .74, 1.63, and 1.53, respectively. A likelihood ratio test comparing model 1 to one where the three between-rater variances were restricted to be equal indicated that these differences were statistically significant \(\chi^2 = 17.29, p < .001\). To aid the interpretation of these estimates, we calculated and plotted raters’ predicted score differences (i.e., raters’ severity levels) \(\beta_0 + u_j\) (y-axis) in ascending rank order (x-axis) separately for each rater group (see Figure 1). In the figure, the predicted score differences were plotted as data points (black dots) and each one was presented with a 95% confidence interval (vertical thin gray lines) which allowed it to be statistically compared to a score difference of zero. Thus, the confidence intervals allowed us to establish whether each rater scored significantly differently from the expert committee. The figure also plotted the mean predicted score difference for each rater group (horizontal thick black lines).

The vertical variability of the black dots in the figure indicated that the predicted score differences varied substantially across raters for all three rater groups. For example, the most lenient team leaders (those with the highest predicted score differences) were predicted to score almost 4 points higher than the most severe team leaders (those with the lowest predicted score differences). For experienced and new raters, the difference between the most lenient and most severe raters was approximately 5.5–6 points. The wide 95% confidence intervals for these predictions indicated that these predictions were imprecise. However, this was expected due to the small number of scores assigned by each rater in our data (the average rater assigned scores to just 25 essays). The majority of the confidence intervals crossed the zero
Figure 1. Predicted score differences for individual raters (black dots) and the mean rater (horizontal thick black lines) plotted in ascending rank order for each rater group. Each predicted score difference is presented with a 95% confidence interval (vertical thin gray lines).

line, so scores of the majority of raters were not statistically significantly different from the consensus scores assigned by the expert committee.

The estimated between-essay variances for team leaders, experienced raters, and new raters were .95, 3.63, and 3.08, respectively. A likelihood ratio test comparing model 1 to a model where the between-essay variances were restricted to be equal indicated that these differences were statistically significant ($\chi^2 = 11.51, p = .003$).

We used a similar approach to interpret the magnitude of these variances as we did in Figure 1 to interpret the magnitude of the between-rater variances. We calculated and then plotted the predicted score difference $\beta_0 + v_i$ for each essay ($y$-axis). However, in Figure 2 we plotted these predicted score differences against the consensus score assigned to that essay by the expert committee ($x$-axis) rather than in ascending rank order based on the level of severity exercised (as was the case for Figure 1). Changing the $x$-axis in this way enabled the figure to provide a graphical indication, albeit not a formal statistical test, of whether there was a central tendency to raters’ scoring. If there were a central tendency bias, the data points in each scatter plot would show a negative slope; this would indicate that essays that were scored low by the expert committee were overscored by the raters and vice versa (essays that were scored high by the expert committee were underscored by the raters). However, the figure must be interpreted carefully as it is important to realize that there also was an inevitable floor and ceiling effect present in the data. For example, an essay with an expert score of zero could only be assigned a zero score or higher by the raters and so could not have been scored severely. (The gray triangles in the figure illustrate all combinations of
Figure 2. Predicted score differences for individual essays (black dots) and the mean essay (horizontal thick black lines) against the consensus scores assigned by the expert committee for each rater group. The gray triangles denote all combinations of predicted score differences and consensus scores assigned by the expert committee which were not possible due to the floor and ceiling effects present in the data.

predicted score differences and consensus scores assigned by the expert committee which were not possible due to these effects.)

Figure 2 shows that essays with positive predicted score differences were assigned scores that were on average higher than the consensus scores assigned to the same essays by the expert committee; these essays therefore were scored leniently relative to the expert committee. Essays with negative predicted score differences were assigned scores that were on average lower than the scores assigned to the same essays by the expert committee; these essays therefore were scored severely relative to the expert committee. Unlike Figure 1, we did not plot 95% confidence intervals for the predicted score differences as our focus was on discerning whether there was a central tendency to raters’ scoring rather than on identifying which individual essays had rater and expert committee scores that were significantly different.

The figure clearly shows a strong negative association between the predicted score differences and the consensus scores assigned by the expert committee for each rater group (correlations of –.58, –.69, and –.74 for team leaders, experienced raters, and new raters, respectively). Thus, essays that were assigned low scores by the expert committee were overscored by each type of rater, while essays assigned high scores by the expert committee were underscored by each type of rater. As discussed above, caution must be exercised when interpreting such negative associations as evidence of a central tendency to raters’ scoring because the plotted negative associations were biased due to the floor and ceiling effects of the scoring scale. However, it
should be noted that the trend of over-scoring lower-quality essays and under-scoring higher-quality essays was observed not just at the extremes of the scoring scale but also across the middle where the bias from the floor and ceiling effects was at a minimum; this was the case for all rater groups. Regression to the mean also may have played a role in these results, as neither the average scores assigned by raters nor the consensus scores assigned by the expert committee would have been perfectly reliable measures of the essays’ true scores; both would have been measured with some random error. However, the high number of raters that were in the sample (135 team leaders, 372 experienced raters, and 182 new raters) and the fact that consensus scores were agreed upon by a panel of experts rather than being assigned by a single individual both would have been expected to minimize this potential problem.

Further indication of a central tendency to raters’ scoring was provided by comparing, for each rater group, the variance of predicted scores for essays with the variance of expert consensus scores for the same essays. If there were a central tendency to raters’ scoring, we would have expected predicted scores for essays to be less variable than consensus scores. Variances based on predicted scores were estimated by refitting model 1 with scores (instead of score differences) as the response variable; these estimates were 50.5, 48.5, and 48.0 for team leaders, expert raters, and new raters, respectively. The corresponding variances based on the expert consensus scores were 64.9, 67.9, and 67.9. The former were substantially lower, and this again suggested that there was a central tendency to raters’ scoring.

Differences across the rater groups also are evident in Figure 2. The experienced and new raters appeared to make less use of the scoring scale extremes than team leaders; they overscored essays with low expert scores and underscored essays with high expert scores to a greater extent than did the team leaders. Importantly, we did not expect the floor and ceiling effects of the scoring scale to differentially bias the negative associations for the different rater groups. Thus, the stronger negative associations observed for experienced and new raters suggests a greater central tendency bias to these raters’ scoring compared to team leaders. However, this result was not found to be statistically significant ($\chi^2 = 1.93, p = .380$).

Figure 2 also showed that many essays were scored with different degrees of severity, even when they had been assigned the same score by the expert committee. One interpretation was that the raters incorrectly scored to different degrees on different essays. An alternative, although perhaps less likely possibility, was that the expert committee occasionally assigned the wrong score to an essay. For example, if the expert committee assigned the same score to two essays when one essay was in fact better than the other, we then would expect raters to assign higher scores to the better essay.

For each rater group, the residual variances were by far the largest component of variation and accounted for approximately two-thirds of the total variability in score differences. The majority of variation in score differences therefore was due to systematic and unsystematic sources of variation not captured by the rater or essay effects. This would have included variation due to rater disagreements as to the relative merit of the essays and temporal variation whereby the same rater would have scored an essay differently had it been scored on a different occasion. Since live
Table 2
Parameter Estimates for Model 2 Fitted to the Essay Score Differences

<table>
<thead>
<tr>
<th></th>
<th>Team Leaders</th>
<th>Experienced Raters</th>
<th>New Raters</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Est.</td>
<td>SE</td>
<td>p</td>
</tr>
<tr>
<td>Fixed part</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept ($\beta_0$)</td>
<td>.75</td>
<td>.34</td>
<td>.029</td>
</tr>
<tr>
<td>Time ($\beta_1$)</td>
<td>−09</td>
<td>.12</td>
<td>.463</td>
</tr>
<tr>
<td>Random part</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Between-rater intercept variance ($\sigma^2_{u0}$)</td>
<td>1.25</td>
<td>.28</td>
<td>1.82</td>
</tr>
<tr>
<td>Between-rater intercept slope covariance ($\sigma_{u01}$)</td>
<td>−.25</td>
<td>.09</td>
<td>−.22</td>
</tr>
<tr>
<td>Between-rater slope variance ($\sigma^2_{u1}$)</td>
<td>.09</td>
<td>.04</td>
<td>.11</td>
</tr>
<tr>
<td>Within-rater-between-check variance ($\sigma^2_c$)</td>
<td>.35</td>
<td>.13</td>
<td>.90</td>
</tr>
<tr>
<td>Between-essay variance ($\sigma^2_v$)</td>
<td>.87</td>
<td>.27</td>
<td>3.76</td>
</tr>
<tr>
<td>Residual variance ($\sigma^2_e$)</td>
<td>4.54</td>
<td>.14</td>
<td>9.13</td>
</tr>
</tbody>
</table>

Note. $-2 \log L = 89233$. Est. = parameter estimate, SE = standard error, and $p = p$-value. $P$-values are not provided for the variance terms as the Wald test for these parameters is only approximate.²

papers in the national curriculum tests were scored only once, these factors may have had a serious impact on the scores allocated to an individual student.

Model 2: Drift and Stability in Rater Severity and Rater Experience

Table 2 presents parameter estimates for model 2 fitted to the essay score differences, where again we fitted separate parameters for each rater group. As discussed in the Method section, model 2 extended model 1 by allowing raters’ predicted score differences to change from check-to-check and to be modeled as nonlinear trajectories. As model 1 was nested in model 2, we statistically compared the two models using a likelihood ratio test. The test indicated that this extension significantly improved the fit of the model to the data($\chi^2_{12} = 89, 633 − 89, 233 = 400, p < .001$).

In the fixed part of the model, the estimated intercept parameters $\beta_0$ gave the predicted score difference at the first check for each rater group while the estimated slope parameters $\beta_1$ for the time variable gave the predicted change in score difference at each check. To aid the interpretation of these estimates, we first calculated the average predicted score difference at each check for each rater group using $\beta_0 + \beta_1 time_{ij}$. We then plotted these predictions (y-axis) against time (x-axis) and we connected the data points with thick black straight lines in (see Figure 3). The resulting trajectories provided a graphical indication of whether raters’ levels of severity, on average, drifted over time. (We introduce and discuss the interpretation of the thin gray lines below.)

As indicated in Table 2 (and shown in Figure 3), team leaders started off, on average, with a significant score difference of .75 and changed by −.09 points at each
check. The effect of time was not significant, however, so there was no evidence that
team leaders, on average, changed in their level of severity over time. Experienced
raters started off, on average, with a score difference of .30 and changed by just .01
points each check. However, neither the intercept nor the effect of time was signif-
cient. There was therefore no evidence that experienced raters scored differently, on
average, from the expert committee at any check. New raters started off, on average,
with a score difference of –.07 and changed by .09 points each check. As for the
experienced raters, neither the intercept nor the effect of time was significant; as a
result, there also was no evidence that new raters scored differently, on average, from
the expert committee. Finally, a Wald test did not reject the null hypothesis that the
three rater groups shared a common trajectory ($\chi^2_{4} = 4.23$, $p = .375$). Thus while
the score differences for the three rater groups appeared, on average, to converge
over time, this pattern was not statistically significant.

In the random part of the model, the estimated rater variances $\sigma^2_{u0}$, $\sigma^2_{u1}$, and covariance $\sigma_{u01}$ described the variability in rater linear time trends about the average linear
time trend for their rater group. The estimated within-rater-between-check variance $\sigma^2_c$ described the variability in raters’ predicted score differences at each check which
was not accounted for or explained by their individual linear time trends. Thus, this
parameter allowed for nonlinear changes in rater severity over time. To aid the in-
terpretation of these estimates, we calculated the predicted score difference for each
rater at each check using $(\beta_0 + u_{0j}) + (\beta_1 + u_{1j}) \text{time}_{ij} + c_{ij}$. Figure 3 plots these
predictions (y-axis) against time (x-axis). In the figure, each rater’s series of pre-
dicted score differences is connected with straight lines. The resulting lines provided

Figure 3. Predicted score differences for individual raters (thin gray lines) and the mean rater
(thick black lines) against time, for each rater group. For clarity the plot shows a random 25%
sample of raters.
a graphical indication of how each individual rater’s level of severity changed from check to check. Thus, raters with positive predicted score differences across all five checks assigned scores to essays that were on average higher than the expert committee’s consensus scores for the same essays; these raters scored leniently relative to the expert committee at all five checks. In contrast, raters with negative predicted score differences across all five checks assigned scores to essays that were on average lower than the expert committee’s consensus scores for the same essays; these raters scored severely relative to the expert committee at all five checks.

The figure shows that raters’ predicted score differences at each check varied considerably about the average linear time trends for their group; we already showed that this variation was statistically significant. The heterogeneity in scoring severity seen for experienced and new raters was substantial and larger than that seen for team leaders. The most lenient experienced and new raters at any given check were predicted to score over 5 points higher than the most severe raters; this is consistent with our finding from Figure 1.

There also is some suggestion in the figure that the predicted lines fanned inwards over time, and this was supported by the estimated rater covariances (which were negative for each group). A likelihood ratio test comparing model 2 to a model where the three rater covariance parameters were constrained to zero indicated that the rater covariances were statistically significant ($\chi^2_3 = 24.126$, $p < .001$). The three types of raters therefore became significantly more similar in their predicted score differences over time.

Finally, the figure shows that changes in raters’ predicted score differences were highly nonlinear. For example, very few raters’ score differences appeared to monotonically increase or decrease over time, suggesting that the scoring severity of each rater was very unstable from check to check.

**Discussion**

**Severity Drift**

Our results showed that, on average, raters’ levels of severity did not drift significantly over time (Table 2 and Figure 3). However, behind these average trends lay very different individual trends for different raters (Figure 3). Previous literature was divided on this topic: one study showed increased leniency over time (Lunz and Stahl, 1990), two studies showed increased severity (Congdon & McQueen, 2000; Pinot de Moira et al., 2002), and another study showed both positive and negative drift for a small proportion of their raters (Myford and Wolfe, 2009).

In terms of raters’ individual severity trends, we found that these trends significantly fanned in over time (Table 2 and Figure 3). This result indicated that raters became more homogenous the more essays they scored. Hoskens & Wilson (2001) argued that a drift toward the mean was caused in their study by feedback to raters on their performances. This causal explanation also is possible for our results, as feedback was given to raters following poor performance. An alternative explanation is that this increased homogeneity might have been due to a rater practice effect: raters simply may have become more accurate with practice.
Consistent with previous research, we detected significant within-rater variability in scoring severity over time (Congdon & McQueen, 2000; Harik et al., 2009; Hoskens & Wilson, 2001; Lamprianou, 2006). This result calls into question the extent to which rater severity effects are stable traits or unstable states (Figure 3). Little is known about what causes high levels of within-rater between-check variability. In our study, one possible cause was interaction effects between raters and essays. Another was that each check on raters was based upon a small sample of six essays and so the within-rater-between-check variability may well have been less had more essays been used.

Central Tendency

We found a central tendency to raters’ scoring; on average, raters overscored low quality essays and underscored high quality essays (Figure 2). This finding is consistent with the literature (Myford & Wolfe, 2009; Engelhard, 1994; Knoch et al., 2007; Eckes, 2005). There also was some suggestion that experienced and new raters’ scoring exhibited a greater central tendency than team leaders’ scoring, but this result was not significant.

Rater Experience

We found that the three different types of raters did not differ significantly from one another in terms of their scoring severity. Further, the experienced and new raters did not, on average, score significantly differently from the consensus scores assigned by the expert committee. However, the team leaders, on average, significantly overscored by half a point (out of a maximum of 30 points for the test). It is troubling that it was the team leaders (the most experienced group of raters) who scored most differently from the expert committee. Previous research (Eckes, 2008; Cumming, 1990; Wolfe et al., 1998) indicated that more experienced raters took into account factors that were not in the scoring rubric and it is possible that this also might have occurred here. It could be argued that some of these factors may be meaningful and should have been included in the scoring rubric. However, as they were not recognized, it simply is unacceptable for raters to implement their own criteria and to depart form the scoring rubric in this way. More research on this issue is needed to replicate the current findings and to investigate whether more experienced raters do indeed introduce their own criteria.

As discussed above, each rater group did not, on average, become significantly more or less severe over time. This result contrasts previous findings in the literature which showed that inexperienced raters tended to start out severe but, with training, became more accurate (Weigle, 1999). Myford and Wolfe (2009) called for research on the impact of rater experience on the stability of rater severity. Our findings from model 2 indicated that while rater severity was unstable for all three rater groups, there was more instability when raters are less experienced.

Limitations

Our research was a secondary data analysis of an operational data set and was not a designed study. As such, there were several limitations to the available data. First,
it would have been useful to have data on rater and student characteristics and also on the teams within which the raters were trained. It also would have been interesting to examine the scoring of raters whose employment was terminated, even though their data did not affect students’ final scores. Another point to note is that, in the live scoring of the national curriculum tests, raters were assigned essays in batches of whole schools. This certainly could have affected rater behavior and would be a very interesting avenue for future research, particularly given the high-stakes uses of these students’ scores for school accountability purposes.

Unfortunately, our data only related to the scores assigned to the check essays and not to students’ actual essays. There therefore are limitations as to the extent to which our results generalize to other contexts. Our analysis of rater effects was based on a single writing paper relating to a specific curriculum. Clearly, replication in other contexts would be required before firm conclusions can be drawn.

It also is important to realize that the raters knew that they were being monitored. The check essays which they scored were scanned and presented on-screen to them, whereas the actual essays in live scoring were presented on paper. Our findings can be generalized to the extent that rater behavior was the same in these two contexts. However, it was possible that the play-it-safe strategy (Wolfe et al., 2007) under monitoring conditions might have produced more pronounced central tendency effects, particularly in the less-experienced raters, than would have been observed in live scoring. If this were the case, our findings might generalize only to conditions under which raters know they are being monitored. If, for example, rating accuracy were lower for the check essays (due to the fact that the checks were obvious), it would suggest that a move to disguise the checks from the raters is strongly needed given that we found several significant results for different types of rater effects. One way to implement such a move simply would be to present all essays (live and check) on-screen. The frequency and quantity of check papers then can be increased to determine exactly what is happening in live scoring.

Acknowledgments

We are grateful for the very helpful comments provided by the current and the previous editor and the three referees.

Notes

1The data did not identify students. This prevented students’ scores from being summed across the two essays or from being analyzed jointly.

2Wald tests on the variance and covariance parameters are only approximate as the sampling distributions of these parameters are only approximately normal (Goldstein, 2010; Raudenbush & Bryk, 2002).

3These predictions were calculated by substituting sample estimates for $\beta_0$ and posterior estimates (or best linear unbiased predictions, empirical Bayes estimates, or shrunken estimates; see Goldstein, 2010; Raudenbush & Bryk, 2002) for the rater random effects $u_j$. The key property of posterior estimates is that they are shrunk toward the overall mean, making them more stable than simple average score differences, particularly when they are based on small samples of data as was the case here.
The 95% confidence intervals were calculated in the standard way: by assuming normality and adding or subtracting 1.96 times the standard error.

References


**Authors**

GEORGE LECKIE is a Lecturer in Social Statistics, Graduate School of Education, University of Bristol, 2 Priory Road, Bristol, BS8 1TX; g.leckie@bristol.ac.uk. His primary research interests include multilevel modeling.

JO-ANNE BAIRD is a Professor of Educational Assessment, Graduate School of Education, University of Bristol, 35 Berkeley Square, Bristol, BS8 1JA; jo-anne.baird@bristol.ac.uk. Her primary research interests include rater effects, standard-setting and assessment policy.