Monitoring Rater Performance Over Time: A Framework for Detecting Differential Accuracy and Differential Scale Category Use

Carol M. Myford
University of Illinois at Chicago
Edward W. Wolfe
Pearson

In this study, we describe a framework for monitoring rater performance over time. We present several statistical indices to identify raters whose standards drift and explain how to use those indices operationally. To illustrate the use of the framework, we analyzed rating data from the 2002 Advanced Placement English Literature and Composition examination, employing a multifaceted Rasch approach to determine whether raters exhibited evidence of two types of differential rater functioning over time (i.e., changes in levels of accuracy or scale category use). Some raters showed statistically significant changes in their levels of accuracy as the scoring progressed, while other raters displayed evidence of differential scale category use over time.

When students respond to complex performance tasks such as essay exams, human raters typically render judgments regarding the quality of student responses. Unfortunately, errors in rater judgment (rater effects) may influence the accuracy of the assigned ratings. Previous research has indicated that rater effects are common, that their influence may be diminished through rater training and monitoring efforts, and that multifaceted Rasch measurement models may provide useful information for detecting and understanding the nature of these effects (e.g., Engelhard & Myford, 2003; Wolfe, Chiu, & Myford, 2000). In the past, researchers have tended to portray rater effects as static characteristics of raters (i.e., as if a rater effect influences every student’s performance in precisely the same manner). However, an individual rater’s behavior may change over time.

The acronym DRIFT—differential rater functioning over time (Wolfe & Moulder, 1999)—refers to changes in rater performance as a scoring project proceeds. A number of studies have determined that raters’ severity levels change systematically over time (e.g., Braun, 1988; Hoskens & Wilson, 2001). That is, they exhibit differential severity. In some assessment programs, raters participate in ongoing training or are periodically recalibrated. If these activities are not effective, a rater’s accuracy could change over time (i.e., the rater exhibits differential accuracy). DRIFT can also influence the variability of ratings. As a scoring project progresses, some raters may use the central categories more frequently, exhibiting a gradual restriction in the range of categories they employ (i.e., exhibit differential scale category use). In our experience, it is more common for the variability of raters’ ratings to decrease over time than to increase.

Most large-scale assessment programs have systems in place for monitoring rater performance, but those systems typically focus only on the ratings each rater assigns...
during a particular time period. There is no attempt to transform these measures of static rater effects into measures of dynamic rater effects. Unfortunately, these systems provide little information to help supervisors identify raters whose standards may be drifting. This is potentially problematic because raters may appear to be functioning appropriately within a given time period even though they are actually changing their rating behavior significantly over time.

A General Framework for Monitoring Rater Performance Over Time

Monitoring rater performance over time requires that administrators of assessment programs make four key decisions that may influence data collection, analysis, and reporting procedures.

1. **Segmenting Rater Performance into Time Periods.** Assessment program administrators must consider the scoring project’s length and determine how best to segment it into time periods. For example, if a scoring project lasts only a day or a few hours, administrators might look for changes in rater performance from hour to hour. On the other hand, if a scoring project lasts for a week or more, capturing changes in rater performance across days may be appropriate. Once administrators have decided on the number of time periods, a method for documenting the particular time period in which raters assign each rating must be devised.

2. **Choosing a Strategy for Detecting and Measuring Changes in Rater Behavior.** There are two primary strategies for comparing raters’ performances. In either strategy, the analyst first computes statistical indices summarizing how each rater performed during each time period. The **baseline comparison** strategy uses one time point as the reference point. The analyst chooses one time period as the baseline and creates DRIFT indices to detect and measure changes in each rater’s behavior between that baseline time period and the remaining time periods. Alternatively, the **adjacent comparison** strategy compares a rater’s performance during a particular time period to performance during each adjacent time point by basing DRIFT indices on measuring changes between the adjacent time points.

3. **Selecting a Frame of Reference for Depicting DRIFT.** The choice of the frame of reference determines how the metric of the parameter estimates is established and how connectivity is established in the ratings. Connectivity problems arise when two raters do not jointly rate any performances, making it difficult to compare each rater’s performance to the other rater (Myford & Wolfe, 2004). One option for establishing a frame of reference is to depict rater performance in terms of the degree to which the ratings of a particular rater agree with the ratings that other raters assign (i.e., norm referenced to the other raters in that group)—an **internal frame of reference.** This frame of reference is based on the assumption that raters assign valid ratings, on average, and that individual ratings are free of rater effects when they agree with the ratings of other raters (Wolfe, 1998). Assessment programs must have multiple raters evaluate at least some performances to employ this frame of reference. However, if a program selects only a sample of performances to double score (to contain costs) and
only a subset of the raters is involved in the double scoring, one must verify that there are no disconnected subsets of raters.

By contrast, an external frame of reference depicts rater performance in terms of the degree to which the ratings of a particular rater agree with scores on an external criterion (e.g., an objectively scored examination or experts’ ratings of benchmark performances). This frame of reference is suitable for assessment programs that cannot afford to have more than one rater evaluate each performance. Rater connectivity is established through the students’ responses to the common set of external items or through the expert ratings of the student performances. This framework is based on the assumption that the rater’s ratings and the external criterion scores measure the same construct. Care must be taken to carefully select the external criterion because, if the scores on the external criterion are not highly correlated with the performance ratings, or if the average rating of the student performances is different for the external measure, then using scores on that external criterion may not be appropriate for interpreting rater effects.

4. Deciding Whether to Employ Anchoring when Scaling the Data. Assessment program administrators must also decide whether or not to employ anchoring (i.e., fixing some parameter estimates on values external to the rater’s ratings) when scaling the data. If an assessment program does not employ anchoring, an analyst can scale data using either internal or external frames of reference. However, an analyst can only apply anchoring to data obtained using an external frame of reference. A second consideration is whether to use all (or only parts) of the data to establish the scale. When an assessment program employs anchoring, the analyst uses only a subset of the data to establish the scale’s metric (i.e., to set the center and spread of the scale). The analyst then estimates the parameters for the remaining data to conform to those constraints. By contrast, if an assessment program decides not to employ anchoring, the analyst uses all the data to create the scale’s metric.

Multifaceted Rasch Measurement as an Approach for Detecting and Measuring Rater DRIFT

In a multifaceted Rasch measurement analysis, the mathematical models for detecting dynamic rater effects may take several forms (Linacre, 1989). One form is a multifaceted Rasch model containing a time facet (referred to hereafter as the Time Facet model),

\[ LN \left( \frac{\pi_{nrtx}}{\pi_{nrtx-1}} \right) = A_n - S_r - R_t - C_x, \]

where

- \( \pi_{nrtx} \) = the probability of the performance of student \( n \) receiving a rating of \( x \) from rater \( r \) at time \( t \),
- \( \pi_{nrtx-1} \) = the probability of the performance of student \( n \) receiving a rating of \( x - 1 \) from rater \( r \) at time \( t \).
$A_n =$ the level of achievement for student $n$,
$S_r =$ the severity of rater $r$,
$R_t =$ the deviation of ratings at time $t$ from the average of all ratings, and
$C_x =$ the difficulty of the threshold between scale categories $x$ and $x - 1$.

The *Time Facet* model depicts each rater’s severity as a static characteristic across time but corrects for shifts in the mean of the ratings across time (via $R_t$). The *Time Facet* model would be useful for detecting gross changes in the mean rating across all raters as time progresses, although changes in $R_t$, which depicts the overall average change in ratings between times, may be due to either differences in rater severity over time, or to differences in the overall achievement of students who are rated during each time period (which can be assumed to be negligible when performances are randomly assigned to time periods). The *Time Facet* model is appropriate for settings in which raters evaluate students’ responses to a single performance task, although one can expand the model to include an item (or task) location parameter for contexts in which students respond to more than one assessment item (or task). The *Time Facet* model can also be expressed as a probability function,

$$
\pi_{nrtx} = \frac{\exp \sum_{k=0}^{x} (A_n - S_r - R_t - C_k)}{\sum_{j=0}^{m} \exp \sum_{k=0}^{j} (A_n - S_r - R_t - C_k)},
$$

where

- $j, k =$ counting indices with the indicated ranges, and
- $m =$ the maximum value of the scale categories beginning with the value of zero.

A problem with the *Time Facet* model when applied to DRIFT analyses is that the model does not directly estimate each rater’s performance at each time period. It is also possible to reformulate the *Time Facet* model to calculate a separate estimate of each rater’s severity at each time point (referred to hereafter as the *Separate* model),

$$
\ln \left( \frac{\pi_{nrtx}}{\pi_{nrtx-1}} \right) = A_n - S_{rt} - C_x,
$$

where

- $A_n =$ the level of achievement for student $n$,
- $C_x =$ the difficulty of the threshold between scale categories $x$ and $x - 1$, and
- $S_{rt} =$ the severity of rater $r$ at time $t$.

In the *Separate* model, a unique severity parameter is estimated for each rater at each time point, allowing one to compare a rater’s severity estimate at a given time point to an estimate at any other time point. It is important to note that an analyst can
establish a common scale for all time points in the *Separate* model by treating each rater at each time period as a unique individual when scaling the data.

The *Facets* computer program (Linacre, 2008) estimates parameters for the *Time Facet* and *Separate* models. The program simultaneously but independently estimates parameters for each facet included in the model, and the joint calibration of facets makes it possible to measure rater severity on the same scale as student achievement. For each element of each facet, the analysis provides a measure in log-odds (logit) units, a standard error (information about the precision of that measure), and fit indices (information about how well the data fit the measurement model’s expectations). Descriptions of how to interpret these indices are provided elsewhere (Wolfe & Dobria, 2008).

**Detecting Static Rater Effects**

Previous research using a multifaceted Rasch measurement approach to examine rater effects has concentrated on detecting static effects. Most investigations of dynamic effects focus on differential severity, and several studies explain how analysts can detect these rating patterns using particular rater effect indices (Engelhard, 1994; Wolfe, 2004, 2005). Unfortunately, few studies have systematically evaluated the statistical power or error rates of these static indicators of rater effects (Wolfe et al., 2000; Wolfe, Moulder, & Myford, 2001). Prior research on static rater effects allows us to speculate about how DRIFT effects other than differential severity may manifest themselves in multifaceted Rasch analyses. In the remainder of this section, we introduce two indices that may be useful for identifying differential accuracy and scale category use.

The single rater-rest of the raters correlation (*SR-ROR*) \( r_{SR-ROR} \) (Myford & Wolfe, 2003), a generalization of the Pearson correlation, depicts accuracy as a static characteristic of rater performance within the adopted frame of reference. The *SR-ROR* is computed using the ratings that a single rater assigned and the ratings that the rest of the raters assigned. The *SR-ROR* coefficient summarizes the degree to which a particular rater’s rank ordering of students’ performances is consistent with the rest of the raters’ rank ordering of those students’ performances. If a rater ranks students’ performances randomly in comparison to the ratings of other raters, the value of *SR-ROR* will be close to .00, an indication of rater inaccuracy. Conversely, if a rater ranks students’ performances in a manner similar to that of other raters, the value of *SR-ROR* will be close to 1.00, an indication of rater accuracy. It is worth noting that the value of *SR-ROR* should not be influenced by static central tendency because central tendency results in an “accurate” rank ordering of student performances—the ratings are simply compressed about their mean.

An index useful for detecting central tendency is the correlation between the residuals and the model-based expected ratings for a particular rater, \( r_{res,exp} \) (Wolfe, 2004, 2005). At the simplest level, one can examine the residual of a rating that a particular rater \( r \) assigned to a particular student’s performance \( n \) and the associated model-based expected rating for that pair,

\[
R_{nr} = X_{nr} - E_{nr},
\]
where

\[ R_{nr} = \text{the residual}, \]
\[ X_{nr} = \text{the rating that rater } r \text{ assigned to the performance of student } n, \text{ and} \]
\[ E_{nr} = \text{the model-based expected rating of rater } r \text{ for the performance of student } n, \]

and

\[ E_{nr} = \sum_{k=0}^{m} k\pi_{nrk} \]

where \( \pi_{nrk} \) resembles the Time Facet model expressed as a probability function.

One can compute a residuals-expected ratings (R-ER) correlation \( (r_{res,exp}) \) for each rater as a measure of that rater’s tendency to overuse the central scale categories by first scaling the data to the relevant model, outputting model-to-data expected ratings and residuals, and then, for each rater, correlating the rater’s residuals and expected ratings. When a rater exhibits evidence of static central tendency, the rater’s ratings of high-achieving students’ performances would be lower than the ratings that the multifaceted Rasch model predicted (Wolfe, 2004, 2005). Conversely, the rater’s ratings of low-achieving students’ performances would be higher than the multifaceted Rasch model predicted values. As a result, the scatter plot of residuals (y-axis) and expected ratings (x-axis) should show a negative slope for a rater exhibiting central tendency (i.e., a negative value of \( r_{res,exp} \)). It is worth noting that, in the case of static inaccuracy, this correlation should be close to .00. Therefore, the R-ER correlation would be sensitive to a static central tendency effect but would not be influenced by a static inaccuracy effect.

**Detecting Dynamic Rater Effects**

It is possible to transform static rater effect indices into statistics sensitive to DRIFT effects by comparing the values obtained for a particular rater at two time points (e.g., one can conduct a statistical test to compare two SR-ROR coefficients). Recall that when a rater is highly accurate, the SR-ROR is high in comparison to a rater who is not as accurate. Therefore, a rater who becomes more accurate over time will produce ratings that will result in a higher SR-ROR, and a rater who becomes less accurate over time will produce ratings that will result in a lower SR-ROR. One can test the null hypothesis that two correlation coefficients are equal by standardizing the correlations via Fisher’s z transformation,

\[
Z_{SR-ROR} = \frac{LN (1 + r_{SR-ROR}) - LN (1 - r_{SR-ROR})}{2}.
\]

This transformation produces positive values of \( Z_{SR-ROR} \) when \( r_{SR-ROR} \) approaches 1.00, negative values of \( Z_{SR-ROR} \) when \( r_{SR-ROR} \) approaches −1.00, and values of \( Z_{SR-ROR} \) that approach .00 when \( r_{SR-ROR} \) approaches .00. One can then subject the transformed correlations to a hypothesis test to determine whether the difference is
Monitoring Rater Performance Over Time

statistically significant,

\[ Z_{SR-RORc,SR-RORb} = \frac{Z_{SR-RORc} - Z_{SR-RORb}}{\sqrt{\frac{1}{Nc-3} + \frac{1}{Nb-3}}} \]

where

\[ Z_{SR-RORc} = \text{the Fisher-transformed SR-ROR at time } c, \]
\[ Z_{SR-RORb} = \text{the Fisher-transformed SR-ROR at time } b, \]
\[ Nc = \text{the number of ratings upon which } Z_{SR-RORc} \text{ is based, and} \]
\[ Nb = \text{the number of ratings upon which } Z_{SR-RORb} \text{ is based.} \]

The value of this statistic can be compared to a standard normal distribution. Assuming that Time \( b \) occurs prior to Time \( c \), values greater than 1.96 indicate that the SR-ROR at comparison Time \( c \) is significantly higher (at the .05 level) than the correlation at Time \( b \), signaling increasing accuracy over time. Values less than \(-1.96\) indicate that the SR-ROR at Time \( c \) is significantly lower than the correlation at Time \( b \), signaling decreasing accuracy over time. Values close to \(.00\) indicate very little change in accuracy over time.

One can depict changes in the R-ER correlation over time to measure differential scale category use. Recall that the correlation between the residuals and the expected ratings is close to \(.00\) in the case of both accuracy and inaccuracy effects. By contrast, when a central tendency effect exists, the R-ER correlation is negative. Hence, a Fisher transformation of \( r_{res,exp} \) produces negative values for \( Z_{res,exp} \) when a rater’s ratings exhibit central tendency, values of \( Z_{res,exp} \) that approach \(.00\) when there is little evidence of central tendency, and positive values for \( Z_{res,exp} \) when the spread of the rater’s ratings is relatively large (i.e., the opposite of central tendency).

Applying this information to a statistical test like that shown in the prior model, we have

\[ Z_{res,exp_c, res,exp_b} = \frac{Z_{res,exp_c} - Z_{res,exp_b}}{\sqrt{\frac{1}{Nc-3} + \frac{1}{Nb-3}}} \]

where

\[ Z_{res,exp} = \text{the Fisher-transformed R-ER correlation at time } c, \]
\[ Z_{res,exp_b} = \text{the Fisher-transformed R-ER correlation at time } b, \]
\[ Nc = \text{the number of ratings upon which } Z_{res,exp} \text{ is based, and} \]
\[ Nb = \text{the number of ratings upon which } Z_{res,exp_b} \text{ is based.} \]

Comparing this statistic to a standard normal distribution and assuming Time \( b \) occurs prior to Time \( c \), values less than \(-1.96\) signal a decrease in the dispersion of ratings and a concomitant increase in the rater’s use of the central scale categories over time. By contrast, values greater than 1.96 indicate an increase in the dispersion of ratings over time and a concomitant decrease in the rater’s use of the central scale
categories over time. Values close to .00 would indicate very small changes in the dispersion of ratings over time.

To demonstrate how to detect and measure rater DRIFT using these indices, we analyzed data from the College Board’s 2002 Advanced Placement English Literature and Composition (AP ELC) Examination. We sought to determine whether raters exhibited changes in levels of accuracy and use of scale categories over time.¹ The research questions that framed this portion of our study were

1. Do AP ELC raters exhibit differential scale category use? If so, how prevalent is this DRIFT effect, at what point in the essay scoring does it occur, and is the change sustained or temporary?
2. Do AP ELC raters exhibit differential accuracy? If so, how prevalent is this DRIFT effect, at what point in the essay scoring does it occur, and is the change sustained or temporary?

Method

Participants

Raters. From the 649 raters involved in the 2002 operational AP ELC scoring, we selected 101 to participate in our study (16%). Table 1 summarizes demographic characteristics of both the rater population and our sample. As shown, our sample was reasonably representative of its population. About two-thirds were female, and a large majority of the raters were White. A little more than half were secondary teachers, and the average number of years of AP scoring experience was about four (slightly higher in the population than in our sample).

Students. From the 210,964 students who took the 2002 AP ELC examination, the 101 raters in our sample rated responses of 51,233 students (about 24%) to Question 1 (the prose prompt) of the AP ELC. Descriptive statistics indicated that the sample was reasonably representative of the population. About two-thirds were female, and a large majority of the students were White. In addition, the student sample and population performed comparably on the multiple-choice section of the examination.

Instrument

The 2002 AP ELC examination consisted of two sections. Section I contained 55 multiple-choice questions, and Section II consisted of three free-response questions. This study focused on the scoring of essays students wrote for the first free-response question. Students read a brief excerpt from a British novel and then wrote an essay to analyze how the author achieved a comic effect (College Entrance Examination Board [CEEB], 2002a).

Essay Scoring Process

Each rater read essays for only one question. The raters used a nine-point scale (referred to as “scoring guidelines” in the AP Program), assigning a single holistic rating (CEEB, 2002b). A single rater rated each student’s essay, except in the case of the benchmark essays used to monitor rater accuracy. During the first
morning, the raters reviewed the scoring guidelines and read sample essays that highly experienced raters had previously scored. The raters then assigned ratings to these training essays. The raters reviewed the ratings they assigned and, when there were disagreements, discussed their rationales for assigning their ratings. Raters repeated the process with additional folders of essays until each rater demonstrated appropriate application of the scoring guidelines. The Exam Scoring section of the online AP Research Technical Manual (CEEB, n.d.) provides a description of the processes of developing the scoring guidelines and training the raters. During the essay scoring, raters sat at tables to evaluate the students’ essays.

Collecting Additional Data to Facilitate Detection of Rater DRIFT

We established connectivity in our design by having raters in our sample rate a common set of essays. Prior to the AP ELC essay scoring, several ETS AP staff and AP ELC leaders (i.e., persons with experience in designing essay prompts, developing scoring guidelines, training and supervising raters, and rating essays for this program) met to select 28 essays used to create connectivity among the raters in our sample. The leaders identified clear examples of essays reflecting performance at each point on the nine-point AP scale and then provided consensus ratings for each essay. Trios of leaders first assigned ratings to each benchmark independently. Then the entire group of leaders discussed each benchmark essay individually and came to
a verbal agreement on a consensus rating. We made multiple, identical photocopies of those essays. We divided the 28 benchmark essays into 8 sets, each containing 7 essays and having at least 1 essay in common with each other set.

Beginning on the second day, each rater in our sample received a folder of benchmark essays to score twice per day for 4 days—one in the morning, and one in the afternoon (i.e., each time period in this scoring project consisted of one-half of a work day). The raters rated the same benchmark essays during each time period, writing their rating for each benchmark on the photocopied essay. They also recorded the date and time they completed each folder of essays during the 4-day data collection period. Other than distributing benchmarks and documenting time indices, the procedures our rater sample followed were the same as for the rater population.

**Analyses**

We used the baseline comparison strategy to detect and measure changes in rater behavior over time. Specifically, we compared each rater’s performance in subsequent time periods to their performance during the first time period (i.e., the morning of the first day of data collection). We employed an internal frame of reference for detecting changes in accuracy and scale category use, basing our analyses on only the ratings our rater sample assigned.

Using the *Separate Model* as implemented in *Facets* software (Linacre, 2003), we calculated a SR−ROR correlation, $r_{SR−ROR}$, and an R-ER correlation, $r_{res,exp}$, for each of the 101 raters for each of the 8 time periods. Next, we standardized the correlations via Fisher’s $z$ transformation. We then determined whether the differences between pairs of standardized correlation coefficients for an individual rater were statistically significant (at the .05 level) to detect changes in rater accuracy and in scale category use.

**Results**

Figures 1 and 2 display histograms of the marginal distributions (i.e., ignoring time period) of the values of $r_{res,exp}$ and $r_{SR−ROR}$, respectively. The $r_{res,exp}$ indices are somewhat negatively skewed, ranging from a low of $−.90$ to a high of $.59$ (Mean $= .03$, $SD = .18$). This indicates that a small percentage of raters exhibited a relatively large central tendency effect in their ratings (e.g., some raters had $r_{res,exp}$ values $< −.50$—evidence of fairly extreme compression of the distribution of ratings). The $r_{SR−ROR}$ indices are quite negatively skewed, ranging from a low of $.00$ to a high of $.99$ (Mean $= .84$, $SD = .14$). This indicates that a small percentage of raters exhibited a relatively large inaccuracy effect when compared to the remaining raters (e.g., some values of $r_{SR−ROR}$ were $< .50$).

Table 2 displays the mean and $SD$ of the $r_{res,exp}$, $Z_{res,exp−res,expb}$, $r_{SR−ROR}$ and $Z_{SR−RORc,SR−RORB}$ statistics for each time period. The average $r_{res,exp}$ index does not vary much across time, suggesting that there is not a pervasive trend toward differential scale category use among the raters. However, on the afternoon of the final day, there is some evidence of a slight increase in the variability of the ratings (i.e., an increase in the average value of $r_{res,exp}$ from $.03$ to $.06$). There is evidence of a practice effect, suggesting increasing accuracy over the first day and a half of
scoring (i.e., the average value of $r_{SR-ROR}$ increased from Time 1 to Time 2, and then again from Time 2 to Time 3). There also appears to be evidence of within-day fatigue effects (i.e., decreasing accuracy) during the second and third days (i.e., the average value of $r_{SR-ROR}$ decreased during the afternoons of these days—from Time 3
### TABLE 2

Descriptive Statistics for Differential Scale Category Use and Differential Accuracy Indices

<table>
<thead>
<tr>
<th>Statistic</th>
<th>1 vs 2</th>
<th>1 vs 3</th>
<th>1 vs 4</th>
<th>1 vs 5</th>
<th>1 vs 6</th>
<th>1 vs 7</th>
<th>1 vs 8</th>
</tr>
</thead>
<tbody>
<tr>
<td>$r_{res,exp}$</td>
<td>0.03</td>
<td>0.02</td>
<td>0.05</td>
<td>0.02</td>
<td>0.01</td>
<td>0.01</td>
<td>0.03</td>
</tr>
<tr>
<td>($0.17$)</td>
<td>($0.22$)</td>
<td>($0.25$)</td>
<td>($0.17$)</td>
<td>($0.19$)</td>
<td>($0.16$)</td>
<td>($0.14$)</td>
<td>($0.16$)</td>
</tr>
<tr>
<td>$Z_{res,exp} - res,exp_b$</td>
<td>0.00</td>
<td>0.14</td>
<td>-0.06</td>
<td>-0.11</td>
<td>-0.06</td>
<td>0.03</td>
<td>0.14</td>
</tr>
<tr>
<td>($1.02$)</td>
<td>($1.38$)</td>
<td>($1.11$)</td>
<td>($1.09$)</td>
<td>($1.02$)</td>
<td>($1.03$)</td>
<td>($1.22$)</td>
<td></td>
</tr>
<tr>
<td>$r_{SR-ROR}$</td>
<td>0.78</td>
<td>0.82</td>
<td>0.91</td>
<td>0.80</td>
<td>0.89</td>
<td>0.79</td>
<td>0.85</td>
</tr>
<tr>
<td>($0.17$)</td>
<td>($0.13$)</td>
<td>($0.07$)</td>
<td>($0.18$)</td>
<td>($0.11$)</td>
<td>($0.17$)</td>
<td>($0.11$)</td>
<td>($0.07$)</td>
</tr>
<tr>
<td>$Z_{SR-ROR_c,SR-ROR_b}$</td>
<td>0.13</td>
<td>0.74</td>
<td>0.14</td>
<td>0.54</td>
<td>0.05</td>
<td>0.31</td>
<td>0.67</td>
</tr>
<tr>
<td>($0.81$)</td>
<td>($0.90$)</td>
<td>($0.88$)</td>
<td>($0.75$)</td>
<td>($0.80$)</td>
<td>($0.81$)</td>
<td>($0.78$)</td>
<td></td>
</tr>
</tbody>
</table>

*Note:* The values shown in this table are the mean and the (standard deviation) for each index at each time period. The values of $Z_{res,exp} - res,exp_b$ and $Z_{SR-ROR_c,SR-ROR_b}$ depict change from the baseline (Time 1) time period.

### TABLE 3

DRIFT Flag Rates for Differential Scale Category Use and Accuracy

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Flag</th>
<th>1 vs. 2</th>
<th>1 vs. 3</th>
<th>1 vs. 4</th>
<th>1 vs. 5</th>
<th>1 vs. 6</th>
<th>1 vs. 7</th>
<th>1 vs. 8</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Z_{res,exp} - res,exp_b$</td>
<td>% Negative</td>
<td>4</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>3</td>
<td>3</td>
<td>7</td>
</tr>
<tr>
<td>% Positive</td>
<td>3</td>
<td>10</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>$Z_{SR-ROR_c,SR-ROR_b}$</td>
<td>% Negative</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>% Positive</td>
<td>1</td>
<td>8</td>
<td>4</td>
<td>3</td>
<td>1</td>
<td>3</td>
<td>5</td>
<td></td>
</tr>
</tbody>
</table>

*Note:* The values in the table indicate the percentage of indices that meet the criteria for statistical significance (at the .05 level) for each time-period comparison.

...performing these comparisons...
Table 4 displays the percentages of raters showing statistically significant changes in a DRIFT index across multiple time periods. Using these results, we can determine whether raters exhibiting evidence of DRIFT showed temporary or sustained changes over time. In all cases, the majority exhibited no evidence of sustained DRIFT. Most commonly, raters who exhibited evidence of statistically significant DRIFT showed change in their rating behavior for only one time-period comparison—the largest percentage flagged at more than one time period for a particular statistic is 4%. This suggests that when raters changed their rating behavior, they tended to do so temporarily. In the case of positive values of $Z_{SR-RORc,SR-RORb}$, this is unfortunate because it means that statistically significant increases in accuracy were short-lived. Fortunately, the same can be said for negative values of $Z_{SR-RORc,SR-RORb}$ (decreases in accuracy). Changes in use of the scale categories were more likely to be maintained over time—about 7% exhibited an increase or decrease in the value of $r_{res,exp}$ that continued beyond a single time period.

Table 5 summarizes the rating patterns for two raters who exhibited these sustained changes. Rater A exhibited a decrease in the R-ER correlation over time, suggesting increased use of the central scale categories as the scoring project progressed. For five of the seven comparisons with Time 1 (i.e., Time 2 through Time 6), the rater’s R-ER correlation is statistically significantly lower (i.e., $Z_{res,exp, res,expb} < -1.96$). However, the rater’s R-ER correlation returns to its near-zero value during Time 8. Notice that the R-ER correlations are somewhat large—about .49 on average for Time 2 through Time 6. The third row in this example shows how this DRIFT behavior affected the ratings that Rater A assigned. Over 85% were 4, 5, or 6 at Times 2 through 6. By contrast, only 59% fell into those three central categories at Time 1. This shift in scale category usage initially occurred during the first day, continued for the second and third days (during which Table Leaders dedicated considerable time and effort to monitoring and retraining raters), and then began to dissipate during the fourth day.

The second example, Rater B, presents a case in which the rater’s level of agreement with other raters increased over time (i.e., the value of the $Z_{SR-RORc,SR-RORb}$ index increased). At Time 1, this rater’s ratings showed a very low level of
## TABLE 5
Differential Scale Category Use and Accuracy Examples

<table>
<thead>
<tr>
<th>Pattern</th>
<th>Statistic</th>
<th>Time Period (Comparison)</th>
<th>1</th>
<th>vs. 2</th>
<th>vs. 3</th>
<th>vs. 4</th>
<th>vs. 5</th>
<th>vs. 6</th>
<th>vs. 7</th>
<th>vs. 8</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Rater A</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Increasing central tendency</td>
<td>$Z_{\text{res.exp} - \text{res.exp}}$</td>
<td>-1.98*</td>
<td>-3.85*</td>
<td>-2.20*</td>
<td>-2.46*</td>
<td>-2.52*</td>
<td>-1.23</td>
<td>-1.17</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$r_{\text{res.exp}}$</td>
<td>.12</td>
<td>-.46</td>
<td>-.71</td>
<td>-.42</td>
<td>-.41</td>
<td>-.43</td>
<td>-.16</td>
<td>.09</td>
<td></td>
</tr>
<tr>
<td></td>
<td>% ratings $\geq 4$ but $\leq 6$</td>
<td>59</td>
<td>90</td>
<td>94</td>
<td>91</td>
<td>88</td>
<td>90</td>
<td>66</td>
<td>71</td>
<td></td>
</tr>
<tr>
<td><strong>Rater B</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Increasing accuracy</td>
<td>$Z_{\text{SR-ROR}, \text{SR-ROR}}$</td>
<td>1.49</td>
<td>2.20*</td>
<td>1.99*</td>
<td>1.59</td>
<td>.10</td>
<td>2.49*</td>
<td>1.99*</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$r_{\text{SR-ROR}}$</td>
<td>.49</td>
<td>.92</td>
<td>.97</td>
<td>.96</td>
<td>.93</td>
<td>.54</td>
<td>.98</td>
<td>.96</td>
<td></td>
</tr>
</tbody>
</table>

*p < .05.
agreement with other raters’ ratings (i.e., $r_{SR-ROR} = .49$). The rater’s level of agreement with the other raters increased dramatically, and Rater B continued to show high agreement levels for six of the remaining seven time periods—note that four of these increases in the $Z_{SR\sim RORc,SR\sim RORb}$ indices were statistically significant (at the .05 level). In fact, the average value of $r_{SR\sim ROR}$ over the remaining seven time periods was .89. Only two of the $Z_{res,exp_c-res,exp_b}$ comparisons for Rater B are statistically significant (not shown in Table 5). It is likely that initially Rater B had difficulty applying the nine-point scale but learned to apply it in a consistent manner by the middle of the first day. There is also something suspicious about the sudden and temporary decline in performance at Time 6—perhaps for some reason Rater B had a bad afternoon on the third day, or maybe fatigue or boredom set in temporarily.

Finally, it is worth noting that in both cases shown in Table 5, supervisors of raters would reach very different decisions about the existence and impact of rater effects if they ignored these DRIFT effects and directed their attention solely at static depictions of rater effects. For example, on average, about 81% of Rater A’s ratings fell into the middle three scale categories. By comparison, 69% of the ratings that all raters in our study assigned were in these three categories. Comparatively speaking, the percentage of Rater A’s ratings in the central categories seems somewhat high but perhaps not alarming. However, during the time periods flagged for DRIFT, on average 91% of Rater A’s ratings fell into these three categories. Clearly, this distribution of ratings is markedly different from the distributions of ratings that other raters assigned. Similarly, if we had depicted rater accuracy in a static manner, Rater B’s overall pattern of ratings would raise no concerns—the average value of $r_{SR\sim ROR}$ (via a Fisher transformation) across the eight time periods is .92 for this rater. However, values of $r_{SR\sim ROR}$ for the individual time-period comparisons display much variability, from a low of .49 to a high of .96. Squaring these values indicates the proportion of variance that Rater B’s ratings shared with the ratings of other raters. The wide range (i.e., a low of .24 and a high of .92) between time periods for Rater B clearly indicates that this rater was not uniformly accurate across time periods.

Conclusions

Our results indicate that DRIFT effects relating to rating scale category use and accuracy do indeed exist in operational ratings, and that these effects can impact those ratings in nontrivial ways. We found that over time, AP ELC raters were more likely to show evidence of differential scale category use than differential accuracy. Raters who showed evidence of differential scale category use tended to show more variability in the ratings they assigned early in the scoring project than they did later. Those raters who exhibited differential accuracy tended to become more accurate over time, particularly during the first 2 days.

In the majority of cases, AP ELC raters exhibited little evidence of change over time in their levels of accuracy or in their scale category usage. Most commonly, raters who exhibited evidence of these changes did so for only one time-period comparison, suggesting that changes in accuracy or scale category use were most often temporary and that, by and large, these raters did not continue to drift over time. However, there were raters who showed evidence of sustained drift, which resulted
in marked changes in their rating behavior maintained across multiple time periods. Those raters more frequently showed changes in their usage of the scale categories than changes in their levels of accuracy.

It is important to point out that we used a single approach to measure differential scale category use and a single approach to measure differential accuracy. That is, we made decisions during data collection and analysis that influenced the approaches we could use to measure these effects. As a result, we were not able to explore other possible approaches for measuring differential scale category use and differential accuracy. In an earlier article (Wolfe et al., 2001), we proposed other indices that researchers could use to identify these same DRIFT effects: (a) the ratio of two mean-square rater fit statistics from two time periods as an indicator of differential accuracy and (b) the ratio of the variances of a rater’s ratings from two time periods as an indicator of differential scale category use. Future research should compare the statistical power and Type I error rates of all four of these statistics to determine which are most suitable for various contexts. Also, note that we scaled our data using the Separate model, but there are additional multifaceted Rasch models (i.e., the Time Facet model, among others) that researchers could employ, and statistical indices comparable to the ones we used could be developed.

Additionally, future research might focus on whether the use of different frames of reference for depicting rater performance alters conclusions regarding the prevalence of DRIFT effects in rating data. Other studies might explore the comparability of anchored and unanchored results. Still other studies might vary the length of the time periods under which DRIFT is examined to compare results from hourly, daily, or weekly depictions of DRIFT. Finally, future research might also seek to determine whether rater characteristics such as level of experience, amount of training or education, or cognitive styles are related to a rater’s tendency to exhibit DRIFT. This would be a particularly useful extension of the current work because, if such relationships exist, then knowledge of the nature of those relationships may help guide rater selection, training, and monitoring.

When considering our findings, it is important to keep in mind the study’s limitations. First, we studied rater performance in only one AP program, and within that program, the performance of raters who read essays written for only one essay question. Thus, one can generalize these findings to other scoring projects to the degree that the essay type, student population, rater population, scoring guidelines, and training and scoring processes are comparable. Second, we acknowledge that using photocopies of students’ handwritten essays may not have allowed us to monitor rater accuracy in the most convincing fashion—because the benchmark essays were visibly different from the other essays rated during the operational scoring, raters may have altered their behavior when rating those essays. The nature of paper-based scoring makes it difficult to introduce benchmark essays seamlessly and unobtrusively into this type of scoring process. Computer-based essay distribution makes this easier to carry out in an inconspicuous fashion because the appearance of a benchmark essay on a computer screen would be no different from the appearance of any other essay. We also acknowledge that in using only 28 benchmark essays that we bundled into sets of 7 and reused over 8 time periods, it is possible that some of the raters were able to remember ratings they previously assigned to at least
some of the papers. Given these methodological limitations, the rater accuracy indices we reported may have been inflated. In reality, raters’ levels of accuracy during the operational scoring may have been lower. Nevertheless, even under these less-than-optimal conditions, we identified some raters who drifted over time in terms of both accuracy and scale category usage. A challenge for researchers in the future will be to conduct quality control monitoring studies similar to ours but using ratings obtained from computer-based scoring projects to determine whether rater DRIFT is more (or less) prevalent in that context.

We believe that monitoring rater performance for evidence of DRIFT and taking corrective actions based on that information is an important step in the evolution of rater monitoring practices. Questions arise concerning when to take action, and what to do when such evidence exists. The Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999) recommend that assessment administrators monitor rater performance to determine whether there is evidence of DRIFT (Standard 3.23), and that they take reasonable steps to remove or reduce the impact of sources of irrelevant variance on reported scores (Standard 3.17). To implement these recommendations, administrators need to establish upper- and lower-control limits for statistical indices for use in detecting at least three rater effects—severity, scale category use (e.g., central tendency), and accuracy. The quality control limits will depend on the stakes associated with decisions that assessment users make from reported scores and will, therefore, be program-specific. Rater supervisors could then implement procedures to detect both static and dynamic rater effects using the established quality control limits. Raters showing evidence of either static or DRIFT effects would be targeted for retraining or removal from the scoring project, and supervisors would review the raters’ ratings to determine whether any of the performances those raters evaluated during flagged time periods should be rerated before issuing score reports. A critical challenge for those who monitor rater behavior is to find ways to identify in “real time” (i.e., while a scoring project is in progress) those few raters whose rating standards are drifting.

In the interests of fairness and objectivity, when raters score essays for high-stakes assessment programs, it is important that they apply the scoring guidelines in a consistent fashion, and that their rating behavior not drift over time. Unfortunately, few organizations presently engage in quality control activities for rater monitoring that would enable them to detect DRIFT effects if they were present. This study provides an example of how such monitoring could be carried out. Additionally, the study moves the quality control monitoring field forward methodologically in its exploration of ways to transform statistical indices for measuring static rater effects into indices for measuring dynamic rater effects.

Acknowledgment

Carol Myford and Edward Wolfe share joint first authorship on this article. The authors agreed upon the sequence of names.

This article is based on a related report, Monitoring Reader Performance and DRIFT in the AP English Literature and Composition Examination Using
Benchmark Essays (Wolfe, Myford, Engelhard, & Manalo, 2007). The research was supported by the College Board. Any opinions, findings, conclusions, or recommendations expressed are those of the authors and do not necessarily reflect the views of the College Board, the Educational Testing Service, Pearson, or the University of Illinois at Chicago.

We would like to acknowledge the helpful advice of Mike Linacre regarding the use of the Facets computer program to analyze the data. We are grateful to the readers of the AP ELC Exam and the AP Program’s administrative personnel, without whose cooperation this project could never have succeeded.

Notes

1As part of this investigation, we also looked at whether raters’ levels of severity changed over time. We present our findings from that investigation elsewhere (Myford & Wolfe, 2008; Wolfe, Myford, Engelhard, & Manalo, 2007).

2Prior to the meeting, we performed a power analysis to determine the minimum number of benchmark essays needed to achieve reasonable sensitivity to DRIFT effects. Based on the results from our analysis, we determined that 28 benchmark essays would be sufficient. We created sets of seven benchmark essays in order to ensure that we would have sufficient overlap of essays in our rating design so that we could connect raters across all eight time periods.

References


Authors

CAROL M. MYFORD is an Associate Professor of Educational Psychology at the University of Illinois at Chicago, 1040 W. Harrison St., MC 147, Chicago, IL 60607; cmyford@uic.edu. Her areas of specialization include scoring issues in large-scale performance and portfolio assessments, the detection and measurement of rater effects, scoring rubric design, and quality control monitoring of rater performance.

EDWARD W. WOLFE is a Senior Research Scientist, Assessment & Information, Pearson, 2510 N. Dodge St., Mailstop 125, Iowa City, IA 52245-9945; ed.wolfe@pearson.com. His primary research interests include psychometrics and test development.