Psychology Science Quarterly, Volume 50, 2008 (3), pp. 417-432

# A LLTM approach to the examination of teachers' ratings of classroom assessment tasks

KAREN DRANEY[1] & MARK WILSON

## Abstract

This paper investigates the use of a specific case of the Linear Logistic Test Model, known as the rating scale rater model, in which the item parameter is conceptualized to include an item difficulty parameter, plus a rating severity parameter. Using this model, the severity of groups of teachers is investigated when they scored sets of 321 pretests and posttests designed to be congruent with an embedded assessment system. The items were included in a linked design involving multiple booklets randomly allocated to students. Individual teachers were found to differ in overall severity, but also showed a reasonable amount of consistency within two of the three district moderation groups. Teachers also showed some mean differences between districts. There is also evidence that the model may be too tightly constrained, and further exploration using a less constrained model is indicated.

Key words: IRT, LLTM, rater effects, teacher effects

[1] Karen Draney, EAEDU - School of Education, 4323 Tolman, Berkeley, CA 94720, USA; email: kdraney@berkeley.edu

## Introduction

The linear logistic test model (LLTM; Fischer, 1983) has long been used as a tool in psychology to examine the effects of various properties on the difficulties of items. The LLTM decomposes the difficulties of specific items into linear combinations of elementary components. Although such a model is potentially very useful in education as well, it is used less often in educational contexts, and particularly in contexts such as classroom and large-scale assessment. This article demonstrates how the LLTM can be used to examine some of the assumptions behind teachers' use of a classroom assessment system.

In recent years, alternative forms of assessment have become more widely used, not only in large scale assessments, but at the classroom level as well. Classroom assessments based on various forms of embedded tasks, portfolios, and other complex tasks, are advocated to promote student learning, particularly the higher-order learning promoted in state and national standards (e.g. Wilson & Sloane, 2000; Black & Wiliam, 1998; Brown, Campione, Webber, & McGilly, 1992; Resnick & Resnick, 1992).

As such models of assessment are used more frequently within classrooms, teachers will need to become familiar and comfortable with the use of scoring guides, both to provide feedback to students and to generate data for grading and accountability purposes.

In contrast to the situation for trained raters, teachers often receive only a small amount of training in assessment, and undertake the assessment of student work with little or no feedback to assist them in developing their assessment skills (Wolf, Bixby, Glenn, & Gardner, 1991). Also in contrast to the situation for trained raters, whose behavior has been investigated in many studies (e.g. Saal, Downey, & Lahey, 1980; Braun, 1988; Lunz, Wright, & Linacre, 1990; Engelhard, 1996; Wolfe & Myford, 1997; Wilson & Case, 2000; Hoskens & Wilson, 2001) we know little about the consistency with which teachers apply the standards contained within a scoring guide to the work their students generate in the classroom.

Whenever raters are involved in scoring student responses to assessment tasks such as written response questions or performance tasks, a number of assumptions about rater behavior are made in the interpretation of the resulting test scores. However, we know that there are inevitably rater effects (Saal, Downey, & Lahey, 1980). These can take the form of inter-rater variability, such as different raters showing different levels of severity (e.g. Lunz, Wright, & Linacre, 1990; Engelhard, 1996); or intra-rater variability, sometimes known as "rater drift" (e.g. Hoskens & Wilson, 2001; Wilson & Case, 2000; Wolfe & Myford, 1997; Braun, 1988). Thus, the assumptions that we make about teacher use of scoring guides should be tested.

Studies of rater effects are generally performed within the context of large-scale testing. With the recent emphasis on the importance of classroom assessment and Assessment for Learning (AfL; Black, Harrison, Lee, Marshall, & Wiliam, 2004), and the resultant necessary increase in the use of scoring guides, the possibility of rater effects associated with classroom teachers cannot be overlooked. While the concerns about teacher rating effects on student scores from formative in-class assessments may not be very great (after all, there are many opportunities for such errors to be corrected during classroom work) the context in which teacher ratings need scrutiny is where the ratings are to be used as part of an accountability system (Wilson, 2004). Systems such as this have been used in Australia (Masters & Forster, 1996).

This study investigates the assumptions involved in the use of scoring guides for a group of middle school science teachers. The possibility that teachers who work together on a regular basis have severities that are more similar than do teachers who are from different locations and do not work together is investigated. Implications of rating severity differences for the use of the assessment system are explored.

## The BEAR Assessment System

The Berkeley Evaluation and Assessment Research (BEAR) Assessment System is an example of an assessment system that is based in large part on performance assessment, and also, on assessments that are embedded in the course, rather than administered as formal "tests". The BEAR assessment system is based on the ideas of developmental assessment (Wilson, 2005; Wolf, Bixby, Glenn, & Gardner, 1991). Elements of the system are based on four principles, described in detail in Wilson & Sloane (2000):

1. A developmental perspective on student learning;
2. A match between instruction and assessment;
3. Management by teachers;
4. Assessments that uphold standards of reliability and validity.

Central to the developmental perspective is the idea of *progress variables*, which are a major focus of instructional and assessment activities. A progress variable is an achievement continuum defined operationally by the assessment tasks in which students participate, and that can be used to chart student progress over time (Masters, Adams, & Wilson, 1990). It is the variables-based approach to assessment that facilitates the developmental perspective.

Progress variables are operationalized as *scoring guides*, which also embody the third principle above. All performance-style assessments related to the same variable are ideally scored with the same or similar scoring guides. Scoring guides are hierarchical in nature. A higher score represents a qualitatively better performance: not just more factual knowledge, but a deeper understanding. This too reflects the developmental perspective of the assessment system.

The BEAR Assessment System has been adopted as an integral part of a yearlong middle school science curriculum, *Issues, Evidence and You* (IEY), developed by the Science Education for Public Understanding Program (SEPUP). IEY is structured around assessments that are embedded throughout the teaching and learning activities that make up the course. These assessments allow student progress to be tracked throughout the year. A detailed description of the implementation of the BEAR Assessment System as implemented in IEY can be found in Roberts, Wilson, & Draney (1997).

The progress variables that are central to IEY are the following:

- Evidence and Tradeoffs (ET): Identifying objective, relevant scientific evidence, and evaluating the advantages and disadvantages of different possible solutions to a problem based on the evidence available.
- Designing and Conducting Investigations (DCI): Designing a scientific experiment to answer a question or solve a problem, selecting appropriate laboratory procedures to collect data, accurately recording and logically displaying data (e.g. in graphs and tables), and analyzing and interpreting the results of an experiment.

- Understanding Scientific Concepts (UC): Recognizing and applying relevant scientific concepts (e.g. threshold, measurement, properties of matter) to an investigation or problem solution.
- Communicating Scientific Information (CSI): Organizing and presenting results, arguments, and conclusions in a way that is free of technical errors and effectively communicates with the chosen audience.
- Group Interaction (GI): Developing time management skills, the ability to work together with teammates to complete a task (such as a lab experiment) and to share the work of an activity.

Each of these variables is composed of two to four sub-parts known as elements. For example, the Evidence and Tradeoffs variable is composed of two elements: Using Evidence, and Using Evidence to Make Tradeoffs. The scoring guide for this variable is shown in Figure 1.

| **Score** | ***Using Evidence:*** Response uses objective reason(s) based on relevant evidence to support choice. | ***Using Evidence to Make Tradeoffs:*** Response recognizes multiple perspectives of issue and explains each perspective using objective reasons, supported by evidence, in order to make choice. |
|---|---|---|
| **4** | Response accomplishes Level 3 AND goes beyond in some significant way, such as questioning or justifying the source, validity, and/or quantity of evidence. | Response accomplishes Level 3 AND goes beyond in some significant way, such as suggesting additional evidence beyond the activity that would further influence choices in specific ways, OR questioning the source, validity, and/or quantity of evidence & explaining how it influences choice. |
| **3** | Response provides major objective reasons AND supports each with relevant & accurate evidence. | Response discusses *at least two* perspectives of issue AND provides objective reasons, supported by relevant & accurate evidence, for each perspective. |
| **2** | Response provides *some* objective reasons AND some supporting evidence, BUT at least one reason is missing and/or part of the evidence is incomplete. | Response states at least one perspective of issue AND provides some objective reasons using some relevant evidence BUT reasons are incomplete and/or part of the evidence is missing; OR only one complete & accurate perspective has been provided. |
| **1** | Response provides only subjective reasons (opinions) for choice and/or uses inaccurate or irrelevant evidence from the activity. | Response states at least one perspective of issue BUT only provides subjective reasons and/or uses inaccurate or irrelevant evidence. |
| **0** | No response; illegible response; response offers no reasons AND no evidence to support choice made. | No response; illegible response; response lacks reasons AND offers no evidence to support decision made. |
| **X** | Student had no opportunity to respond. | |

**Figure 1:**
Evidence and Tradeoffs scoring guide

The functional definitions of the progress variables are contained in the scoring guides, which describe the kind of achievement needed to reach each scoring level on the elements of the progress variables. Teachers use the scoring guides to rate student performance into 5 ordered, qualitatively different categories, labeled 0 through 4. The scoring guides describe the kind of performance that can be expected from students at each of the performance levels. Although each scoring guide is specific to the variable for which it was developed, there is a common structure. Generally, a score of 0 indicates an off-task or missing response; a score of 1 indicates performance that is incorrect; a score of 2 indicates performance that is generally correct but missing something important; a score of 3 is complete and correct performance; and a score of 4 indicates performance that goes above and beyond what is asked of the student.

In addition to the assessments embedded in the day-to-day classroom activities in IEY, the curriculum also contains what are called *link tests*. Link tests are composed of items that are designed to measure one or more progress variables, but are also less directly tied to specific course content than are the embedded assessments. As for the embedded assessments, they are, following principle 2, designed to have features in common with central elements in the curriculum. They are intended to be given at major course transitions, and can also be used to compare student performance across multiple contexts (i.e. several different curricula based on similar progress variables). An example of an item from a link test (a link item) is shown in Figure 2.

You run the shipping department of a company that makes glass kitchenware. You must decide what material to use for packing the glass so that it does not break when shipped to stores. You have narrowed the field to three materials: shredded newspaper, Styrofoam® pellets, and cornstarch foam pellets. Styrofoam® springs back to its original shape when squeezed, but newspaper and cornstarch foam do not. Styrofoam® floats in water. Although Styrofoam® can be reused as a packing material, it will not break down in land fills. Newspaper can be recycled easily, and cornstarch easily dissolves in water.

Present the properties of the three materials in an organized way.

Based on this data:

- Discuss the advantages and disadvantages of each material.
- Which material would you use? Be sure to describe the trade-offs made in your decision.

Item measures both DCI (organizing data) and ET (using evidence, and using evidence to make tradeoffs) progress variables

**Figure 2:**
Sample link item

Link items were designed to measure four of the five progress variables. GI was not measured using link items, due to the group-level nature of the variable – as link tests are to be completed by the student individually, it was determined not to be practical to include link items associated with the GI variable.

## Data collection

The data for this paper were collected as part of a larger research project set up to examine the effects of various components of the BEAR Assessment System as implemented in IEY. As part of this study, the original set of link items (see Wilson, Roberts, Draney, & Sloane, 2000) were revised and enhanced. A total of 22 link items resulted, each designed to be scored on multiple variable/element combinations. A calibration study was then conducted for these items, for the purpose of determining the relative difficulty of all of the items, as well as to study the rater effects which are the subject of the current paper.

For the calibration study, twenty of the 22 link items were divided into 8 booklets, each containing 5 items (two of the link items were considered too content-specific for a pretest, and were used in the posttest only; these two items are excluded from the analyses in this paper). Each of the remaining 20 items appeared in 2 of the 5 booklets. Each of the 8 booklets was linked to 5 other booklets. This is further illustrated in Table 1. Items are labeled 1 through 20, and each item is represented by a row; an X appears in the column representing the booklets in which the item appeared.

**Table 1:**
Distribution of Items by Booklets

| | Booklet | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Item | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| 1 | X | | | | | | X | |
| 2 | X | X | | | | | | |
| 3 | X | | | | | X | | |
| 4 | X | | | X | | | | |
| 5 | | | | X | X | | | |
| 6 | | X | | | | | | X |
| 7 | | X | X | | | | | |
| 8 | X | | | | | | | X |
| 9 | | | | | X | | X | |
| 10 | | | X | | | X | | |
| 11 | | X | | X | | | | |
| 12 | | | X | | X | | | |
| 13 | | | | X | | X | | |
| 14 | | | | | X | | X | |
| 15 | | | | | | X | | X |
| 16 | | X | | | | | X | |
| 17 | | | X | | | | | X |
| 18 | | | | X | | | X | |
| 19 | | | | | X | | | X |
| 20 | | | X | | | X | | |

This elaborate linking design was constructed in order to make sure that there was enough overlapping data between all of the forms so that an IRT analysis could be used to put all of the items on a common scale. These forms were spiraled and administered at the beginning of the year, as a pretest, to a total of 321 students in three different centers in the United States (Los Angeles, California; Louisville, Kentucky; and Sand Springs, Oklahoma)[2].

Ten teachers participated in this study: three from Oklahoma, five from Kentucky, and two from Los Angeles. Each teacher administered a set of pretests to at least one of their classes that would be using the IEY curriculum during the 1998-99 school year, and scored the results on all of the associated variable/element combinations. In addition, each of these tests was scored by one of seven UC Berkeley project staff members trained in the use of the scoring guides, and a random sample (stratified on teacher and test form) of approximately 50% of the tests was scored by a second staff member. Thus, all of the tests were scored at least twice (by one teacher and one UCB staff member), and approximately one third to one half were scored three times (by one teacher and two UCB staff members).

As part of their participation in the study, the teachers in these centers participated in a process called *moderation*. Moderation is the process in which a local group of teachers come together to score and discuss student work. This process serves two purposes. First, it improves technical quality by encouraging discussion and agreement among teachers as to the appropriate score level for a particular student response (i.e., local consensus building to set standards of student performance). Second, it provides for teacher professional development in assessment (Roberts, Sloane, & Wilson, 1996). In local assessment moderation sessions, teachers discuss the scoring, interpretation, and use of student work, and make decisions regarding standards of performance and methods for reliably judging student work relative to those standards Further, this process provides a forum in which teachers may discuss common mistakes or difficult concepts that can then be addressed in subsequent instruction.

The current study took place during the third year of a four-year project funded by the US National Science Foundation. All of the teachers in this study had been involved for at least one year prior to the gathering of the calibration data. This involvement included attending a week-long Summer Institute in assessment the summer before each school year, using the SEPUP IEY curriculum with at least one classroom of students, scoring selected embedded assessments, administering pretests, link tests, and posttests to these students and scoring the results, and participating in assessment moderation with the other teachers in their center approximately once every six weeks. Thus, each teacher in the current study had participated in an assessment moderation session in his or her center at least six times in the year prior to the calibration study, and had been actively scoring student work for at least one academic year.

[2] The school names are kept anonymous, to assure privacy

## The rating scale rater model

The model that was used to estimate rater effects for the four variables measured by the link tests was a specific version of the LLTM model (Fischer, 1983) called the rating scale rater model, as described by Wilson & Case (2000), and similar to Linacre's multifacet Rasch model (Linacre, 1989). In this model, we envision the overall difficulty of each item, as taken by the students in a particular classroom, and scored by the teacher in that classroom, as decomposed into a difficulty for the item and a severity for the teacher. We can then test the assumption that each of the teachers is using the scoring guides consistently with all of the other teachers.

In addition, because great emphasis is placed in the IEY curriculum on consistent use of the scoring guides across many different items and activities, it is also assumed that the difficulty of achieving one score versus another (say, a 3 versus a 2) is not affected by the particular activity to which the score is being assigned.

Thus, this model may also be seen as extending the rating scale model given by Andrich (1978) to include a rater severity parameter. If $\theta$ is the person proficiency, $\delta_i$ is the difficulty of item *i*, $\tau_j$ is the step parameter describing the step from level *j-1* to level *j*, and $\lambda_k$ is the severity of rater *k*, then the probability that a person rated by rater k at level *j* to item *i* is given by:

$$P(X_{ik}=j)=\frac{\sum_{l=0}^{j}\exp(\theta-\delta_i-\tau_l-\lambda_k)}{\sum_{l=0}^{J_i}\exp(\theta-\delta_i-\tau_l-\lambda_k)}, \tag{1}$$

where $J_i$+1 is the number of response categories in item i. In this model, the concept of the "item" is generalized to include each rater x item combination as a new generalized "item."

Constraints on the various parameter sets are necessary for model identifiability. In the analyses to be discussed in this paper, rater severities were centered on zero, and the person distribution was assumed to have a mean of 0. Considered as a LLTM model, the new generalized item (i.e. item x rater combination) is modeled to have two facets – one for the original item and one for the rater.

The ConQuest software (Wu, Adams, Wilson & Haldane, 2007) was used in all analyses described in this paper. This software fits all of the models described below.

## Results

Of particular interest in this investigation is the use of the LLTM, and in particular the rater severity parameters, particularly those for the teachers. As a first step, overall model fit was investigated by comparing a model containing rater severity parameters with a rating scale model containing no additional rater parameters. As the rating scale model with rater parameters is a more general case of the rating scale model, the two models can be compared using the likelihood ratio test. This is done by taking the difference in the deviance for the two models, which approximately follows a chisquare distribution, with degrees of freedom equal to the number of additional parameters in the more general model (in this case, the sixteen extra parameters needed to estimate rater severities). For these two models, the difference in the deviances results in a $\chi^2$ of 916.2 with sixteen degress of freedom; this is statistically significant at the .05 level, indicating that we achieve significantly better model fit by including the rater parameters. We will therefore use this model in the remainder of the analyses to be performed in this paper.

Overall precision of person estimates can be given by the EAP/PV reliability coefficient (EAP/PV reliability is explained variance according to the estimated model divided by total person variance, and is provided by the ConQuest software). For this data, the EAP/PV reliability is .95, indicating that estimates of person proficiency have good precision.

Fit of the models can additionally be examined by using meansquare and *t*-statistics to examine the fit of the various sets of parameters (items, raters, and steps). Detailed discussion of the calculation of fit statistics is given in the manual for the ConQuest program (Wu, et al, 2007); however, a quick summary of these statistics will be helpful. Fit statistics are a summary of the degree to which actual responses to items deviate from their expected values (calculated using the estimated model parameters), summed across the various facets of the data (items, raters, etc.). These statistics may be expressed as *t*-values, which allow an approximate significance test (misfit is statistically significant if $t > 1.96$ or $t < -1.96$), or as mean-squares, which give a measure of effect size. As in statistics generally, *t*-statistics are dependent on sample size, while mean-square statistics are not. Large mean-square statistics are considered those which are less than 0.75, or greater than 1.3 (Wilson, 2005; Adams & Khoo, 1996). Mean-squares less than 0.75, and *t*-statistics less than –1.96, suggest that there is less variability than expected; *t*-statistics greater than 2, and mean-squares greater than 1.3, suggest that there is more variability than expected. The latter are generally considered more serious, or, at least the ones that should be attended to first.

Summaries for the fit statistics for the various sets of parameters (items, raters, and steps) are shown in Table 1, including the average mean-square statistic for that set of parameters, the number of parameters for which the *t*-statistics were out of range (either > 1.96, or < -1.96), and the total number of parameters estimated.

The fit statistics in this Table show that, with the exception of DCI, the fit is reasonably good for item and rater parameters. None of the sets of rater parameters, or of item parameters for ET, UC, or CSI, show more than one parameter with significant misfit. DCI has 13 item parameters (41%) showing significant misfit, of which 6 show negative misfit and 7 show positive misfit. This suggests that one might want to examine the DCI items more closely, to try to understand why the misfit is occurring.

The fit statistics for the step parameters tell another story. All of these fit statistics show significant misfit, and in all cases, the misfit is positive. This would seem to suggest that the scoring guides do not seem to be used consistently across items. Although identical scoring guides are being used for each item associated with a given element of a variable, and strong emphasis is placed on consistent use of the scoring guides, and on teacher and student interpretation of the scores regardless of the particular item, it might prove better to use a partial credit model augmented with rater parameters. This provides evidence that the assumption of consistent use and interpretation of scores is, at least in part, not justified for these items as scored by these teachers.

**Table 2:**
Fit results for the four progress variables

| | | Average mean-square | # parameters with t-statistics out of range | Total # parameters estimated |
|---|---|---|---|---|
| DCI | items | 1.01 | 13 | 32 |
| | raters | 1.02 | 1 | 16 |
| | steps | 1.27 | 3 | 3 |
| ET | items | 1.00 | 1 | 24 |
| | raters | 1.11 | 0 | 16 |
| | steps | 1.21 | 3 | 3 |
| UC | items | 1.02 | 1 | 22 |
| | raters | 0.99 | 1 | 16 |
| | steps | 1.29 | 3 | 3 |
| CSI | items | 1.01 | 0 | 40 |
| | raters | 0.93 | 0 | 16 |
| | steps | 1.46 | 3 | 3 |

### *Variability of rater severity by site*

Figure 3 shows sets of boxplots of rater severity, grouped by site, with a boxplot of rater severity for the UC Berkeley staff raters included for comparison.

These boxplots show the overall differences in severity between sites for the groups of teachers. The two teachers at the Los Angeles site were consistently more severe than the teachers at the other sites. The three teachers in Oklahoma and the five teachers in Kentucky were similar in median severity, although the Oklahoma teachers were usually slightly more severe. However, the five teachers in Kentucky showed a much wider range of severity; those in Oklahoma tended to be quite similar to one another. This is quite interesting, in that in contrast to the situation for the teachers in Los Angeles and in Oklahoma where the teachers were in the same schools, the teachers in Kentucky were in several different schools.

For comparison, five of the seven raters at UC Berkeley also tended to be quite similar to one another. One of the raters, Rater 7, tended to be severe (shown as an outlier on two of the four variables, and was also the most severe of the raters on the other two), and another, Rater 3, showed a similar pattern in the opposite direction (was an outlier in the lenient direction for three of the four variables, and was the most lenient rater on the fourth variable).

Figure 4 provides another look at the rating severities of teachers within each site. This figure shows the severity estimates, with a two-standard-error wide band, for all variables within a teacher (the order is CSI, DCI, ET, UC), organized by site. Although the standard error bands are rather wide, we can clearly see that the LA teachers are quite severe, and that

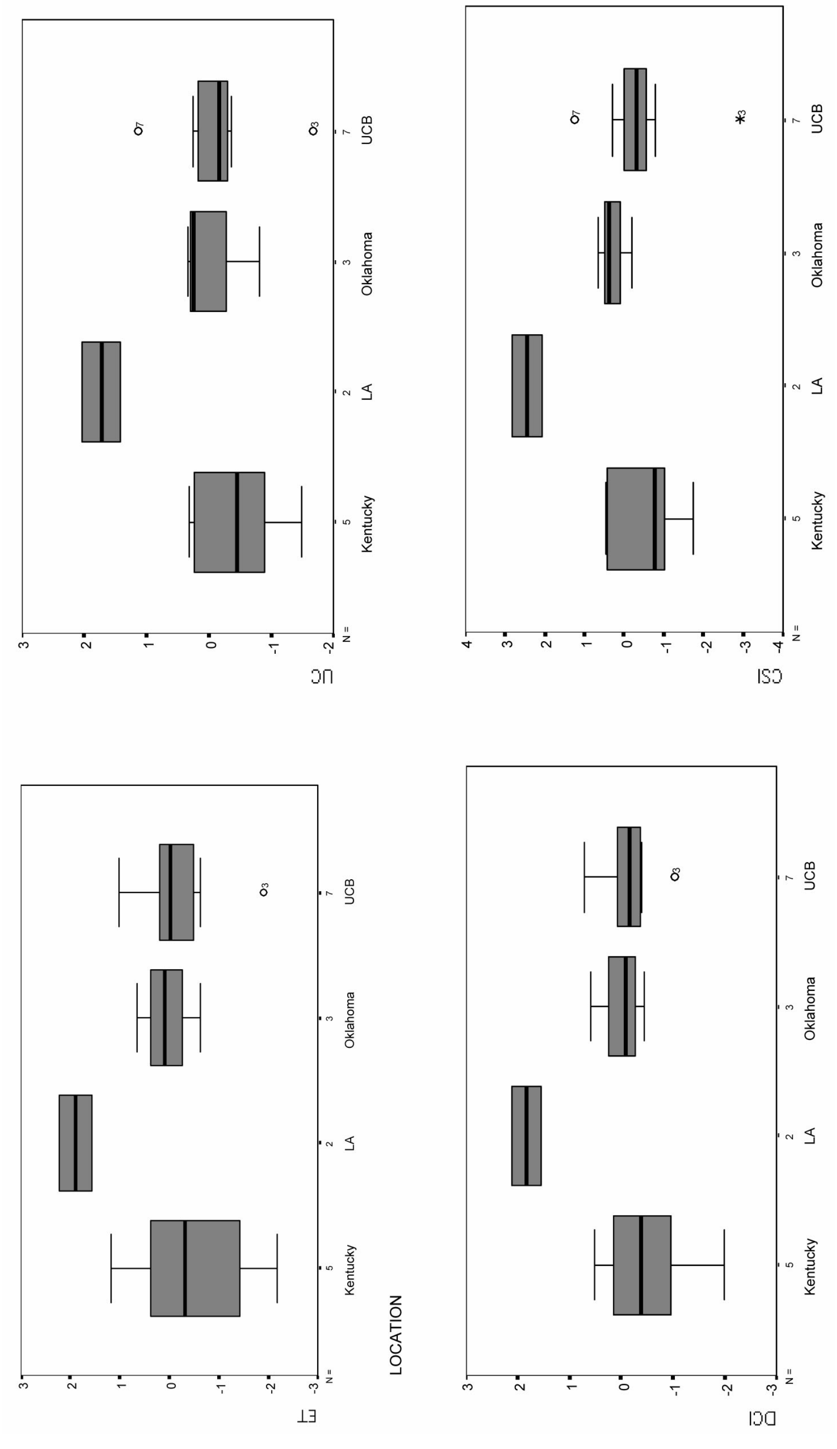


**Figure 3:**
Boxplots of rater severities by site

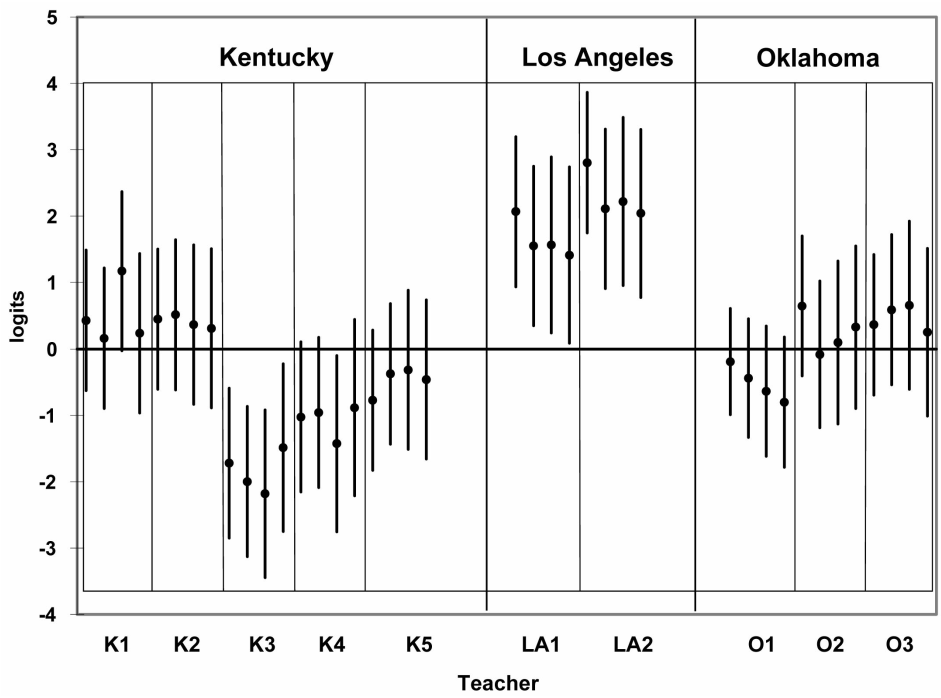
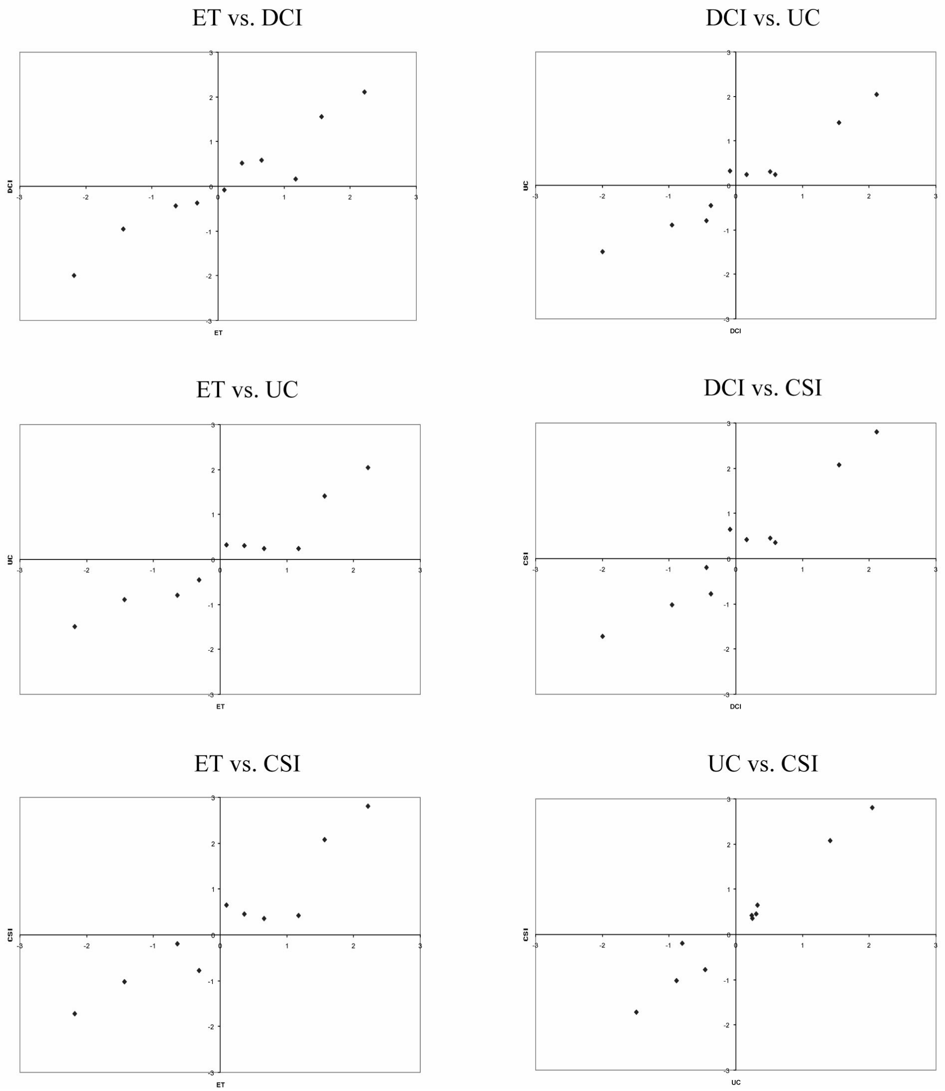


**Figure 4:**
Teacher Severities with standard errors
(Note: For each teacher, severities are shown in the following order: CSI, DCI, ET, UC)

the Oklahoma teachers all show similar severity, and are clustered near the average of zero. Interestingly, there appear to be two groups of teachers in Kentucky: three lenient teachers and two somewhat severe teachers – although these groups overlap quite a bit.

### *Consistency of teacher rater severity across variables*

Teacher severity appears to be quite consistent across variables. Table 1 gives the correlations between the estimated severities for the 10 teachers across the four variables (these analyses do not include the severity estimates for the project staff). The correlations are all very high, ranging from .94 (for ET with CSI) to .98 (for UC with CSI), indicating that the teachers are quite consistent in their scoring severity across the four variables. This is further evident in Figure 5, which shows scatterplots of the estimated teacher rating severities for each pair of variables. As is evident from the correlations, the patterns of severity for each pair of variables is roughly linear.


ET vs. DCI

DCI vs. UC

ET vs. UC

DCI vs. CSI

ET vs. CSI

UC vs. CSI

**Figure 5:**
Scatterplots of teacher severities for pairs of variables

**Table 3:**
Correlations between teacher severities for four variables

| | ET | DCI | UC |
|---|---|---|---|
| DCI | .96 | | |
| UC | .95 | .97 | |
| CSI | .94 | .96 | .98 |

*Comparison to multidimensional model*

As a final, exploratory step in the current investigation, a multidimensional rater model was fit to these data. Each of the four progress variables was assigned a separate dimension. As in the unidimensional model, each progress variable (i.e. each dimension) was assigned its own separate set of step and rater parameters. When comparing this model to the unidimensional rating scale rater model, the difference in the deviances is 505.9, and the number of additional parameters is 12; when compared to a chi-square distribution with twelve degrees of freedom, this is statistically significant at the .01 level, indicating that we achieve significantly better model fit by using a multidimensional model.

In addition, the multidimensional model provides disattenuated correlations between the dimensions (Adams, Wilson, & Wang, 1997). In this analysis, the correlations between the progress variables as represented by the dimensions are quite high, ranging from .87 (ET with DCI) to .99 (UC with CSI). In this case, the correlations can serve as an effect size for dimensionality; and although the multidimensional model is statistically different than the unidimensional model, the strength of the correlations indicates that there is likely little difference between the dimensions in practice.

## Discussion

This paper provides a case for the use of the LLTM in examining the assumptions involved in the use of classroom assessments by groups of teachers, and perhaps in controlling for situations in which those assumptions are not met – particularly in large-scale testing situations.

The analyses described in this paper have revealed several important characteristics of the rating behavior of the teachers involved in this study. First, as in most other situations involving ratings, there are inter-individual differences in overall rating severity between classroom teachers. In addition, there is evidence for lack of consistency in the use of scoring guides across the items in this pretest.

Second, there is notable consistency in severity differences across variable – those teachers who are most and least severe in their ratings tend to be so regardless of the variable and scoring guide being used.

In addition, there is significant consistency in the severity of teachers within a site. In this study, the LA teachers both showed similar high severities, while the Oklahoma teachers were also similar to one another, and neither particularly severe nor particularly lenient. Both of these sets of teachers worked in a single school. The Kentucky teachers, who worked in several different schools, showed the least consistency in their overall severities. Although all of the teachers in the study met together at the beginning of the year and participated in extensive discussions of the use of the scoring guides, it would seem that differences by site and by individual school remained.

Such within-site consistency could be based on several factors. As part of the larger study, all teachers within a site had been participating in moderation sessions, in which they met as a group, discussed student work, and attempted to come to consensus on the scores they assigned. This process is designed to produce consistency of scoring among the teachers working together, in part so that students within a site know clearly what is expected of them. In addition, the process is designed as a form of professional development for the teachers, so that teachers are more aware of their assessment practices, the reasons for the scores they assign to student work, and the teaching practices they use to improve student performance. It may be that this process was in some way more successful in the more consistent sites.

There may also be other reasons for the observed consistency in teacher severity. Teachers who worked within the same school may converse informally, and thus develop a shared set of expectations. The climate of the school or district in which they work may also influence their expectations. It would be interesting to study two groups of teachers in similar school settings, one group that engaged in regular assessment moderation sessions and one group that did not. One could then begin to ascertain the effect of moderation on teacher rating severity.

It would also be valuable to study teacher rating severities at several points in time, to determine whether and how teacher expectations change within a school year. It has been shown that in other contexts (e.g. Braun, 1988; Wolfe & Myford, 1997; Wilson & Case, 2000; Hoskens & Wilson, 2001) that rater severities change over a short time span (several hours to several days); such a study would provide us with the opportunity to observe potential changes over a longer span of time.

There are a number of extensions of the version of the LLTM used in this paper which might prove useful in additional analysis of data of this type. The most obvious extension would be to use the partial credit model instead of the rating scale model, to account for the inconsistency in the use of the scoring guides across items.

The second extension would be to add rater parameters not just for each progress variable, but by item and/or by step within item. We could thus examine whether raters are consistently severe across items, or in their use of particular levels of the scoring guide.

Finally, we could examine the use of a latent regression model such as that discussed by Adams, Wilson, & Wu (1997) to examine group effects (such as teacher) on student performance on the assessments, and other, more complex models, for both item and student group membership, such as those discussed in De Boeck & Wilson (2004).

## References

Adams, R. J., & Khoo, S. T. (1996). *Quest.* Melbourne, Australia: Australian Council for Educational Research.

Adams, R. J., Wilson, M., & Wang, W. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement*, 21, 1-23.

Adams, R. J., Wilson, M., & Wu, M. (1997). Multilevel item response models: An approach to errors in variables regression. *Journal of Educational and Behavioral Statistics, 22*, 47-76.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, 43, 561-573.

Braun, H. I. (1988). Understanding scoring reliability: Experiments in calibrating essay readers. *Journal of Educational Statistics*, 13.

Black, P., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2004). Working inside the black box: Assessment for learning in the classroom. *Phi Delta Kappan, 86*, 8-21.

Black, P., & Wiliam, D. (1998). *Inside the black box: raising standards through classroom assessment.* London: School of Education, King's College.

Brown, A. L., Campione, J.C., Webber, L. S., & McGilly, K. (1992). Interactive learning environments: A new look at assessment and instruction. In B. R. Gifford & M. C. O'Connor (Eds.), *Changing assessments* (pp. 121-212). Boston: Kluwer Academic Publishers.

De Boeck, P., & Wilson, M. (Eds.). (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. New York: Springer-Verlag.

Engelhard, G. J. (1996). Evaluating rater accuracy in performance assessments. *Journal of Educational Measurement, 33,* 56-70.

Fischer, G. (1983). Logistic latent trait models with linear constraints. *Psychometrika, 48*, 3-26.

Hoskens, M., & Wilson, M. (2001). Real-time feedback on rater drift in constructed-response items: an example from the Golden State examination. *Journal of Educational Measurement, 38,* 121-145

Linacre, J. M. (1989). *Multi-facet Rasch measurement.* Chicago: MESA Press.

Lunz, M., Wright, B. D., & Linacre, J. M. (1990). Measuring the impact of judge severity on examination scores. *Applied Measurement in Education, 3*, 331-345.

Masters, G. N., Adams, R. J., & Wilson, M. (1990). Charting of student progress. In T. Husen & T. N. Postlethwaite (Eds.), *International Encyclopedia of Education: Research and Studies.* Supplementary Volume 2 (pp. 628-634). Oxford: Pergamon Press.

Masters, G. N., & Forster, M. (1996). *Developmental assessment: Assessment resource kit*. Hawthorn, Australia: ACER Press.

Resnick, L. B., & Resnick, D. P. (1992). Assessing the thinking curriculum: New tools for educational reform. In B. R. Gifford & M. C. O'Connor (Eds.), *Changing assessments* (pp. 37-76). Boston: Kluwer Academic Publishers.

Roberts, L., Wilson, M., & Draney, K. (1997). *The SEPUP Assessment System: An overview*. University of California, Berkeley: BEAR Report Series, SA-97-1.

Roberts, L., Sloane, K., & Wilson, M. (1996). *Assessment Moderation: Supporting Teacher Change while Improving Professional Practice*. BEAR Report Series, SA-96-2, University of California, Berkeley.

Saal, F. E., Downey, R. G., & Lahey, M. A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin, 88*, 413-428.

Wilson, M. (2005). *Constructing measures: An item response modeling approach.* Mahwah, NJ: Lawrence Erlbaum Associates.

Wilson, M. (2004). Assessment, accountability, and the classroom: A community of judgment. In M. Wilson (Ed.), *Toward coherence between classroom assessment and accountability.* 103rd yearbook of the National Society for the Study of Education. Chicago, IL: The University of Chicago Press.

Wilson, M., Roberts, L., Draney, K., & Sloane, K. (2000). *SEPUP Assessment Resources Handbook.* Berkeley, CA: Berkeley Evaluation and Assessment Research Center.

Wilson, M., & Sloane, K. (2000). From principles to practice: An embedded assessment system. *Applied Measurement in Education, 13*, 181-208.

Wolf, D., Bixby, J., Glenn, J. III, & Gardner, H. (1991). To use their minds well: Investigating new forms of student assessment. *Review of Research in Education, 17*, 31-74.

Wolfe, E. W., & Myford, C. M. (1997). *Detecting rater effects with a multi-faceted rating scale model.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, Chicago, IL.

Wu, M., Adams, R.J., Wilson, M., & Haldane, S. (2007). ACER *ConQuest 2.0* [computer program]. Hawthorn, Australia: ACER.