

Effects of Different Training and Scoring Approaches on Human Constructed Response Scoring

Walter D. Way

Daisy Vickers

Paul Nichols

Pearson¹

Paper presented at the annual meeting of the National Council on Measurement in Education,
New York City, April 2008

¹ The authors are grateful to the North Carolina Department of Public Instruction and the Texas Education Agency for permission to utilize and report data related to their testing programs.

Abstract

This paper summarizes and discusses research studies related to the human scoring of constructed response items that have been conducted recently at a large scale testing company. These studies addressed approaches and procedures related to training scorers and the scoring process itself. Topics included image-based scoring, online scorer training approaches, the use of distributed scoring procedures, the use of annotations in constructed response item scoring, and training procedures associated with scoring constructed response items administered by computer. Although sometimes inconclusive and limited in scope, these studies add to the literature on constructed response scoring and are helpful in informing clients about issues important to their assessment programs.

Introduction

Human scoring of constructed response items is commonly accepted in large-scale assessment programs. In fact, the use of constructed response items has especially proliferated in statewide assessments in recent years as educators have looked for alternatives to the ubiquitous multiple-choice item type. Clearly, the processes required for humans to reliably score high quantities of constructed response items can be massive in scope. Our company, Pearson, maintains a total of 13 permanent performance scoring facilities throughout the country in addition to a 25,000 square foot facility at our home office in Iowa City, Iowa. Our regional scoring centers can house almost 4,000 human scorers, and from 2005 to 2007 we assigned approximately 90 million scores to student responses annually. Although perhaps not to the scale of Pearson, other testing vendors have established similarly formidable operations for constructed-response scoring.

Given such large constructed-response scoring undertakings, how do testing vendors maintain central measurement tenants such as reliability and validity? What sorts of research issues are these organizations addressing and what sorts of issues should they address? In this paper, we provide some answers to these questions based on our experiences working with state departments of education on large-scale assessments. We begin by reviewing what the testing standards say about scoring constructed response items. We then review a number of studies related to scoring constructed response items that have recently been conducted over the past several years at Pearson. Finally, we discuss additional studies that we believe are needed in the future.

Constructed Response Item Scoring: Standards and Processes

Skimming through the 15 chapters in the current *Standards for Educational and Psychological Testing* (AERA, APA, NCME, 1999), one will find 264 unique standards. Interestingly, few of these standards directly address constructed response item scoring, and the ones that do are not very prescriptive. Standard 2.10 calls for evidence on “both inter-rater consistency in scoring and within-examinee consistency over repeated

measurements”. Standard 3.14 addresses the need for documenting the criteria used for scoring extended-response items and Standards 3.22, 3.23, and 3.24 address the documentation of scoring procedures, processes for selecting, training, and qualifying scorers, and responsibilities for documenting procedures when constructed response scoring is done locally. Standard 5.9 calls for scoring rubrics and states that, “Adherence to established scoring criteria should be monitored and checked regularly. Monitoring procedures should be documented”. Taken in total, these standards say little about what level of inter-rater consistency is acceptable or what makes for good or bad constructed response scoring procedures and processes.

Nevertheless, test vendors and their clients pay a great deal of attention to the processes and procedures utilized for constructed response item scoring. Although specific constructed response scoring procedures vary slightly based on client conventions and requirements, certain components are universally addressed, including rubric development, rangefinding, scorer selection procedures, scorer training and qualification, and scorer monitoring.

Rubric development for constructed response items is done at the outset of a program. Depending upon the content and type of rubric (i.e., holistic or analytical) either a single rubric may be developed and applied to all items, or item-specific rubrics may be developed and in some cases, a holistic rubric may be developed for a content area with item specific rubrics for each item. Once constructed response items are developed and field-tested, rangefinding is carried out. Rangefinding is the process used to determine how to apply the rubric to student papers and therefore determines the standards that are used to score constructed response items. The process may also define both a range of response types or performance levels within a score point on the rubric and the threshold between score points. In other words, the rangefinding committee determines where one score point ends and another begins. There are aspects of rangefinding meetings that are similar to standard setting, that is, the rangefinding process defines the papers that are characteristic of the various score points represented by the rubric.

Rangefinding is generally conducted by a committee of content experts selected by the client. They look at a pool of responses which cover the range of score points for a particular item and through scoring and discussion come to a consensus score on each

response. The contractor makes notes from the discussion and uses these notes for annotating the responses to use to train scorers. The contractor and the client can be assured that the scorers are scoring based on the standards set by the committee when responses with consensus scores are used as anchor and training papers.

Scorers are recruited based on requirements developed in conjunction with clients. Depending on the client and the assessment, specific educational and experience requirements may have to be met (e.g., college degree, experience as a classroom teacher, etc.). Scorers are trained using comprehensive training materials developed by the test vendor's scoring experts and that are typically approved by the client. In most large scale constructed response scoring operations, scorers must pass a qualifying test for the prompt they will score.

Once the scoring process has begun, scorers are monitored in a variety of ways. In most programs, some sample of constructed responses (if not all responses) are scored twice, providing the basis for a number of statistics related to inter-rater reliability (e.g., perfect agreement, perfect plus adjacent agreement, spearman correlations, kappa statistics, etc.). In addition, papers previously scored by experts are distributed to scorers and form the basis for validity indices that are similar to the inter-rater reliability statistics. Validity papers may be specifically chosen because of certain features that can test, for example, whether scorers are consistently applying the consensus-based logic to borderline papers. Finally, scorers may be monitored using back-reading, whereby a project leader will re-score papers from a certain scorer or scorers that are performing at a marginal level of reliability. Through this process, the project leader can provide specific feedback or additional training in real time.

The scorer monitoring approaches described above have been significantly enhanced by technology, such as image-based scoring systems and electronic real-time capture of rater scores. The use of technology in the scoring process has led to changes in the way constructed response scoring occurs, and for Pearson has been the focus of a number of research studies. We summarize and discuss several of these studies in the next section of this paper.

Constructed Response Scoring Research at Pearson

Pearson has invested heavily in technology related to large scale constructed response item scoring, and in many cases technology has affected the scoring processes and procedures used. To implement many of these changes, it has been necessary to conduct research to evaluate their impact on the reliability and validity of the scoring process. Topics of study have included image-based scoring, online scorer training approaches, the use of distributed scoring procedures, the use of annotations in constructed response item scoring, and the training procedures associated with scoring constructed response items administered by computer. In the sections below, we summarize and discuss results from several studies addressing these topics.

Image-Based Scoring

Image-based scoring was a significant development in the scoring of constructed response items in large scale assessments. In these systems, image technology captures student responses (grayscale or bitone images) electronically. Most image-based scoring systems were developed in the late 1990s. Pearson's current system has been in place since 2000 and has been instrumental in streamlining and shortening scoring schedules.

The use of image technology includes the following features:

- Automatic routing of student responses to scorers based upon scorer qualification and project requirements and automatic, transparent routing of responses for second scoring, resolution, and validity reads. The image-based system can route responses so that only those scorers qualified for the item in a given stage of scoring will rate the item.
- Generation of online status reports that provide information on work that has been completed compared with how much is left to score. These reports significantly enhance the ability to complete the scoring within an assigned schedule.
- System capture of assigned scores, along with the scorer's ID; whether the score was a first, second or resolution score; and the time the score was assigned. This results in the entire scoring process being fully documented.

- A unique image ID allows the association of the online image with the actual student document. This makes it easy to retrieve original documents after scoring has been completed.
- Project leaders are able to “back-read” (i.e., re-read selected papers that have already been scored) online, and the system performs an automatic comparison of validity scores with “live” scores using online inter-reader reliability tools. This feature makes it possible to continuously monitor the accuracy of scoring at both group and individual levels. The real-time monitoring of scorers allows project leaders to identify and correct scoring trends before large volumes of responses have been scored.
- The image-based system can produce a full complement of on-demand online reliability reports throughout the scoring process.

The development of the image-based scoring system raised some concerns with Pearson’s clients about comparability with the traditional paper-based scoring processes that had previously been in place. In response to these concerns, Nichols (2002) carried out a study to compare image versus paper-based scoring procedures. One aspect of the study examined paper-based versus image based scoring. A second aspect compared two techniques for capturing images, gray-scale and bitone. Gray-scale permits the image to be represented in various shades, while bitone represents every pixel in the image as either black or white. Nichols found no differences between conditions in agreement rates, either among raters or between raters and expert scoring leaders. However, the image-based system resulted in a 15% reduction in the time needed to complete scoring compared with the traditional paper-based approach. Since that study, virtually all constructed response item scoring done by Pearson has transitioned to the image-based system.

Online Training and Distributed Scoring

The advent of image-based systems has resulted in related constructed response scoring innovations. One innovation has been online training. Traditional essay scoring procedures follow a stand-up, face-to-face approach. In this approach, the trainer typically leads a large group in classroom style training, or a group of trainers might each

deliver similar training to various smaller groups. In online, image-based training, standardized training material is delivered online to scorers. Vickers and Nichols (2005) compared online and standup training. They noted that online training has a number of advantages compared to conventional stand-up training:

- Client representatives can review and approve all training to which scorers will be exposed without having to observe training at a scoring site in person or interact with a particular project leader.
- Scorers can train at their own rate without having to wait for the slowest person in the room to complete each segment of the training and eventually qualify. This adds to the efficiency of training and is particularly effective when the scorer pool includes a combination of new and returning experienced scorers.
- Online trainees can receive immediate feedback on their performance on training sets, rather than waiting for scoring staff to manually tabulate results and give feedback to the entire group.
- A larger group of scorers can participate because the size of the training session need not be an issue. Scorers can train at a centralized site or they can train in their homes in a totally distributed model.
- Online training can be more flexible because it is not trainer-dependent. For example, additional scorers can be brought in once scoring is underway and still receive the same standardized training as the initial scorers.

Vickers and Nichols' (2005) study involved 63 raters scoring 35,534 essays written in response to a seventh-grade reading item with a four-point rubric for scoring. A group of 32 raters was randomly assigned to the online training condition and a group of 31 raters was randomly assigned to the standup training condition. Each group then scored the papers assigned to them until all papers were scored. Results indicated that the online and standup groups achieved comparable results in terms of reliability (measured by agreement of first and second raters) and validity (measured by agreement with expert raters on validity papers seeded into the scoring activity). In addition, the group trained online was able to score about 10 percent more responses than the standup group in the

same period of time, suggesting that online training may have promoted more efficient scoring as compared to the standup training.

Similar studies comparing online and face-to-face training were recently published by Elder, Barkhuizen, Knoch, and Von Randow (2007) and Knoch, Read, and Von Randow (2007). These investigators also found online training to be about as equally effective as face-to-face training, although they did report individual variations in receptiveness to the training. In a large scale setting, different scorer perceptions of online versus face-to-face training is certainly a factor to consider in developing online training. It is not unreasonable to assume that some scorers respond better to human interaction as they go through training. On the other hand, it is probably not efficient in a large scale scoring setting to allow scorers to choose how they prefer to be trained. One strategy to deal with this issue is to administer standup and online training presentations in as similar a fashion as possible and to provide scorers that train online with access to additional human-based feedback during the subsequent scoring activity.

A second innovation in constructed response scoring that has been made possible with image-based systems is distributed scoring. Distributed scoring allows raters to score constructed response items from any location, including their homes. In this model, the image-based system transmits student responses to the scorers via a secure website. As part of this model, an extended-hours scoring support center provides quality monitoring, scorer feedback and user technical support.

Although the advantages of distributed scoring in terms of flexibility and efficiency are obvious, some of Pearson's clients have been hesitant to implement distributed scoring because of concerns that accuracy of scoring may be reduced. Pearson recently conducted a study in conjunction with one of our clients to investigate this issue. Nichols, Kreiman, and Kanada (2006) reported results of this study, which compared scorers under three conditions. The first condition involved standup training and scoring in a common regionally-based location. The second condition involved online training and distributed scoring. The third condition involved online training and regional scoring. For each of two grades (4 and 10), a set of 25 writing responses for each of 17 final score points ranging from 4 to 20 were drawn at random from a recent administration of a large state-wide assessment and utilized for the study. This selection resulted in 425 writing

responses per grade. Under this design 20 readers in each condition (traditional, online training and distributed) scored the full set of 425 writing responses per grade. This resulted in a completely crossed design in which every scorer scored every paper. This resulted in a total of 60 raters at each grade equally distributed among the three training/scoring conditions. Each prompt was scored for both content and conventions. A final score was computed by combining the content and convention score so that the content score was given twice the weight as the convention score. Figure 1 summarizes the study design.

	Scorer	Training/Scoring Condition														
		Standup Training/					Online					Online Training				
		1	2	3	...	20	1	2	3	...	20	1	2	3	...	20
Paper	1															
	2															
	.															
	.															
	425															

Figure 1. Study design for grades 4 and 10.

Nichols, Kreiman and Kanada (2006) evaluated the results of the study in terms of inter-rater agreement, validity (i.e., agreement with an expert scorer), and two measures of association, Spearman rank correlations and Kappa coefficients. Their results were inconclusive in that some statistically significant differences between the online training/distributed scoring and online training/regional scoring conditions were found for grade 4 but not for grade 10. The conclusion of the paper stated,

Whereas there was a suggestion of a pattern in the data, this study found no consistent statistically significant differences in validity or reliability between distributed scoring and traditional local scoring. Because of these findings, any apparent pattern may be due simply to chance (p. 17).

Further analyses of the data collected in this study were reported by Kreiman (2007), who utilized the FACETS program (Linacre, 1989) and included the following scorer variables: training/scoring condition, years of experience, college major, and highest degree held. Results of this study indicated that all of these facets were found to contribute significantly to the severity of ratings received by students in both the grade 4 and grade 10 cohorts. However, there did

not seem to be clear interpretations of the findings and several limitations to the study were identified, including a caution about the original sampling of papers uniformly across total scores.

We chose to further examine the data from this study to better understand the trends in the findings. In our analyses we were interested in the differences between the ratings of the 20 scorers in each condition and the expert scorer involved in the study. Specifically, we calculated mean differences (bias), standard deviations of differences, and total root mean square errors, where:

$$Bias = \frac{1}{N_p} \sum_{i=1}^{N_p} (D_i),$$

Where $D_i = (\text{rater score for paper } i - \text{expert score for paper } i)$, N_p is the total number of papers,

$$SD_{DIFF} = \sqrt{\frac{1}{N_r} \sum_{i=1}^{N_r} (D_i - Bias)^2},$$

and

$$RMSE = \sqrt{Bias^2 + SD_{DIFF}^2}.$$

Figures 2 and 3 present bivariate plots of scorers' standard deviations of differences against their bias values. In each figure, the upper graph is for the content scores (on a 0-4 scale) and the lower graph is for the conventions score (on a 0-2 scale).

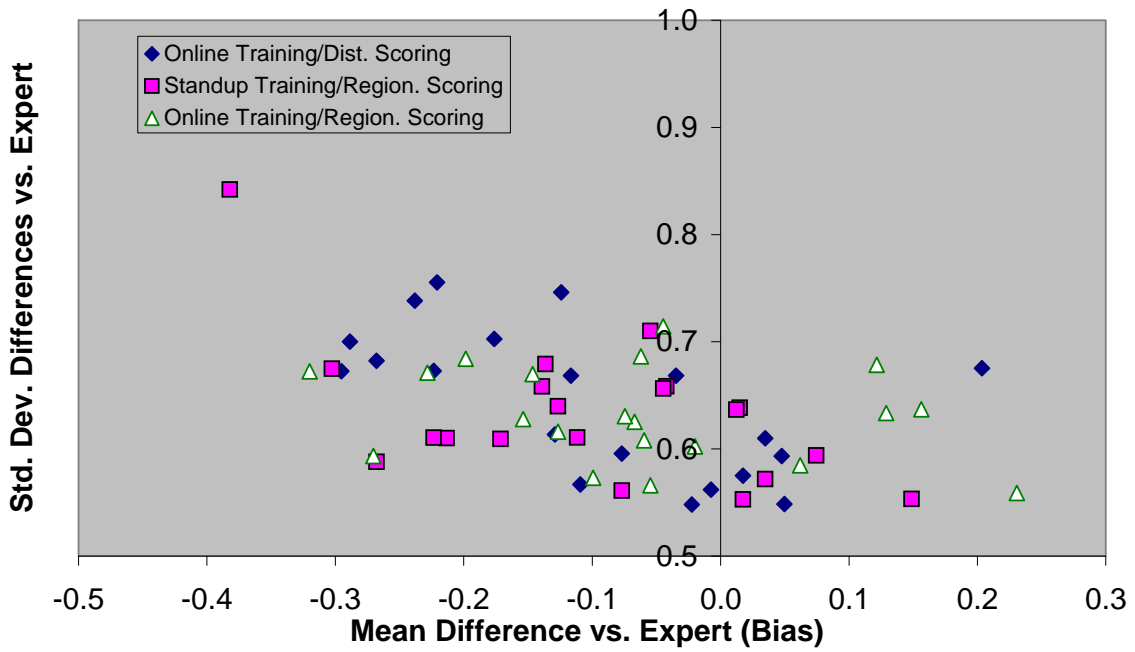
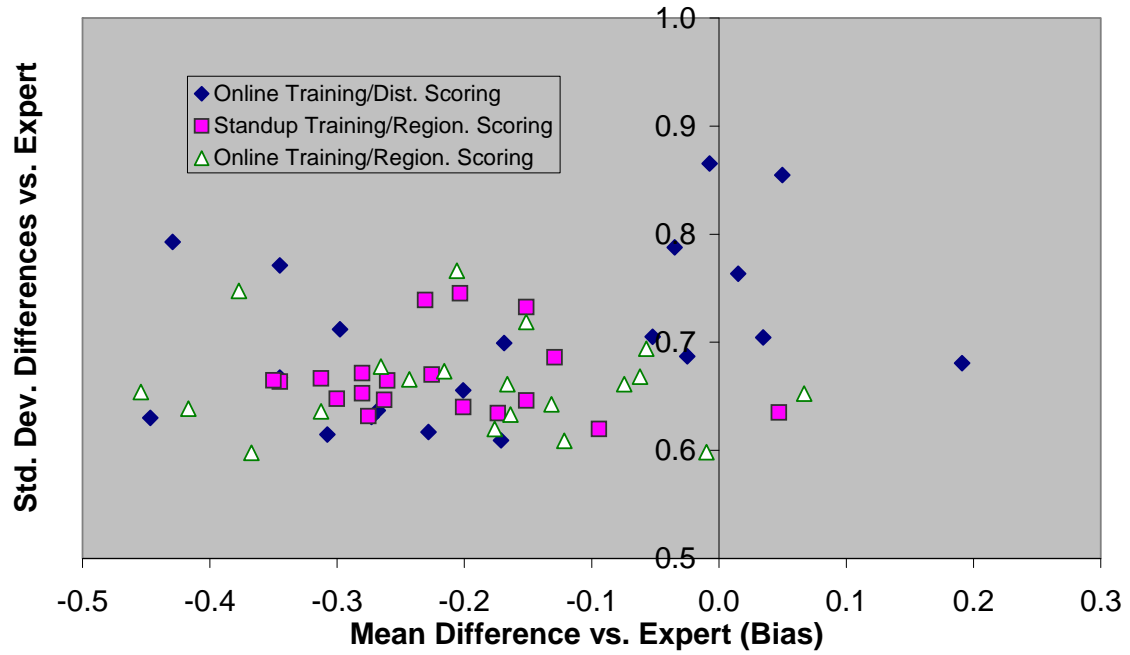


Figure 2. Grade 4 Scorers' Standard Deviations of Differences plotted against Bias for Content Scores (Upper Graph) and Conventions Scores (Lower Graph)

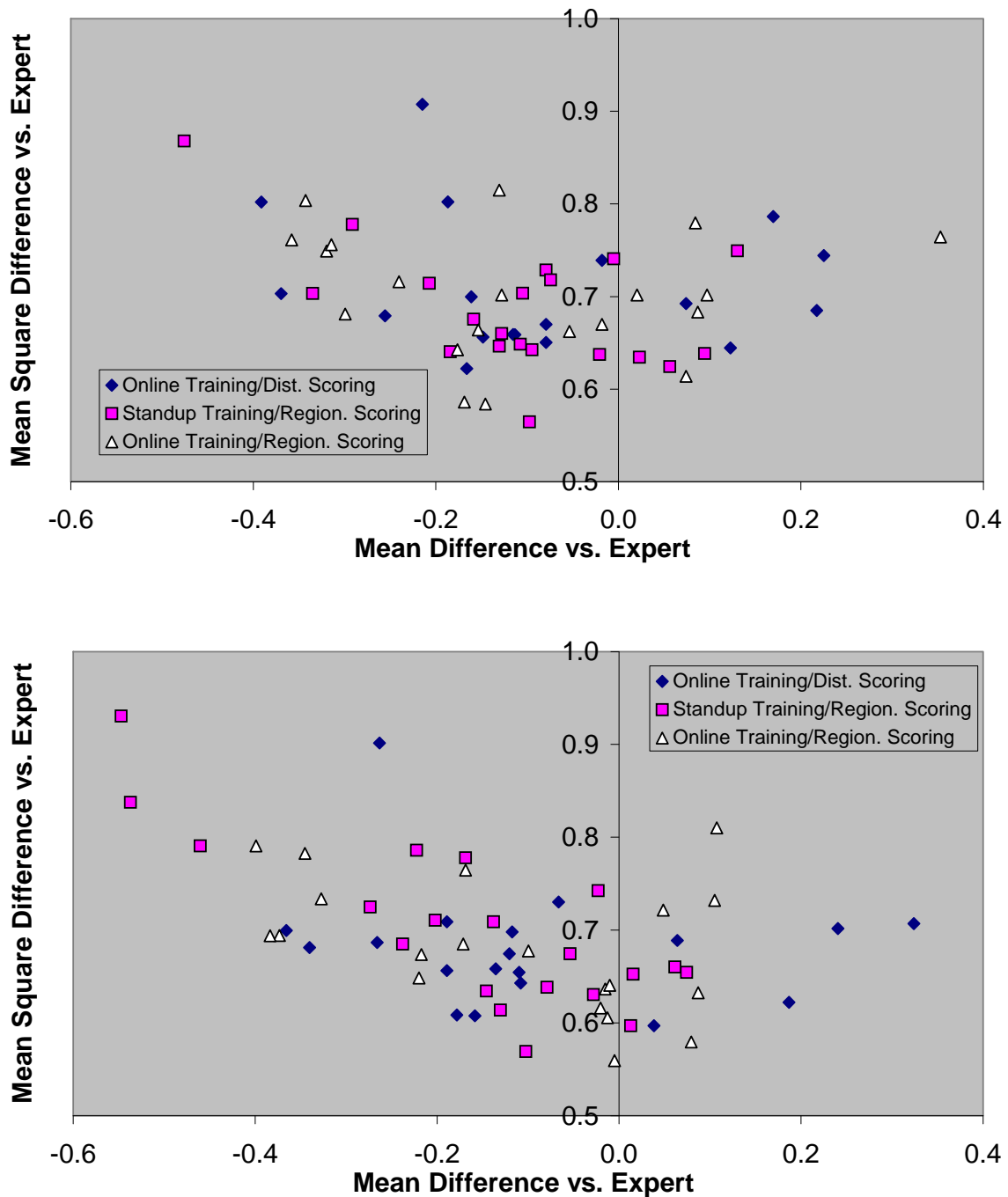


Figure 3. Grade 10 Scorers' Standard Deviations of Differences plotted against Bias for Content Scores (Upper Graph) and Conventions Scores (Lower Graph)

These plots indicate a negative bias (i.e., a tendency for the scorers to rate the essays more harshly than the expert) under all conditions of the study. This pattern was likely due to two reasons. One reason is the uniform distribution of scores in the papers

drawn for the study. This distribution was artificial and represented a much higher mean than the natural distribution of scores in the full sample of papers. Thus, it is likely that regression to the mean effects contributed to the negative bias. A second reason was that the monitoring process (e.g., validity checks, back-reading, etc.) that typically occurs in operational scoring was suspended for the study. This was done because it was thought that the monitoring process would contaminate the comparison across conditions. In general, the patterns of scorer error are very similar across the three conditions of study.

Table 1 summarizes the means of the bias, standard deviation of differences, and RMSE values across all scorers in each condition. The online training/distributed scoring condition had the lowest overall mean scorer bias compared with the other two conditions across all scores and grades. However, the online training/distributed scoring condition had the highest mean standard deviations of differences and overall mean RMSEs, with the exception of the conventions score at grade 10.

Table 1. Mean Scorer Bias, Standard Deviation of Differences, and RMSE values by Condition

Grade 4 – Content Score				Grade 4 – Conventions Score			
Condition	Bias	SD DIF	RMSE	Condition	Bias	SD DIF	RMSE
OT/DS	-0.17	0.70	0.74	OT/DS	-0.10	0.64	0.66
ST/RS	-0.23	0.67	0.71	ST/RS	-0.10	0.63	0.65
OT/RS	-0.20	0.66	0.70	OT/RS	-0.06	0.63	0.65
Grade 10 – Content Score				Grade 10 – Conventions Score			
Condition	Bias	SD DIF	RMSE	Condition	Bias	SD DIF	RMSE
OT/DS	-0.05	0.69	0.73	OT/DS	-0.09	0.65	0.68
ST/RS	-0.11	0.66	0.69	ST/RS	-0.16	0.66	0.70
OT/RS	-0.11	0.67	0.70	OT/RS	-0.12	0.65	0.68

The analyses summarized in Table 1 and Figures 2 and 3 provide some additional detail about the results of the study, and support the conclusion from the original study by Nichols, Kreiman and Kanada (2006). Because Pearson considers the use of distributed scoring to be an important operational enhancement to our constructed response scoring services, we are planning additional research to compare it with the more traditional regional scoring approach.

Use of Annotations in Constructed Response Scoring

In many large scale assessment programs involving constructed response item scoring, annotations are used to help illustrate the reasons why a particular paper received a particular score. Generally, annotations are used both in scorer training and as part of disclosed responses that are made available to the public. However, Pearson has had some conversations with our clients about incorporating annotations into constructed response scoring. The advantage of including annotations is to provide students and educators with information about performance in addition to the single reported score that is typically reported for constructed response items that are part of an assessment.

Nichols (2005) reported on a study done at Pearson to investigate the use of annotations in the scoring. He investigated both the impact of training on assigning annotations using holistic scoring and the effect of assigning annotations during holistic scoring. The study used a set of approximately 600 essays from an assessment administered to high school students. The essays were written in response to one of two narrative prompts that will be referred to as Prompt 1 and Prompt 2. Possible scores ranged from 1 to six and non-scorable and off-topic essays were excluded from the study. Specific annotations were developed for each score point based on three major criteria: complexity of thought, substantiality of development, and facility with language. There were up to nine annotations per score point from which the reader selected.

The results of the study indicated that training readers on annotations had no meaningful impact on the reliability, validity or rate of essay scoring. In addition, applying annotations during scoring had no meaningful impact on the reliability or validity of essay scoring. However, as might be expected, applying annotations had a large impact on rate of human scoring, increasing scoring time by 42 percent. In a large-scale setting, this study suggests that the time and cost implications to including annotations as part of essay scoring are too great to justify the interpretative benefit to students and educators.

One alternative for including more diagnostic information in the feedback associated with human constructed response scoring in the future might be to utilize automated scoring procedures. Pearson, as well as most large scale test vendors, now

offer automated scoring options that are supported by a growing body of research (c.f., Phillips, 2007; Dilki, 2006). However, these methods are limited in that they can only be applied to constructed responses from computer-based assessments.

Human Scoring of Online and Paper Constructed Responses

The research literature suggests a tendency for raters to rate typed essays more stringently than handwritten essays (Breland, Lee, & Muraki, 2005; Yu, Livingston, Larkin & Bonet, 2004; Hollenbeck, Tindal, Stieber, & Harniss, 1999; Powers, Fowles, Farnum, & Ramsey, 1994; Sweedler-Brown, 1991). Breland, Lee and Muraki (2005) speculated that the most plausible explanation for these findings seems to be that typed essays are more likely to be perceived by scorers as final drafts, and therefore expectations are slightly higher and errors in grammar or spelling are judged more harshly than they are in handwritten essays.

In many studies, comparisons of students responding to constructed response items by computer versus by paper-and-pencil have been confounded by potential rater effects in rating typed essays. Pearson researchers recently reported results of such a study (Way & Fitzpatrick, 2006). In that study, a variety of approaches were taken to investigate reasons for lower performance for student testing online versus by paper-and-pencil on the essay portion of the grade 11 Texas Assessment of Knowledge and Skills (TAKS) in English language arts. These included ANCOVAs using multiple-choice item performance and responses to survey questions as covariates as well as analyses of online essay responses and typed versions of paper-based responses using automated essay scoring technology. Results of analyses comparing text characteristics (e.g., word count, sentence count, words per sentence, readability measures, and other linguistic measures) suggested some structural differences between the typed and handwritten essays. However, patterns in the results also suggested the possibility of rater biases in the scoring of the online essays.

Working with essays administered in North Carolina, Fan et al. (2007) conducted a study designed to disentangle mode and rater effects in evaluating the comparability of paper and online writing assessment by controlling for rater effects in one training condition. Specifically, the study compared paper essay performance with computer-

based essay performance under conditions involving conventional training versus modified training conditions developed by Pearson to explicitly address rater tendencies to be more stringent with typed essays. The modified training materials addressed several possible scoring biases typically associated with typed responses:

- Typed essays generally appear shorter than identical handwritten responses. Two pages of handwritten material may take up only a screen or less when the response is typed. Raters must be sensitive to these differences in response length.
- When essays are handwritten, scorers may give writers the benefit of the doubt if spelling or punctuation, for example, is not clear. Because errors may be more easily discerned in typewritten essays, raters must take care not to be unduly influenced by the more visible errors.
- Errors made more evident by typing usually represent only one aspect of the scoring rubric -- that which relates to grammar, usage, and mechanics -- and raters should be certain to evaluate essays based on all criteria on all aspects of the scoring guide.
- Just as the style, tidiness, size, or any other characteristic of student handwriting can inappropriately influence the scores assigned by raters, the greater ease of reading typed responses can affect impressions of student essays. Raters must be constantly reminded to evaluate student responses based on the anchor papers and scoring guides rather than the ease or difficulty of reading responses.

The results from ANCOVAs (using multiple choice scores as covariates) and t-tests did not indicate a clear pattern of results. For some comparisons, statistically significant differences in writing scores favored the paper-and-pencil mode. However, effect sizes for these statistically significant results were small. Study results also supported the hypothesis that modified training reduced or eliminated the rater effect and so minimized lack of comparability across the paper-and-pencil and online modes. In general, differences between paper-and-pencil and online scores were smaller under the modified training condition compared to the conventional training condition.

Future Constructed Response Scoring Research

There are almost an unlimited number of studies related to the human scoring of constructed response items that can be imagined and considered. Because most of our constructed response scoring is for custom projects involving client-owned assessments, our research priorities tend to focus on questions that come up in our client-based work. Our list of high-priority studies for the future includes the following:

- An additional distributed scoring study. As previously mentioned, we plan to conduct an additional comparability study of online versus standup training and regional versus distributed scoring in both reading and writing content areas. The design of the new study will be informed by results the study reported in this paper, especially with regard to considerations such as the sample of papers to be scored and the use of monitoring procedures during the course of the study.
- Reader background study. We are interested in conducting a study to determine the effects of readers' level of education, content of degree program, and scoring experience on the scoring of specific content-based assessments. One question we hope to address in this study has to do with transfer of scoring expertise across content domains. We have gathered some limited data suggesting that experienced and expert constructed response scorers in one content area may not necessarily be effective in another content area. Although obtaining solid research data on this question will be difficult, we hope that some of it can be obtained from mining our scorer databases.
- Portfolio annotation study. In one program that we have recently become involved with, portfolio entries are scored using extensive written annotations. Although Nichols' (2005) study indicates that the including annotations or not in scoring has no effect (except in scoring efficiency), it is not clear how well these findings generalize to a more elaborate portfolio assessment. We are therefore interested in investigating the effect of removing the written annotations as part of scoring on the reliability, validity, and scoring efficiency of this assessment.
- Portfolio scoring study. The abovementioned portfolio assessment utilizes a scoring design in which different portfolio entries made by the same student are

each scored independently by different scorers. An alternate approach to scoring this assessment might be to allow the same scorer to evaluate all of the different portfolio entries and to allow some holistic judgment of the complete set of essays to enter into the scoring process. To some extent, this is a validity question: should portfolio entries be treated as independent “items” or should entries be judged as a complete body of work? There are clearly philosophical aspects to this question as well as implications for scoring efficiency and cost, but a more concrete initial measurement question is, how do the two scoring approaches compare?

- Trait vs. holistic rubrics/scoring study. We would like to conduct research comparing trait-based scoring with holistic scoring. This study would evaluate different content areas and consider scoring reliability, validity, and comparability of scores.
- Combined scoring of items and variants. Some constructed response items are based on another item and modified only in minor ways (e.g., values manipulated in a mathematics problem). Sometimes the relationship between a base item and variants is referred to as “parent-child”. In one program, we conduct independent rangefinding, training, and scoring for parent and child constructed response items. However, we are interested in studying whether parent and child constructed response items could share common scoring activities.

Summary

In this paper, we summarized and discussed a number of research studies related to the human scoring of constructed response items. These studies have addressed approaches and procedures related to training scorers and the scoring process itself. Topics included image-based scoring, online scorer training approaches, the use of distributed scoring procedures, the use of annotations in constructed response item scoring, and the training procedures associated with scoring constructed response items administered by computer. Although sometimes inconclusive and limited in scope, the studies add to the literature on constructed response scoring and are helpful in informing clients about issues important to their assessment programs.

References

- American Educational Research Association (AERA), American Psychological Association (APA), and the National Council on Measurement in Education (NCME). (1999). Standards for educational and psychological testing. Washington, DC: AERA.
- Breland, H., Lee, Y.W., & Muraki, E.(2005). Comparability of TOEFL CBT essay prompts: response-mode analyses. Educational and Psychological Measurement, *65* (4), 577-595.
- Dikli, S. (2006). An overview of automated scoring of essays. Journal of Technology, Learning, and Assessment, *5*(1). Retrieved December 13, 2007 from <http://www.jtla.org>.
- Elder, C., Barkhuizen, G., Knoch, U., & von Randow, J. (2007). Evaluating rater responses to an online training program for L2 writing assessment. Language Testing *24*(1) 37–64. Available at: <http://ltj.sagepub.com/cgi/content/abstract/24/1/37>.
- Fan, M., Kanada, M., Alagoz, C., Harms, M., Meyers, J., & Nichols, P. (2007, July). NC comparability study: Paper and pencil vs. online at grades 7 and 10. Report prepared for the North Carolina Department of Public Instruction.
- Hollenbeck, K., Tindal, G., Stieber, S., & Harniss, M. (1999). Handwritten vs. word processed statewide compositions: Do judges rate them differently? Retrieved March 16, 2006, from http://brt.uoregon.edu/files/Hdwrtn_vs_Typed.pdf.
- Knoch, U., Read, J., & von Rensdow, J. (2007). Re-training writing raters online: How does it compare with face-to-face training? Assessing Writing, *12*(1), 26-43. Available at: <http://www.sciencedirect.com/science/journal/10752935>.
- Kreiman, C. (2007). Investigating the effects of training and rater variables on reliability measures: A comparison of standup local scoring, online distributed scoring, and online local scoring. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.
- Linacre, J. M. (1989). Many-facet Rasch measurement. Chicago: MESA Press.

- Nichols, P. (2002). Paper versus image-based scoring. Paper presented at the 32nd National Conference on Large-Scale Assessment, Palm Desert, CA.
- Nichols, P. (2005). Evaluating the use of annotations when scoring essays. Paper presented at the 35th National Conference on Large-Scale Assessment, San Antonio, TX.
- Nichols, P., Kreiman, C., & Kanada, M. (2006). Evaluating computer-delivered scorer training and scorer performance for scoring constructed responses. Paper presented at the annual meeting of the Iowa Educational Research and Evaluation Association, Iowa City, IA.
- Phillips, S.M. (2007). Automated essay scoring : A literature review. (SAEE research series #30). Kelowna, BC: Society for the Advancement of Excellence in Education.
- Powers, D. E., Fowles, M. E., Farnum, M., & Ramsey, P.(1994). Will they think less of my handwritten essay if others word process theirs? Effects of essay scores of intermingling handwritten and word-processed essays. Journal of Educational Measurement, 31 (3), 220-233.
- Sweedler-Brown, C. (1991). Computers and assessment: The effect of typing versus handwriting on the holistic scoring of essays. Research & Teaching in Developmental Education, 8(1), 5–14.
- Vickers, D., & Nichols, P. (2005). The comparability of online vs. standup training. Paper presented at the 35th National Conference on Large-Scale Assessment, San Antonio, TX.
- Way, W. D., & Fitzpatrick, S.(2006). Essay responses in online and paper administrations of the Texas assessment of knowledge and skills. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.
- Yu, L., Livingston, S. A., Larkin, K. C., & Bonett, J. (2004). Investigating differences in examinee performance between computer-based and handwritten essays (ETS Research Report RR-04-18). Princeton, NJ: Educational Testing Service.