Evaluating Computer Automated Scoring: Issues, Methods, and an Empirical Illustration

Yongwei Yang
The Gallup Organization

Chad W. Buckendahl
Buros Center for Testing
University of Nebraska-Lincoln

Piotr J. Juszkiewicz
The Gallup Organization

Dennison S. Bhola
James Madison University

Please direct correspondence regarding this manuscript to:

Chad W. Buckendahl
Buros Center for Testing
21 TEAC, UNL
Lincoln, NE 68588-0353
cbuck1@unl.edu

Abstract

With the continual progress of computer technologies, computer automated scoring (CAS) has become a popular tool for evaluating writing assessments. Research of applications of these methodologies to new types of performance assessments is still emerging. While research has generally shown a high agreement of CAS system generated scores with those produced by human raters, concerns and questions have been raised about appropriate analyses and validity of decisions/interpretations based on those scores. In this paper we expand the emerging discussions on validation strategies on CAS by illustrating several analyses can be accomplished with available data. These analyses compare the degree to which two CAS systems accurately score data from a structured interview using the original scores provided by human raters as the criterion. Results suggest key differences across the two systems as well as differences in the statistical procedures used to evaluate them. The use of several statistical and qualitative analyses is recommended for evaluating contemporary CAS systems.

Keywords: automated scoring, computerized testing, structured interviews, validity

Evaluating Computer Automated Scoring: Issues, Methods, and an Empirical Illustration

Introduction

A major challenge faced by testing programs pertains to the scoring of various types of constructed-response items. In recent years, computer automated scoring (CAS) systems have emerged as one solution to this challenge. While different CAS procedures have yielded a range of results, there is generally a high level of correspondence between the scores produced by human scorers and CAS systems (see, for example, a review by Khaliq [2003]). Analyses of the scores produced by human raters and CAS systems can provide valuable information when addressing reliability and validity issues.

While validity issues have been discussed (e.g., Clauser, Kane, & Swanson, 2002) and general framework for validation design has been offered (e.g., Bennett & Bejar, 1998; Yang, Buckendahl, Juszkiewicz, & Bhola, 2002), a more systematic examination of contemporary validity concepts, current practices and available methodologies is needed. There are a number of questions about CAS systems raised by the research community and general public. Some of these stakeholder groups are concerned with the appearance of conflict of interest because the published research in the area is mostly completed by groups that develop and market CAS systems. Others are concerned about generalizability of the CAS studies across domains and populations. Others, yet, are simply not convinced that methodologies used to examine the meaningfulness of CAS generated scores are most appropriate.

In this paper, we discuss validity issues in CAS, describe methods for evaluating validity evidence to support CAS-produced scores, and offer an illustration of how these methods were applied to a structured interview used for employment testing. Our goal with this study is to extend the validity literature pertaining to CAS to other types of tests. In doing so we examined

the efficacy of two CAS systems with constructed-response data from an instrument measuring personality and attitude aspects, rather than measuring knowledge, skill, and ability. We also evaluated these systems as they applied to multiple open-ended items from structured interviews rather than individual or a small number of writing prompts or tasks that are used in most applications. Finally, we examine the decisions about the meaningfulness of the CAS systems with respect to both item and total scores. Throughout, we explain our rationale for choosing certain methodologies and make general recommendations on designing and conducting studies involving CAS systems.

A secondary goal of this study is to illustrate appropriate analyses that can be conducted on pilot data to empirically compare two or more CAS systems. We focus our illustrations on analyses that can be done during a testing program's piloting phase to evaluate the feasibility of CAS systems. Practitioners are looking for guidelines and examples of how to carry out adequate analysis to determine whether CAS systems can perform appropriate scoring tasks. We also hope that these illustrations will encourage potential users of CAS systems to conduct independent analyses, instead of relying on research provided by system developers.

Contemporary CAS Systems

In the field of education, attempts to develop CAS systems date back to the late 1960s when Ellis Page developed the first generation of Project Essay Grader (PEG) (Kukich, 2000; Page, 1966; Page, 1994). The development of PEG consisted of three major steps (Page, 1994). First, a set of measurable *proxy* features for assessing essay quality was identified. Next, multiple regression was used to find the optimal combination of these features that best predicts the ratings of human experts. Finally, the features and the optimal combination were translated into computer programs. The current version of PEG may be used to provide holistic scores of essays as well as

trait scores for instructional and diagnostic feedback (Shermis, Koch, Page, Keith, & Harrington, 1999).

In recent years, several other CAS systems have been developed to score essay responses, including Intelligent Essay Assesor (IEA) by Knowledge Analysis Technologies, Intellimetric by Vantage Learning, and E-rater by ETS-Technologies. Like PEG, these systems claim to provide both scores and some amount of instructional feedback. The IEA (Landauer, Laham, & Foltz, 2001; Laham, 2001) applies latent semantic analysis (LSA) to assess writing quality. LSA methodology is used to judge the semantic relatedness among documents such as essays (Landauer & Dumais, 1997; Landauer, Foltz, & Laham, 1998). To generate scoring models for essays, the IEA engine first processes a large body of text in the domain of interest. Then, each scoring model is calibrated on a number of essays that human experts have rated. In the end, IEA predicts how the human experts would have scored the semantic content of a new essay by comparing it to essays used in the calibration process (Landauer, Laham, & Foltz, 2000, 2001).

Elliot (2001) and the Intellimetric website provide information about the technologies and models of their respective CAS systems. According to these sources, Intellimetric incorporates various artificial intelligence techniques and statistical methods. It generates a scoring model for an essay prompt by training on a set of pre-scored responses without pre-specifying a set of features or a scoring rubric. Elliot (2001, p. 2) asserts Intellimetric is able to "infer the rubric and the pooled judgments of the human scorers" from the training materials.

The core technologies of E-rater came from research in the areas of natural language processing and information retrieval (Burstein, Kukich, Wolff, Lu, & Chodorow, 1998; Burstein, Kukich, Wolff, Lu, Chodorow, Braden-Harder, & Harris, 1998; Burstein & Marcu, 2000; Burstein, 2001a, 2001b). These technologies were used to develop three modules aimed at

identifying three characteristics of an essay: syntactic variety, topic content, and organization of ideas (Kukich, 2000; Burstein & Marcu, 2000). To generate a scoring model, a scoring rubric is needed. The rubric includes specific descriptions of score levels. E-rater then processes a set of training essays that expert raters have pre-scored. It identifies the salient features and models in the pre-scored essays to find the optimal combination that best predicts expert ratings. The resulting model is used to create the scoring program for new essays.

In addition to those CAS systems developed to score essays, there are also applications developed for scoring other types of constructed-response items. Stephen Clyman and his colleagues at the National Board of Medical Examiners developed a system to score computer-simulated performance assessments of physicians' patient management skills (Clauser, Subhiyah, Nungester, Ripkey, Clyman, & McKinley, 1995; Clauser, Margolis, Clyman, & Ross, 1997; Clauser, Ross, Clyman, Rose, Margolis Nungester, Piemme, Chang, El-Bayoumi, Malakoff, & Pincetl, 1997; Clauser, Swanson, & Clyman, 1999; Clauser, Harik, & Clyman, 2000). National Council of Architectural Registration Boards (NCARB) developed another system to score graphic simulation tasks on the Architect Registration Examination (Bejar, 1991; Bejar & Braun, 1994; Williamson, Bejar, & Hone, 1999). A third system is used in the Dental Interactive Simulation Corporation (DISC) assessment (Johnson, Wohlgemuth, Cameron, Caughtman, Koertge, Barna, & Schultz, 1998; Mislevy, Steinberg, Breyer, Almond, & Johnson, 2002).

CAS systems have faced skepticism since their inception. Some of the criticisms pertain to logistics issues such as the costs of developing CAS systems, and the availability of computers and other technologies required to implement them. Other concerns address more fundamental issues regarding the validity of CAS system generated scores, such as the over-reliance on surface features of responses, the insensitivity to the content of responses and to creativity, and the

vulnerability to new types of cheating and test-taking strategies. However, with the advances in theories and technologies as well as the increasing accessibility to computers, processing speed, and web-based applications, these concerns are being addressed.

In general, two important improvements contribute to the increasing popularity of CAS systems. First, compared to earlier procedures, which relied heavily on surface elements, current CAS systems are designed to analyze deeper structures of responses and to replicate measurement of the constructs of interest. Second, CAS systems continue to receive increasingly sophisticated evaluations regarding their appropriateness and utility. The improvements made under such intense scrutiny further increased the popularity and credibility of the CAS systems. In the context of these improvements, Williamson, Bejar and Hone (1999) presented a number of advantages of modern CAS systems over human scoring. With CAS systems, a given response will always receive the same results (reproducibility); the same scoring criteria is consistently applied to all responses (consistency); specific reasons and processes behind computer scoring can be traced, investigated and manipulated (tractability); items are to be constructed in a more precise fashion (item specification); responses can be evaluated at a higher level of precision and specificity (granularity); scoring criteria are more articulated and much of the subjectivity in human scoring can be removed (objectivity); scoring outcomes are likely to be more reliable (reliability); and the scoring process can be less time-, resource- and cost-demanding (efficiency). These advantages support the expansion of these systems in testing programs. At the same time, recommendations for appropriate validation strategies will continue to dominate discussion.

CAS Validation Frameworks

Bennett and Bejar (1998) presented an original framework for validating computer-based test scores. Using contemporary validity theory as a guide, they argued that the scoring method

should be considered as a dynamic component of a larger testing system. Such a testing system would include interrelated components such as test development tools, examinee interface, tutorials, and reporting methods. The interrelatedness among these components is obvious as the decisions made upon each one of them would directly or indirectly affect the others. Because the emphasis of the study is on CAS systems, we focus on the relationships that include the scoring method.

Clauser et al. (2002) discussed validity issues in details for computer scoring. Yang et al. (2002) added that one should also evaluate the relative importance and appropriateness of certain validity evidence by considering the level of integration of a CAS system to the entire testing program. Differences in the level of integration reflect the differences in the perceptions of the utility and implications stemming from the use of a CAS system. One practical utilization of a CAS system is as merely a replacement or substitute for human scorers. In this situation, validity evidence for the human-generated scores can be used as the basis of validating scores generated with the CAS procedure. When assessing agreement of a CAS system produced scores, multiple statistical indices should be used because different agreement indices provide related, but different information on the performance of a CAS procedure. Additionally, one should evaluate the adequacy and relevance of different indices to a given situation.

Naturally, with the proliferation of CAS systems, there is a growing body of literature on the attempts made to validate the meaning and uses of scores generated by CAS systems. These validation strategies may be classified into three general types. The first type focuses on the relationship among scores given to the same response by different scorers. The second type focuses on the relationship between these scores and external measures. The third type focuses on

the scoring processes and the mental models represented by the CAS systems. For a detailed review of these validation strategies, see Yang, et al. (2002).

In subsequent sections we present an empirical illustration that compares the performance of two CAS systems on pilot data for a structured employment interview that has multiple constructed response items. We sought to illustrate methods that may be used to evaluate the following questions: 1) Could two existing CAS systems be extended to a different constructed-response test? 2) What are appropriate types of reliability evidence at the item and total score levels? 3) Is there sufficient validity evidence to use these scores for the desired inference?

Methods for Evaluating CAS System Generated Scores

The following analyses reflect a validation approach that focuses on the relationship among scores given to the same instrument by different scorers. This approach is consistent with the purpose of the study, that is, to evaluate the efficacy of a CAS system using pilot data. The level of CAS system integration for the testing program was low. According to Yang et al. (2002), at this level of integration, the capability of a CAS system to replicate the scores of expert raters is the major concern. Moreover, even with higher levels of integration, score correspondence is a key starting point for evaluating the performance of computer scoring.

Various methods are used to assess rater agreement in educational, psychological, medical, and biological research. These methods can also be used to evaluate agreement between CAS system generated scores and expert raters' scores. There are four key aspects of rater agreement. First, agreement analyses can be used to evaluate the strength of association among scores (or categories) given by different raters. Second, the introduction of systematic bias from raters can be examined. For example, raters may have different tendencies in terms of leniency, they may have different interpretations of a scoring rubric, or they may use the scoring scale differently. The

comparison of score distributions of raters affords the opportunity to study the presence of biases. Third, the agreement among raters in an absolute sense can be examined, e.g. whether raters are likely to assign the exact same score or category to the same response. In this case, point-by-point agreement analyses are useful. Finally, one may want to examine the pattern and nature of agreement or disagreement among raters. Such information can then be used to improve the scoring rubric and training of human or computer raters. It should be noted that it is a commonly held impression that rater agreement is about the "reliability" of scores. However, evidence on the four aspects described above extends traditional concepts of reliability, which address the consistency of outcomes across replications of the measurement process and address key validity issues. It is also important to note that although we may ascribe a level of confidence to the score when independent judgments yield similar values, rater consistency only reflects one source of potential error in the score and by itself is insufficient to support validity.

In the context of assessing the quality of CAS system generated scores, because we generally rely on expert raters for this purpose, the fundamental question is whether a computer program can provide scores that are accurate with respect to some criterion. With subjectively scored tests, our common criterion is the scores of human experts. Rater agreement analysis is a strategy to evaluate the concurrent validity of the CAS system generated scores. In this sense the first aspect of rater agreement – association between raters – is a rather liberal approach. A high association among scores is not the same as exact correspondence among scores. A computer program may consistently assign one more point to every response and its scores will still have a perfect correlation with the criterion.

The other three aspects may be more critical to evaluate the validity of scoring. Providers and users of CAS systems need to show that the computer system does not assign different scores

(or categories) systematically in terms of mean scores or distributions (e.g., the criterion scores). They also need to show that an individual who should be given a particular score according to the criterion will be likely to receive that score from the CAS system. Finally, understanding the pattern and nature of agreement or disagreement helps to improve the performance of the CAS system, and to support the validity of the inferences made from CAS system generated scores.

Although there are many analytical methods for investigating rater agreement, it is useful to begin the examination of data with simple means and a review of cross-tabulations between item scores produced by the CAS systems and the criteria. It is within this context that the proportion of overall agreement index is commonly used. This measure shows the percent of times when a CAS system assigned the same scores as the rater(s). However, this index has limitations because it can be very high simply by chance (Cohen, 1960; Fleiss, 1975; Fleiss, 1981). To overcome this problem, researchers have proposed agreement indices that take into account the size of chance agreement (Scott, 1955; Cohen, 1960; Landis & Koch, 1977; Maxwell, 1977; Fleiss, 1981; Zwick, 1988). Among these indices, Cohen's κ coefficient (Cohen, 1960) is widely used. This coefficient has its own assumptions and limitations and its application has been cautioned (Cohen, 1960; Maclure & Willett, 1987; Zwick, 1988; Feinstein & Cicchetti, 1990; Cicchetti & Feinstein, 1990; Cook, 1998).

In a discussion of κ and a few other κ-like coefficients, Zwick (1988) suggested the use of a 2-step approach to assess rater agreement. The first task is to perform a test on the similarity of score distributions produced by raters. This analysis can be accomplished by testing the marginal homogeneity of the raters' scores. If marginal homogeneity is rejected, further analysis of the level of agreement between raters is unnecessary. If marginal homogeneity is not rejected, κ, or preferably, according to Zwick (1988), Scott's π coefficient can be used to assess chance-

corrected agreement. Investigating the difference in the distributions of raters' scores provides information regarding rater biases. It is only after biases are ruled out, that the assessment of agreement can become an assessment of the accuracy of scores.

Method

Data Source and Background

To explore the application of CAS systems to a new domain, namely employment testing, organizations that market these systems to analyze and score constructed responses were invited to use their CAS systems to score a structured employment interview. In this study the structured interview measures work motivation, interpersonal skills, and cognitive styles. Together these aspects can be used to predict success in a particular job environment. The structured interview consisted of 60 open-ended items. Some of the questions fell in the range of typical situational or behavior description prompts used in employee selection whereas others were different, such as "How competitive are you?"[1]. Some questions included follow-up probes that asked candidates to give examples, such as "Tell me a time when you were very competitive."

The instrument contained job-related and biographical questions that were related to defined job performance criteria. For each item, expert raters using a scoring rubric scored candidates' responses. The scoring rubric credited a given response as showing either the presence or absence of a desirable aspect. Although the decision was dichotomous (scored as 0 or 1) at the item level, the decision about the employability of a candidate was made at the total score level (ranges between 0 and 60).

The study used a pool of 326 interview transcripts. Expert raters pre-scored the transcripts. Among the sample of transcripts, 286 were randomly selected to form a training set to calibrate

[1] Due to the proprietary nature of the actual interview questions, examples in the text are not directly from the actual interview used, but are similar in structure.

the CAS systems. The remaining 40 were not used for calibration and formed a validation sample to evaluate the performance of scoring models. The analyses were based on the data from these 40 transcripts. Although larger validation sample or additional cross-validation samples of transcripts are desirable to evaluate the performance of the CAS systems, it is also important to have a large sample to initially create the scoring models. In our study, with the limited number of transcripts available, focus was first given to ensure a large enough amount of transcripts were used to train and calibrate the scoring models. The decision was made based on the inputs from the developers of the CAS systems with respect to the optimum size of the calibration sample. Results from two CAS systems[1] (i.e., CAS systems A and B) are reported.

Data Analyses

In the current study, agreement analysis at the item level followed Zwick's (1988) two-step approach. The first step evaluates the differences among item difficulties (proportion correct) computed from each CAS system and from expert raters. Because items were dichotomously scored, item difficulties obtained by two raters can be used to assess the homogeneity of the marginal distributions with the Stuart-Maxwell test (Stuart, 1955; Maxwell, 1970; Zwick, 1988). The null hypothesis of the test is: *Ho:* $p_{i+} = p_{+j}$, where p_{i+} is the marginal proportion of being in row i, and p_{+j} is the marginal proportion of being in column j. As Zwick (1988) noted, in a 2-by-2 table, the Stuart-Maxwell test becomes the McNemar test. The computation of the McNemar test statistic is shown in formula 1.

$$\chi^2 = \frac{(B-C)^2}{B+C} \tag{1}$$

[1] The names of the CAS systems are not identified to protect the confidentiality of the respective organizations in this exploratory study.

In formula 1, *B* represents the number of item responses to which one rater assigned a "0" but the other assigned a "1", and *C* is the number of item responses to which one rater assigned a "1" and the other assigned a "0". The sampling distribution of the above statistic when H_o is true is asymptotically distributed as χ^2 with $df = 1$ (Siegel & Castellan, 1988). The "desirable" result of the McNemar test in this situation is to retain the null hypothesis. That is, the marginal homogeneity holds for an item. In this situation, Type II error (i.e., failure to reject a false null hypothesis) was judged to be a more serious concern than Type I error. A common method for controlling for Type II error is by raising the level of Type I error (alpha) allowed. In our case, a higher alpha level (.10) was chosen. A higher Type I error rate makes it easier to reject the null hypothesis and in turn provides greater confidence in a decision where the null hypothesis is not rejected. We acknowledge that a potential negative consequence of setting a more lenient alpha level is the greater probability of rejecting the null hypothesis suggesting that the distributions are different when perhaps they are not. The consequence of a possible "over-rejection" could be a higher level of scrutiny before accepting the performance of CAS system. However we believe this is desirable in a pilot study where a CAS system's utility to a new situation is investigated.

The second step assesses the item-level decision consistency between CAS systems and the expert raters. Such analyses were conducted only on items that the test of marginal homogeneity was not statistically significant. The Scott's π coefficient (Scott, 1955; Zwick, 1988) was used as a major index for this analysis and is calculated as:

$$\pi = \frac{P_a - P_c}{1 - P_c} \qquad \textbf{(2)}$$

where P_a is the proportion of observed exact agreement. It is computed as the number of incidences where both raters assigned the same score divided by the total number of incidences.

The term P_c is the expected proportion of agreement by chance (Scott, 1955; Zwick, 1988). With dichotomously scored items, it is computed as

$$P_c = \left(\frac{p_{0+} + p_{+0}}{2}\right)^2 + \left(\frac{p_{1+} + p_{+1}}{2}\right)^2, \quad \textbf{(3)}$$

where P_{0+} and P_{1+} are the proportions of one rater assigned a score of 0 and a score of 1, respectively, and P_{+0} and P_{+1} are the proportions of the other rater assigned a score of 0 and a score of 1, respectively. Along with Scott's π coefficient, the overall proportion of agreement (P_a) was also reported.

Finally, the level of agreement between a CAS system and expert rater on each item was classified based on combining the analyses of item difficulty, marginal homogeneity and item level decision consistency. The following decision rules were used for this classification:

1. The agreement between a CAS system and expert rater on an item was classified as poor if the test of marginal homogeneity was statistically significant at .10 level, or if the difference between item difficulties (proportion correct) was greater than .10 as an indicator of practical significance.
2. If the test of marginal homogeneity was not significant *and* item difficulty difference was less than or equal to .10, the level of agreement was still classified as poor if the Scott's π coefficient was less than .40 or P_a was less than .75.
3. If the test of marginal homogeneity was not significant and the difference in item difficulty was less than or equal to .10, the level of agreement was classified as good if the Scott's π coefficient was greater than or equal to .75 and P_a was greater than or equal to .90.
4. All other situations were classified as moderate agreement.

Fleiss (1981) suggested that Cohen's κ values of .75 or larger indicate excellent agreement and values less than .40 indicate poor agreement. Zwick (1988) suggested that when marginal distributions are similar, values of Scott's π and Cohen's κ become very close, with the former always being smaller than or equal to the latter. Thus, Fleiss's (1981) recommendation on Cohen's κ was applied to choose cutoff values of Scott's π. Incidentally, the minimum and maximum values of κ -like statistics (e.g., κ and π) depend on, among other things, marginal distributions (Cohen, 1960; Fleiss, 1981). This should caution researchers against applying universal cutoff values of these statistics. In this study, it was less problematic to apply the same set of cutoff values across different items because Scott's π was used to make classification decisions only when marginal homogeneity was assumed.

Because the employee selection interview data are used to facilitate hiring decisions based on total scores, score agreement at the total score level is also important when evaluating CAS system generated results. Besides item-level analyses, characteristics of the total scores from the CAS systems and expert raters were examined. The analyses included assessing distribution comparability, score correlations, score differences and decision consistency based on total scores. The rationale for these analyses is that when inferences are made about individuals on the basis of any scoring method there should be consistency in those scores to provide empirical evidence for those inferences.

The first series of total score level analyses examined the similarity between the pairwise comparisons of total score distributions. Such analyses were conducted first because if the score distributions were drastically different, then high levels of score agreement are unlikely. Because at this first stage of analysis the agreement between individual scores is not yet of interest, two-sample tests were utilized. Two non-parametric tests, the Kolmogorov-Smirnov (K-S) and

Wilcoxon-Mann-Whitney (W-M-W) tests were used. Each of these tests assumes that the distributions are based on at least ordinal data. Generally, the K-S test is better for smaller samples and the W-M-W test is better for larger samples (Siegel & Castellan, 1988). Our sample of 40 was on the threshold of the distinction between large and small, so both statistical tests were used and statistical significance tests were conducted using an alpha level of .10. Again, the .10 Type I error level was chosen because the desirable outcome of such test is that the score distributions are not significantly different. Next, descriptive analyses were conducted to calculate the mean, range and standard deviation of the total scores to inspect the shape of the distribution.

The next three series of analysis focus on the level of agreement between total scores. First, correlations were computed among the total scores of various scoring methods. A Pearson correlation was chosen because the individual item-level data, when combined, might be interpreted at an interval level of measurement. Evaluating the relationship among scores produced by the various scoring methods is important to ensure that different methods elicited similar structure from the responses. However, it is also important to evaluate agreement between scores individuals received. In this study, the Wilcoxon signed rank test (Siegel & Castellan, 1988) is used to assess difference between scores.

Finally, decision consistency analyses were conducted because the interview was designed to make total-score-based decisions about candidates' advancement in a selection process. The same two-step approach used in the item-level analysis was used to assess total-score-level decision consistency. The McNemar test was used to first test for marginal homogeneity because the decision took on a binary form. When the test is not significant at a .10 level, Scott's π based on pass/fail classification decisions were calculated for each of the pairwise comparisons among the three scoring methods considered. These analyses were conducted on decisions that were

based on a cut score that ranged from approximately half a standard deviation below to approximately half a standard deviation above the mean scores of expert human raters. These cut scores were selected to cover the typical range of cut points used in the applications of similar structured interviews, such as cut points might have been chosen by clients using these instruments to assist hiring decisions. Proportion agreement and differences in the passing rates were also calculated for each pair of comparison.

Results

Item level analyses

Table 1 provides a summary of descriptive analyses of item difficulty differences and the test of marginal homogeneity. When looking at differences in item difficulties obtained from human rater and a CAS system, CAS system A had 20 items whose difficulties differed by more than .10 from those obtained from human rater. CAS system B had 6 such items. Nineteen of the 60 items failed the test of marginal homogeneity ($p < .10$) with system A, and 5 items failed the test with system B.

[Insert Table 1 Here]

Table 2 summarizes the analyses using Scott's π and overall proportion of agreement. Following the rationale of the two-step approach, the summary only includes items for which the test of marginal homogeneity was not statistically significant at .10 level. As shown by this table, among the items passed the test of marginal homogeneity (41 from CAS system A and 55 from CAS system B), the distributions of Scott's π statistics and overall proportion of agreement are quite similar between the two systems.

[Insert Table 2 Here]

Table 3 summarizes, for both systems, the levels of agreement between a CAS system and human rater based on multiple criteria. Performance on each item was classified as "poor", "moderate", and "good" following the four decision rules described earlier in the data analysis section. These decision rules took into account the statistical tests of marginal homogeneity, the numerical differences in item difficulty, and two indices of item-level decision consistency (Scott's π and percent exact agreement). With these decision rules, CAS system A had "good", "moderate", and "poor" agreement with expert raters on 2, 12, and 46 items, respectively. CAS system B had "good", "moderate", and "poor" agreement with expert raters on 4, 15, and 41 items, respectively. The numbers of good, moderate, and poor agreement items are similar between the two systems. However the differences in item difficulties between computer and human scoring were somewhat larger with CAS system A. For example, CAS system A had 19 items that failed the test of marginal homogeneity whereas CAS system B had only 5.

[Insert Table 3 Here]

Table 4 further illustrated that although the overall numbers of good, moderate, and poor agreement items were similar, the two CAS systems did not perform equally on the same items. There were only 9 items that both systems' performance was classified as good or moderate. On the contrary, there were 15 items that one system's performance was good or moderate whereas the other system's performance was poor.

[Insert Table 4 Here]

Total Score Analyses

Table 5 presents the results of the total score distribution comparability analyses. As shown in Table 5 the comparisons of the score distributions were similar for both statistical procedures. None of the comparisons suggests that there were distribution differences across the

three scoring methods (i.e., 2 CAS systems, and expert raters). These results provided guidance and greater confidence for the additional analyses that are described below.

[Insert Table 5 Here]

Table 6 shows the comparison of total score means and standard deviations for each of the three scoring models. The mean total scores for the three methods were very similar with values ranging from 27.08 to 27.88. The standard deviation, though, was somewhat different across the scoring methods, ranging from 3.94 to 6.81. This limited range in variation of scores speaks to the validity of the inferences. It is assumed that the instrument measures a range of performance and that candidates who took the interview had a similar range of performance. Yet, with some scoring methods, there was a limited range in scores. Such a narrow range means that if items are designed to distinguish between higher and lower candidate performance, some of the items are not adding to this distinction using the two CAS systems.

[Insert Table 6 Here]

In practice, at this point one may conclude that the CAS systems produced scores that differ in meaning from those of expert raters (and hence also lack item-level agreement). However to complete the illustration on possible analytical methods, an analysis to assess total-score-level score agreement was performed. These analyses evaluated correlations among total scores, agreement between total scores, and decision consistency.

Table 7 shows the correlations among the total scores of various scoring methods, as well as the results of the Wilcoxon signed rank test, where scores of each candidate's responses (transcript) received were compared across scoring models. The correlations among the total scores of these scoring models and the statistically non-significant results of the test of pairs of scores suggested that there were moderate levels of consistency at the total score level. A stronger

relationship was found between the human rater and CAS system B. The results suggest that the two CAS systems shared some of the variance with the expert raters but there were also differences in the psychometric properties on the scores.

[Insert Table 7 Here]

The use of test scores to make pass/fail decisions requires the inspection of decision consistency. Tables 8 to 10 show the results of decision consistency analysis at three different points where cut scores for "passing" could be set. Table 8 shows the number of candidates for each CAS system that passed or failed compared to the classification decisions of an expert rater. The cut score for the distribution of scores was set at approximately half a standard deviation below the mean (i.e., 24 out of a maximum of 60 possible points).

[Insert Table 8 Here]

As shown in Table 8, with both CAS systems A and B, the test of marginal homogeneity was not statistically significant. Moderate agreement existed between the expert rater and the CAS system A scoring model, with a Scott's π of 0.57 and with 6 candidates misclassified. Low agreement existed between the expert rater and the CAS System B scoring model, with a Scott's π of 0.06 and with 12 candidates classified differently. The calculation of the proportion of agreement between the expert rater and the CAS scoring models produced values of 0.85 and 0.70 for CAS systems A and B, respectively. A similar analysis was conducted to compare the decisions between the expert rater and both CAS systems using a cut score of 28 (mean value of human scoring model). Table 9 contains these results. Again, the test of marginal homogeneity for both CAS systems was not statistically significant.

[Insert Table 9 Here]

As shown in Table 9, there was relatively low agreement between the expert rater and both CAS Systems A and B using a cut score that was set at the mean. Calculating the proportion of agreement between the expert rater and CAS scoring models produced values of 0.63 and 0.60 for models A and B, respectively. For both CAS systems, we found similar levels of agreement with 15 or 16 candidates classified differently depending on the system. If the CAS system-generated total scores were used to make classification decisions, this low level of agreement would pose concern.

[Insert Table 10 Here]

Table 10 shows the comparison of decisions with a cut score set at a half a standard deviation above the mean value of the expert-rated scoring model. According to the π statistic, there was moderate agreement between the expert rater and the CAS systems under this condition. However, the test of marginal homogeneity was significant for CAS system B, which indicates a low agreement between the classifications based on expert rater scores and CAS system B generated scores. Calculating the proportion of agreement between the expert rater and CAS systems produced values of 0.78 and 0.85 for CAS systems A and B, respectively. For both CAS systems, we found slightly different levels of agreement with 6 or 9 candidates classified differently depending on the system. (One interesting observation was that CAS system B seemed to perform better than CAS system A on many comparisons except for the decisions consistency analysis. CAS system B may have performed poorly in the decision consistency analysis because it generated a score distribution that was more different from the human rater scores in terms of mean and standard deviation. This is shown in Table 6.

Overall, the level of agreement on decision consistency was low to moderate at a cut score of 24, low with a cut score of 28, and moderate with a cut score of 31. Although it was not established, a reasonable target for proportion of agreement is 0.90 and a minimum value for π for indicating good agreement may be 0.75. Using these targets, none of the reported comparisons met these standards. These results suggest that there could be less confidence in the classification decisions that would be made about prospective candidates whose responses were scored using the current scoring models of these CAS systems. For the sample used in this study, only 40 participants' scores were included. Many of the comparisons had more than 10 candidates misclassified. The misclassification of candidates raises some concern. If both systems lead to a higher than expected rate of misclassifications, we may not have evidence to select one scoring system over the other.

Discussion

The first objective of our paper was to evaluate the efficacy of two existing CAS systems for scoring multiple open-ended items on a structured employment interview. The findings indicate neither system provided optimal results. The findings are not surprising as the CAS systems were both originally designed primarily to score essays, whereas, the structured interview questions were designed to elicit shorter and more spontaneous verbal responses. These analyses are exploratory in nature and serve only as an initial attempt to extend the existing CAS methodology to a new testing domain. In this study the most intriguing finding is that the different scoring models may be identifying somewhat different characteristics in the underlying constructs or other, perhaps extraneous information, as revealed by the differences in the distributions and relationships of the scores. The dissimilarity in the scoring models reflects a difference in the basic approaches and processes these computer systems used to score responses. Perhaps more

importantly, it emphasizes that the fundamental validity issues, namely construct relevance and representation of the construct, need to be addressed with CAS system generated scores. Because these systems produced different scores, there are questions about how the construct is represented by these scores.

The second goal of the paper was to illustrate how appropriate analyses can be conducted to maximize information gained from limited data that may be collected during the pilot phase of test development. Constraints created by size of the data set and the availability of resources are common in pilot testing, when an organization evaluates the feasibility of using a CAS system.

Although score agreement is usually among the first concern when evaluating CAS system generated scores. The bigger picture, namely the validity of the inferences and decisions based on the assessment outcomes, should always dictate the framework of analysis. Because the CAS systems generated results that could be evaluated differently for the item and total test level, further examination of the characteristics of the scoring algorithm may be needed. For example, how would these results differ if the emphasis of the scoring program was to maximize total score consistency?

A number of methods are available to evaluate score agreement as well as broader validity questions, even when data is limited. Choosing among these methods should be an explicit process that realizes the strengths and weaknesses of each method. We make the following recommendations for selecting among these methods.

1) Select appropriate methods based on the purpose of the analysis. For example, in the initial stage of developing the CAS system, analyzing the patterns and reasons of agreement and disagreement should be a priority.

2) Triangulate the results from multiple, appropriate methods. No single method or index will be able to evaluate all relevant aspects at the same time. Additionally, the inconsistency among the results from different methods is sometimes unavoidable, but may provide unique information.
3) Take into account the actual size of the indices, not just statistical significance to better evaluate the quality of the score or decision.
4) Cross-validate scoring engine calibration activities.
5) Assess the internal structure of test scores, even with descriptive methods where data permit. These analyses provide evidence about the validity (or lack of validity) of the outcomes. They may also point out areas in which scoring methods can be improved.
6) Investigate patterns of agreement/disagreement between scoring models. In the piloting phase of test development, these analyses are helpful to the improvement and defense of these models. Some statistical methods could be applied here, but judgmental review and qualitative analysis are also valuable.

Our study utilized data gathered from a pilot testing of two CAS systems. The availability of data posed a few limitations to the design and analyses illustrated. First, we must caution readers that the pilot testing directly applied the existing computer scoring models to a novel context. Thus the results are in no way an evaluation of these systems performance on the tasks they were designed for, namely well-defined essay questions. Second, the limited sample size forced us to keep a difficult balance between the number of scripts used for calibration and for validation. The small number of scripts set aside for validation prohibited a cross-validation design. Finally, the use of a single human rater rather than a composite or consensus of several

raters as scoring criteria was a result of the restricted resources allocated to the pilot testing, but it limited the generalizability of the findings regarding the performance of the CAS system.

Although the potential for using CAS system to score responses to structured employment interviews was not fully evaluated in the current study, we want to remind readers that one of the purposes of the study was to illustrate how appropriate analyses may be conducted when data are limited, such as in a pilot testing scenario. The data in our study represents an example of the reality that practitioners may encounter. For many potential CAS system users, their evaluation of the systems could be limited by the amount of training and the availability of validation materials, as well as by time, cost and other related resources. The illustrations in this study are intended to provide guidance for evaluating CAS systems with limited data.

We also want to emphasize the importance of considering the evaluation of computer-generated scores as an integral part of a validation plan. As the measurement community continues to create innovative testing approaches that rely on technological solutions for scoring, CAS systems will become more widely used. The subtle distinction between what constitutes reliability evidence and what constitutes validity evidence continues to challenge researchers, particularly as it relates to subjectively scored instruments. It is possible that at the total score level, we are making correct decisions, but what happens at the item level may not reflect the intended measurement. This means that problems with scoring particular items may under- or over-represent examinees' performance; however, if these problems wash out at the total score, the decision of the performance relative to the cut score may not change. As the research in this area expands, the need for choosing appropriate validation methods and using appropriate statistics will endure greater scrutiny. This study contributes to the conversation regarding the choice and

interpretation of appropriate designs and statistics when data collected via computer automated scoring systems are analyzed.

Acknowledgements

The authors would like to acknowledge Stephen G. Sireci and William G. Harris for their helpful comments and suggestions on an earlier version of this article.

References

Bennett, R. E., & Bejar, I. I. (1998). Validity and automated scoring: It's not only the scoring. Educational Measurement: Issues and Practice, winter 1998, 9-17.

Bejar, I. I. (1991). A methodology for scoring open-ended architectural design problems. Journal of Applied Psychology, 76, 522-532.

Bejar, I. I., & Braun, H. I. (1994). On the synergy between assessment and instruction: early lessons from computer-based simulations. Machine-Mediated Learning, 4, 5-25.

Burstein, J. C. (2001a, February). Automated essay evaluation in Criterion. Paper presented at the Association of Test Publishers Computer-Based Testing: Emerging Technologies and Opportunities for Diverse Applications conference, Tucson, AZ.

Burstein, J. C. (2001b, April). Automated essay evaluation with natural language processing. Paper presented at the annual meeting of the National Council of Measurement in Education, Seattle, WA.

Burstein, J. C., Kukich, K., Wolff, S., Lu, C., & Chodorow, M. (1998, April). Computer analysis of essays. In Automated Scoring. Symposium conducted at the annual meeting of the National Council on Measurement in Education, San Diego, CA. Available on-line: http://ftp.ets.org/pub/res/erater_ncmefinal.pdf.

Burstein, J. C., Kukich, K., Wolff, S., Lu, C., Chodorow, M., Braden-Harder, L., & Harris, M. D. (1998, August). Automated scoring using a hybrid feature identification technique. In the Proceedings of the annual meeting of the Association of Computational Linguistics. Montreal, Canada. Available on-line: http://ftp.ets.org/pub/res/erater_acl98.pdf.

Burstein, J. C., & Marcu, D. (2000, August). Benefits of modularity in an automated essay scoring system. In the Proceedings of the Workshop on Using Toolsets and Architectures to Build NLP Systems, 18th International Conference on Computational Linguistics. Luxembourg. Available on-line: http://ftp.ets.org/pub/res/erater_colinga4.pdf.

Cicchetti, D. V., & Feinstein, A. R. (1990). High agreement but low kappa: II. Resolving the paradoxes. Journal of Clinical Epidemiology, 43, 551-558.

Clauser, B. E., Harik, P., & Clyman, S. G. (2000). The generalizability of scores for a performance assessment scored with a computer-automated scoring system. Journal of Educational Measurement, 37, 245-261.

Clauser, B. E., Kane, M. T., & Swanson, D. B. (2002). Validity issues for performance-based tests scored with computer-automated scoring systems. Applied Measurement in Education, 15, 413-432.

Clauser, B. E., Margolis, M. J., Clyman, S. G., & Ross, L. P. (1997). Development of automated scoring algorithms for complex performance assessments. Journal of Educational Measurement, 34, 141-161.

Clauser, B. E., Ross, L. P., Clyman, S. G., Rose, K. M., Margolis, M. J., Nungester, R. J., Piemme, T. E., Chang, L., El-Bayoumi, G., Malakoff, G. L., & Pincetl, P. S. (1997). Development of a scoring algorithm to replace expert rating for scoring a complex performance-based assessment. Applied Measurement in Education, 10, 345-358.

Clauser, B. E., Subhiyah, R. G., Nungester, R. J., Ripkey, D. R., Clyman, S. G., & McKinley, D. (1995). Scoring a performance-based assessment by modeling the judgments of experts. Journal of Educational Measurement, 32, 397-415.

Clauser, B. E., Swanson, D. B., & Clyman, S. G. (1999). A comparison of generalizability of scores produced by expert raters and automated scoring systems. Applied Measurement in Education, 12, 281-299.

Cohen J. (1960). A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20, 37-46.

Cook, R. J. (1998). Kappa. In T. P. Armitage and T. Colton (Eds.), The Encyclopedia of Biostatistics (pp. 2160-2168). New York: Wiley.

Elliot, S. M. (2001, April). IntelliMetric: from here to validity. Paper presented at the annual meeting of the American Educational Research Association, Seattle, WA.

Feinstein, A. R., & Cicchetti, D. V. (1990). High agreement but low kappa: I. The problems of two paradoxes. Journal of Clinical Epidemiology, 43, 543-549.

Fleiss, J. L. (1975). Measuring agreement between two judges on the presence or absence of a trait. Biometrics, 31, 651-659.

Fleiss, J. L. (1981). Statistical Methods for Rates and Proportions (2nd ed.). New York: John Wiley.

Johnson, L. A., Wohlgemuth, B., Cameron, C. A., Caughtman, F., Koertge, T., Barna, J., & Schultz, J. (1998). Dental Interactive Simulations Corporation (DISC): Simulations for education, continuing education, and assessment. Journal of Dental Education, 62, 919-928.

Khaliq, S. N. (2003). A Review and Critique of Automated Scoring for Large-Scale Performance Assessments. Center for Educational Assessment Research Report No. 479, Amherst, MA: School of Education, University of Massachusetts Amherst.

Kukich, K. (2000). Beyond automated essay scoring. IEEE Intelligent Systems [On-line], 15(5), 22-27. Available: http://www.knowledge-technologies.com/papers/IEEEdebate.pdf.

Laham, D. (2001, April). Automated scoring and annotation of essays with the Intelligent Essay Assesor. Paper presented at the annual meeting of National Council of Measurement in Education, Seattle, WA.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychological Review, 2, 211-240.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. Discourse Processes, 25, 259-284.

Landauer, T. K., Laham, D., & Foltz, P. W. (2000). The Intelligent Essay Assesor. IEEE Intelligent Systems [On-line], 15(5), 27-31. Available: http://www.knowledge-technologies.com/papers/IEEEdebate.pdf.

Landauer, T. K., Laham, D., & Foltz, P. W. (2001, February). The Intelligent Essay Assesor: putting knowledge to the test. Paper presented at the Association of Test Publishers Computer-Based Testing: Emerging Technologies and Opportunities for Diverse Applications conference, Tucson, AZ.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. Biometrics, 33, 159-174.

Maclure, M., & Willett, W. C. (1987). Misinterpretation and misuse of the kappa statistic. American Journal of Epidemiology, 126, 161-169.

Maxwell, A. E. (1970). Comparing the classification of subjects by two independent judges. British Journal of Psychiatry, 116, 651-655.

Maxwell, A. E. (1977). Coefficients of agreement between observers and their interpretation. British Journal of Psychiatry, 130, 79 –83.

Mislevy, R. J., Steinberg, L. S., Breyer, F. J., Almond, R. G., & Johnson, L. (2002). Making sense of data from complex assessments. Applied Measurement in Education, 15, 363-389.

Page, E. B. (1966). The imminence of grading essays by computer. Phi Delta Kappan, 48, 238-243.

Page, E. B. (1994). Computer grading of student prose, using modern concepts and software. Journal of Experimental Education, 62, 127-142.

Scott, W. A. (1955). Reliability of content analysis: The case of nominal coding. Public Opinion Quarterly, 22, 321-325.

Shermis, M. D., Koch, C. M., Page, E. B., Keith, T. Z., Harrington, S. (1999, April). Trait ratings for automated essay grading. Paper presented at the annual meeting of National Council on Measurement in Education, Montreal, Canada.

Siegel, S., & Castellan, N. J., Jr. (1988). Nonparametric Statistics for the Behavioral Sciences (2nd ed.). New York: McGraw-Hill.

Stuart, A. (1955). A test of homogeneity of marginal distributions in a two-way classification. Biometrika, 42, 412-416.

Williamson, D. M., Bejar, I. I., & Hone, A. S. (1999). 'Mental model' comparison of automated and human scoring. Journal of Educational Measurement, 36, 158-184.

Yang, Y. W., Buckendahl, C.W., Juszkiewicz, P. J., & Bhola, D. S. (2002). A Review of Strategies for Validating Computer Automated Scoring. Applied Measurement in Education, 15, 391-412.

Zwick, R. (1988). Another look at interrater agreement. Psychological Bulletin, 103, 374-378.

Table 1

Summary results of item difficulty analyses and testing the homogeneity of marginal distributions[1,2]

	CAS System	
	CAS System A	CAS System B
Number of items showing large item difficulty difference[3]	20	6
Largest item difficulty difference	.33	.15
Mean item difficulty difference	.10	.05
Number of items for which marginal homogeneity was rejected[4]	19	5

Note:

1. Total number of items = 60.
2. Number of transcripts = 40.
3. Large difference refers to differences between human and CAS item difficulties > .10.
4. McNemar test described above was used to test the marginal homogeneity. The significance level chosen was .10.

Table 2

Summary results of item level decision consistency analysis[1,2]

	CAS System A		CAS System B	
	π	P_a	π	P_a
Number of Items[3]	41		55	
Minimum	-.14	.48	-.13	.48
Maximum	.84	.98	.84	.95
Mean	.33	.76	.33	.73
Median	.35	.75	.29	.75
Standard Deviation	.25	.11	.23	.12

Note:

1. Number of transcripts = 40.
2. Only items for which marginal homogeneity holds are used in calculating values for this table.
3. Number of items for which marginal homogeneity holds (i.e., test of marginal homogeneity was not statistically significant [$\alpha = .10$]).

Table 3
Item-level analysis decision rules and scoring quality classification results by CAS system

Decision Rule					Number of Items in Each Classification	
Test of Marginal Homogeneity	**Absolute Difference in Item Difficulty**	**Scott's Pi**	**Percent Exact Agreement**	**Classification**	**CAS System A**	**CAS System B**
Statistically Significant at .10 Level	> .100	< .40	< .75	Poor	9	0
			>= .75 and < .90		7	1
			>= .90		0	0
		>= .40 and < .75	< .75		0	0
			>= .75 and < .90		2	1
			>= .90		0	0
		>= .75	< .75		0	0
			>= .75 and < .90		0	0
			>= .90		0	0
	<=.100	< .40	< .75		0	0
			>= .75 and < .90		0	0
			>= .90		1	1
		>= .40 and < .75	< .75		0	0
			>= .75 and < .90		0	0
			>= .90		0	1
		>= .75	< .75		0	0
			>= .75 and < .90		0	0
			>= .90		0	1
Not Statistically Significant at .10 Level	> .100	< .40	< .75	Poor	2	4
			>= .75 and < .90		0	0
			>= .90		0	0
		>= .40 and < .75	< .75		0	0
			>= .75 and < .90		0	0
			>= .90		0	0
		>= .75	< .75		0	0
			>= .75 and < .90		0	0
			>= .90		0	0
	<=.100	< .40	< .75	Poor	13	18
			>= .75 and < .90		8	13
			>= .90		2	0
		>= .40 and < .75	< .75		2	1
			>= .75 and < .90	Moderate	12	14
			>= .90		0	1
		>= .75	< .75	Poor	0	0
			>= .75 and < .90	Moderate	0	0
			>= .90	Good	2	4

Table 4
Comparing performance of the two systems.

CAS System A Item Performance Classification	CAS System B Item Performance Classification			Total
	Good	Moderate	Poor	
Good	1	1	0	2
Moderate	3	4	5	12
Poor	0	10	36	46
Total	4	15	41	60

Table 5
Results of distribution comparisons using Kolmogorov-Smirnov and Wilcoxon-Mann-Whitney tests (n=40).

	K-S value (p)	W-M-W (p)
Human – CAS System A	.67 (.76)	-.49 (.63)
Human – CAS System B	.78 (.57)	-.48 (.63)
CAS System A – CAS System B	.45 (.99)	-.04 (.97)

Table 6
Total score mean and standard deviation for each scoring method (n=40).

	Mean (Range)	Standard Deviation
Human Rater	27.88 (12 – 40)	6.81
CAS System A	27.28 (20 – 36)	4.73
CAS System B	27.08 (18 – 35)	3.94

Table 7

Pearson correlations among scores from each scoring model and Wilcoxon signed rank test of score differences (n=40).

	Human & CAS System A	Human & CAS System B	CAS System A & CAS System B
Pearson r	.62	.70	.65
Wilcoxon Test			
Number of Negative Ranks	23	24	19
Number of Positive Ranks	13	14	17
Number of Ties	4	2	4
Mean Negative Rank	16.85	18.17	19.66
Mean Positive Rank	21.42	21.79	17.21
Z	.86	.95	.64
Probability	.39	.34	.52

Table 8
Comparison of decisions between human and CAS models at a cut score of 24 (n=40).

Human	CAS System A		CAS System B	
	Fail	Pass	Fail	Pass
Fail	6	3	2	7
Pass	3	28	5	26
Proportion Agreement	.85		.70	
Absolute Difference in Passing Rate	.00		.05	
Test of Marginal Homogeneity	.00 (p = 1.00)		.33 (p = .56)	
Scott's π	.57		.06	

Table 9
Comparison of decisions between human and CAS models at a cut score of 28 (n=40).

Human	CAS System A		CAS System B	
	Fail	Pass	Fail	Pass
Fail	14	7	13	8
Pass	8	11	8	11
Proportion Agreement	.63		.60	
Absolute Difference in Passing Rate	.03		.00	
Test of Marginal Homogeneity	.07 (p = .80)		.00 (p = 1.00)	
Scott's π	.25		.20	

Table 10
Comparison of decisions between human and CAS models at a cut score of 31 (n=40).

Human	CAS System A		CAS System B	
	Fail	Pass	Fail	Pass
Fail	24	3	27	0
Pass	6	7	6	7
Proportion Agreement	.78		.85	
Absolute Difference in Passing Rate	.08		.15	
Test of Marginal Homogeneity	1.00 (p = .32)		6.00 (p = .01)	
Scott's π	.45			