DOCUMENT RESUME

ED 393 937 TM 024 972

AUTHOR	Burstein, Jill C.; Kaplan, Randy M.
TITLE	GE FRST Evaluation Report: How Well Does a Statistically-Based Natural Language Processing System Score Natural Language Constructed-Responses?
INSTITUTION	Educational Testing Service, Princeton, N.J.
REPORT NO	ETS-RR-95-29
PUB DATE	Sep 95
NOTE	30p.
PUB TYPE	Reports - Evaluative/Feasibility (142)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	*Computer Assisted Testing; *Constructed Response; Cost Effectiveness; Hypothesis Testing; *Natural Language Processing; Performance Based Assessment; Reading Comprehension; *Scoring; Standardized Tests; Test Construction; *Test Items
IDENTIFIERS	*General Electric Free Response Scoring System

ABSTRACT

There is a considerable interest at Educational Testing Service (ETS) to include performance-based, natural language constructed-response items on standardized tests. Such items can be developed, but the projected time and costs required to have these items scored by human graders would be prohibitive. In order for ETS to include these types of items on standardized tests, automated scoring systems need to be developed and evaluated. Automated scoring systems could decrease the time and costs required for human graders to score these items. This report details the evaluation of a statistically-based scoring system, the General Electric Free-Response Scoring Tool (GE FRST). GE FRST was designed to score short-answer, constructed-responses of up to 17 words. The report describes how the system performs for responses on three different item types: (1) the formulating-hypotheses item; (2) a paraphrase language proficiency item; and (3) a reading comprehension item. For the sake of efficiency, it is important to evaluate systems on a number of item types to see if the system's scoring method can generalize to a number of item types. An appendix shows learning information abut responses recognized by GE FRST. (Contains 7 figures, 13 tables, and 3 references.) (Author/SLD)

RR-95-29

RESEARCH REPORT

GE FRST EVALUATION REPORT: HOW WELL DOES A STATISTICALLY-BASED NATURAL LANGUAGE PROCESSING SYSTEM SCORE NATURAL LANGUAGE CONSTRUCTED-RESPONSES?

Jill C. Burstein
Randy M. Kaplan

Educational Testing Service
Princeton, New Jersey
September 1995

GE FRST Evaluation Report: How Well Does a Statistically-Based Natural Language Processing System Score Natural Language Constructed-Responses?

Jill C. Burstein and Randy M. Kaplan

August 22, 1995

Acknowledgements

The authors would like to thank the following people whose contributions made this evaluation possible. We are grateful to Kelli Boyles, Lois Frankel, and John Hawthorne of SHEP test development for their work on the Formulating-Hypotheses item. They created rubrics for these items, and hand-scored all responses for items used in this study. We would also like to thank Kelli Boyles for writing a summary of the test developers' experience using GE FRST for rubric creation and hand-scoring of responses, and for providing us with information about FRST AID. We would like to thank Susan Nissan and Christine Wright of SHEP test development for their work on the Paraphrase item. They developed this item, and organized the item administration and scoring procedure. They were invaluable in developing rubrics so that the responses could be scored in GE FRST. We are grateful to Susanne Wolff of Research for scoring all of the Paraphrase item responses in GE FRST. We would like to thank Altamese Jackenthal of Research for doing all of the automatic scoring for GUIDES response data and Formulating-Hypotheses response data using GE FRST. We wish to thank Daniel Zuckerman of Research for writing the computer program to calculate Kappas.

Abstract

There is a considerable interest at Educational Testing Service (ETS) to include performance-based, natural language constructed-response items on standarized tests. Such items can be developed, but the projected time and costs required to have these items scored by human graders would be prohibitive. In order for ETS to include these types of items on standardized tests, automated scoring systems need to be developed and evaluated. Automated scoring systems could decrease the time and costs required for human graders to score these items. This report details the evaluation of a statistically-based scoring system, the General Electric Free-Response Scoring Tool (GE FRST). GE FRST was designed to score short-answer, constructed-responses of up to 17 words. The report describes how the system performs for responses on three different item types. For the sake of efficiency, it is important to evaluate systems on a number of item types to see if the system's scoring method can generalize to a number of item types.

Introduction

At Educational Testing Service (ETS), research is currently being done to try to develop performance-based constructed-response items, in which a short-answer natural language response is elicited. To make constructed-response items operational, it is essential to develop automated scoring systems to minimize the time and costs involved to have these items scored by human judges. This report is an evaluation of the General Electric Free Response Scoring Tool (GE FRST), a statistically-based natural language scoring system that is designed to score short answer, constructed-response items. It is the second generation of a linguistically-based scoring program for natural language constructed-responses called the Free-Response Scoring Tool (FRST). The essential difference between GE FRST and FRST is that GE FRST does not examine linguistic (e.g., syntactic and semantic structures) of responses in order to score them, but calculates similarity measures between responses based primarily on the lexical content (i.e., words) used in responses. FRST, on the other hand, makes scoring decisions based on a sublanguage composed of semantic rules that are developed based on responses for a particular item (see Kaplan and Bennett (1994) for a detailed discussion of FRST).

This report discusses GE FRST's scoring capability for three constructed-response item types: (a) an inferencing item called the Formulating-Hypotheses[1] (F-H) item, (b) a language proficiency item called the Paraphrase item, and (c) a reading comprehension item from an instruction and assessment program for remedial and developmental studies (GUIDES) which we will call the GUIDES item in this paper. Three F-H items were used in the evaluation: the *Police Officers item*, the *Minor Dutch Landscape Painters item* and the *Deer item*. Two Paraphrase items were used in this evaluation: the *Morels item* and the *Bebop item*. The F-H and Paraphrase items used in the evaluation are described later in the report. Six hundred and thirty-five responses previously collected for the GUIDES item were scored by GE FRST, and the results are reported in this evaluation.

There were two primary goals for this evaluation. First, we wanted to know how GE FRST's scoring decisions compared to human rater decisions for each set of responses, so that the system's performance could be evaluated. Secondly, we wanted to know how GE FRST's scoring decisions compared to FRST's scoring decisions for the GUIDES data and the F-H Police item data, used in the FRST evaluation. We used Kappas to measure agreement between human rater decisions and machine decisions. By using Kappas we could accomplish both goals. Kappa measurements (see Fleiss (1981) for a detailed discussion) will reveal the amount of agreement between human raters and GE FRST. Furthermore, Kappa measurements had been used in the previous study with FRST, so these results could be compared to the results in this current study.

[1] The F-H item is now called Generating Explanations.

Method

For each set of response data for the F-H item, the Paraphrase item, and the GUIDES item, the entire set of responses to be used in the evaluation was scored by human raters, according to rubrics (i.e., scoring keys) developed to categorize all of the data. The rubrics for the F-H data were complex, and composed of numerous categories. Human raters used these categories to score the F-H response data, and the same rubrics were used in the machine-scoring process. For the Paraphrase items, human raters scored each response as either "correct" or "incorrect." For machine scoring purposes, a more detailed rubric was developed collaboratively by test development staff and Research. For the GUIDES item, a multiple category rubric was created and used for both hand-scoring and machine-scoring.

In order to use GE FRST to score the items, response data was partitioned into a set of training data, a set of evaluation data, and a set of test data. The number of responses in each set was determined by the number of responses in the set of item responses. *Training data* is the set of exemplary hand-scored responses entered into GE FRST before automated scoring begins. The training set is used by GE FRST during the automatic scoring process as a model of response classifications. The *evaluation set* is used during the *Learn* procedure in GE FRST, described later in this section. The *test data set* is uncategorized data that is used to evaluate GE FRST's automatic scoring mechanism. In this study, GE FRST's categorizations for the test set were compared to the human graders' decisions to measure GE FRST's performance. In this study, the set of data chosen to represent training and evaluation data is manually scored using GE FRST.

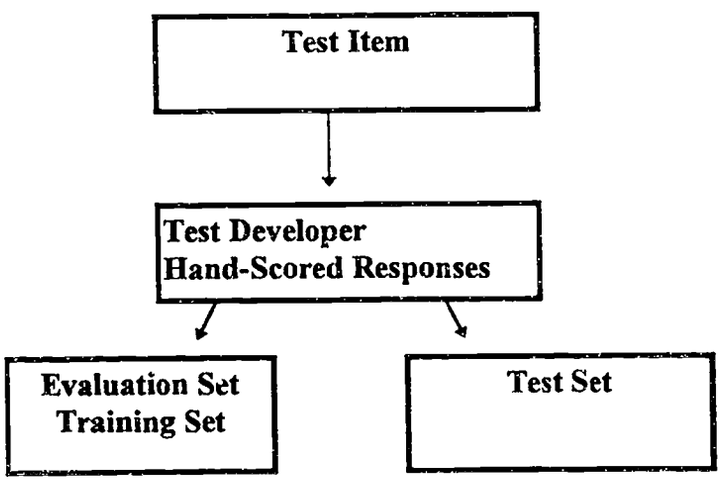

Figure 1: Response Data Partitioning for the GE FRST Evaluation

Next, the Learn procedure, described below, is run before the automatic scoring process begins. The placement of a response into a rubric category is controlled by confidence and matching parameters.[2] These parameters are set by the user of the system. To determine the best parameters, GE FRST has a procedure called Learn. The Learn procedure computes the best parameter configuration by using the scored training data as a model to appropriately categorize responses in the evaluation set. The system user selects the best configuration of parameters indicated by the results of the Learn procedure. An example configuration generated from the Learn procedure is illustrated in Figure 2. In this figure, the top row of numbers would be the *best configuration* since it has the highest level of placement and precision. This configuration would be used for automatic scoring.

POS	ATT	ACT	+COR	+INC	-COR	-INC	?COR	?INC	%ASGN	REC	PRE	INDEX MODE	CONFIDENCE METHOD	CONFIDENCE THRESHOLD
100	75	75	50	25	0	0	0	0	75.00	75.00	100.00	CONCEPT	PURITY	0.65
100	50	50	13	12	12	13	0	0	50.00	50.00	25.00	MORPH	PURITY	0.65

Figure 2: An Example Configuration from a *Learn* Procedure[3]

POS - Total number of responses in test data set
ATT - Total number of test data responses GE FRST *attempted* to categorized
ACT - Total Number of test data responses that GE FRST *actually* categorized
+COR - Responses scored correct by GE FRST and human graders
+INC - Responses scored incorrect by GE FRST and human graders
-COR - False positives
-INC - False negatives
?COR - Responses scored correct by a human which GE FRST did not score
?INC - Responses scored incorrect by a human which GE FRST did not score
%ASGN - Percent of actual category assignment
REC - Recall is the total number of correct category assignments
PRE - Precision is the total number of correct category assignments actually placed[4]

In the final stage of each scoring procedure, after the training data has been hand-scored, and the test data has been machine-scored, Kappas are calculated to determine the amount

---

[2] Matching and confidence parameters are discussed Kud, et al (1994).

[3] The Index Mode MORPH refers to analyses of words and their subparts (e.g., affixes); the Index Mode CONCEPT refers to analyses of terms through a hierarchy of conceptual relations; the Confidence Method PURITY looks for responses in training which have intersecting terms with test responses over each rubric category; Confidence Threshold .65 refers to the level of confidence that must be exceeded by GE FRST for a response to be placed automatically.

[4] During a *Learn*, a response may be placed in multiple categories if the level of confidence is ambiguous.

of agreement between GE FRST's scoring decisions and human rater decisions about the test data. To calculate Kappas, categorizations assigned to each test data response are collapsed into a "correct" or "incorrect" category assignment. A computer program was written to calculate Kappas automatically. Based on the results of the Kappa calculations for each item, GE FRST's performance was assessed.

The Formulating-Hypotheses Item

The Formulating-Hypotheses item (F-H) presents the examinee with a short text passage describing a situation. Based on the information in the passage, the examinee is prompted to produce reasons that explain the situation. Examinees are expected to use inference to generate creative short-answer free-responses.

Formulating-Hypotheses Response Data Set

Three of the eight F-H items (that is, the Police Officers, Minor Dutch Landscape Painters, and Deer items) used for the FRST evaluation (see Kaplan (1992) and Kaplan and Bennett (1994)) were used for the GE FRST evaluation.

A representative sample of 200 responses taken from the response sets of 30 examinee was used as training data. This was the same set of training data that was used in the original FRST study. The training set is divided into 2 sets, each one containing 100 responses: (a) the *training set*, and (b) the *evaluation set*. For the three items, test development staff hand-scored a set of training data and a set of evaluation data. Although it is somewhat confusing, this training set and evaluation set form the original training set used in the FRST study. The splitting of the original training set (the one used to train FRST) into two distinct sets was necessary to conform to GE FRST's training procedure. The test data set contains approximately 300 responses from the same 30 examinees whose responses were used to create the training and evaluation sets.

F-H Rubric Creation and Hand-Scoring F-H Response Data

One difference between the method for scoring F-H items and the other two items, is that the test developers who worked on this item were trained how to use GE FRST for rubric building and hand-scoring. They worked together on the tasks of rubric creation and hand-scoring. The process of rubric creation is described below. After rubrics were created, the test development staff used the hand-scoring mechanism in GE FRST to populate the rubric categories with test responses.

In GE FRST, rubrics are created as three-level trees. The highest node in the tree is a box identifying the *rubric name* which identifies the item; the intermediate nodes contain the *General Categories*. At the bottom level, examinee responses are actually categorized into *Specific Categories* which GE FRST refers to during the scoring process. The figure below illustrate the category structure in GE FRST. Categories can represent "correct" or "incorrect" responses. Below, the shaded boxes are categories for incorrect responses. Categories to represent incorrect responses are typically created to categorize irrelevant or incoherent response.

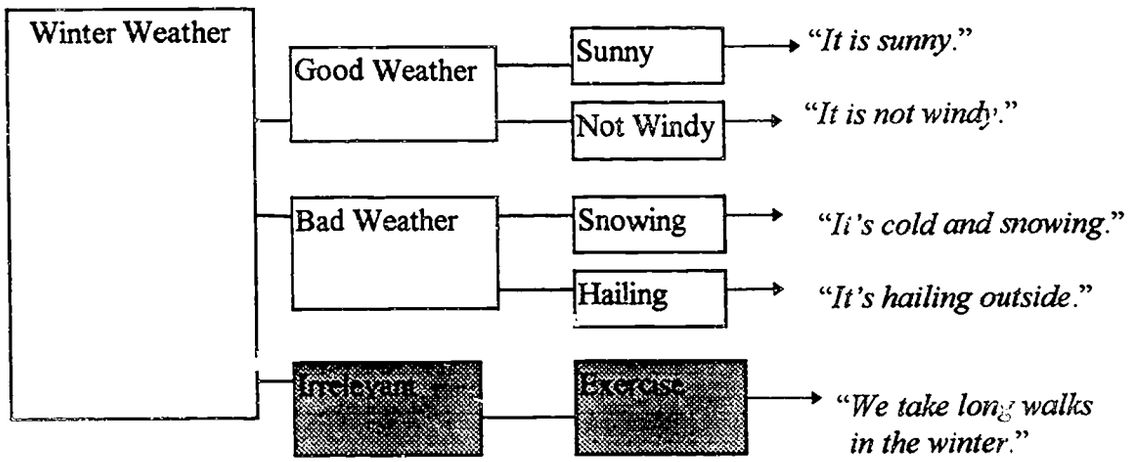

Figure 4: Illustration of GE FRST Categorization Hierarchy and Responses.

The test development staff used the functions provided by GE FRST to create the rubric structure described above for each item. Since more than one person worked on each rubric, copies were made of each rubric so that each person assigned to work on the development of a particular rubric could make appropriate rubric revisions and hand-score the test data. When the test development staff completed a rubric, they compared different versions of a rubric using a comparison program. The comparisons were used to discuss differences between rubrics and to resolve these differences to create a single acceptable rubric for each item.

When a rubric was created for an item, the test development staff used it to hand-score the training, evaluation, and test data.

General Comments About the GE FRST User Interface

The processes of automatic and hand-scoring using GE FRST are fairly straightforward, and once the user is familiarized with these procedures, the system runs fairly smoothly.

Overall, however, the interface is not an intuitive one, that is, a user cannot figure out how to use the system, on the fly, but rather, very detailed instructions must be followed to work with some procedures, such as procedures which involve *rubric editing*. Due to the lack of documentation from General Electric, we all learned how to use the system, for the most part, through verbal communication with the system developers at General Electric and through trial and error.

The test development staff kept a record of detailed notes describing how to facilitate the use of the system for their specific tasks. Based on the test development staff notes over a few months, it appeared that the most significant problems which they encountered in using the interface occurred when they needed to edit rubrics. Rubric creation for F-H items is a lengthy process which, at least, in this case, involved the input of more than one test developer for a single rubric. This being the case, as the test development staff worked both individually and together, the need often arose, to add, delete, and otherwise change the state of the rubric (e.g., rename rubric categories). The test development staff found that the interface mechanisms for rubric editing were problematic in that the procedure was too involved, and the system was not friendly toward any deviation from the instructions. That is, if the instructions were not followed exactly, it could result in a system crash, or even loss of data without warning. The test development staff have managed to work around these difficulties and with the most recent version of GE FRST, system crashes and data loss do not occur frequently.

The test development staff found the following features to be extremely useful in the rubric creation process. GE FRST allows the test developer to view the responses which are contained in the Specific Categories of the rubric hierarchy. Tests developers felt that this feature was crucial for the purposes of developing a conceptually defensible rubric. Since rubrics tend to be large, and often change many times before the final version, the test development staff found it to be useful to have a system which allowed them to see a picture of the rubrics as they were being developed.

FRST AID

To compensate for some of the features which GE FRST lacked, but were required by test developers for rubric creation, the test development staff created a system, FRST AID[5], a front-end human-scoring interface which was designed to be used along with GE FRST. FRST AID does not do automatic scoring. It is a PC-based program which permits use by multiple users.[6] FRST AID offers the following extra features.

(1). A scorer can tag any subset of hypotheses from the large group and assign it to a given category in one operation. (GE FRST offers this function only with the "Browse" feature, which is informative but often constraining.)

---

[5] FRST AID was developed by Lois Frankel of SHEP Test Development.

[6] Since GE FRST runs on Sparc Workstations and there is only one of these for the test development staff, only one test developer could use GE FRST at a time.

(2). Scorer's can review each other's rubrics and enter any discrepant judgments, which are identified as belonging to a second scorer.

(3). It allows scored responses to be marked as scored, and displayed in two places: (a) In the category(ies) to which they have been assigned by all scorers; and (b) In the list of all hypotheses in the data set. The scored hypotheses are marked so they are immediately distinguishable from the unscored ones. From this master list, it is also possible to view the entire history of a given response, across scorers and (where applicable) across multiple categories.

(4) Users can search for an individual response in order to:

(a). Examine its scoring history. (This functionality is crucial for resolving double-scored items and can facilitate Research staff's placement of "unknown" hypotheses into the categories with which human scorers have previously linked them.)

(b). Locate a response/responses containing a given word or phrase for more efficient rubric category-creation and placement of responses in categories.

(5). FRST AID automatically prints reports that show:

(a). A comparison of one or more scorings of a given set (across multiple placements and multiple scorers)

(b). A rubric with all categories and no hypotheses-contents

(c). A rubric with the hypothesis-contents of all categories

(d). A rubric with all hypotheses, and with each scorer identified (so that discrepancies can be resolved).

(e). For responses assigned to multiple categories, cross-reference information in the form of a "See also" note.

(6). Automatically calculates and displays the following scoring information for each candidate and for each scorer:

(a). Number of hypotheses submitted.
(b). Number of correct hypotheses.
(c). Number of incorrect hypotheses.

(d). Number of duplicate hypotheses.

Automatic Scoring of the F-H Response Data

GE FRST was used to automatically score the test data set. The system processes each response from the test data set by attempting to categorize the response into a rubric category. Responses not automatically categorized by GE FRST are left for the user to categorize. The rubrics prepared by the test development staff were used by GE FRST to automatically categorize responses.

Parameters Used to Score F-H Items Using GE FRST

For the Police Officers item and the Minor Dutch Landscape Painters item, the Learn procedure determined that the optimal parameters were: Confidence at .65; Matching Method is *Purity*; Index Mode is *Concept*; and, Minimum Size is 2. This means that GE FRST assigned responses to rubric categories for which it was over 65% confident in the category assignment; it used the Purity method (refer to footnote 3) to do matching; it looked at a semantic concept hierarchy to locate lexical items in training that were conceptually related to those in the test data set; and, Minimum Size is 2 means that GE FRST only considered rubric categories with at least two responses.

Responses for which GE FRST's confidence level did not exceed 65% had to be categorized by the system user. Responses not categorized by GE FRST are referred to as *unknowns*.

To ensure that the unknown test responses are categorized according to the test developers' criteria, we used their hand-scored data to categorize unknowns. A record was kept of unknowns.

Handling unknowns in this way ensured the training of GE FRST was consistent with the test development staff rubrics. Since GE FRST incrementally builds its internal rubric while scoring, it was essential to make sure that human categorization of unknowns was exactly the same as the categorizations assigned in the rubrics.

Table 1 : Number of Responses Automatically Scored (AS) and Manually Scored (MS) and Total Number of Responses (Total) for Each Item

	Minor Dutch Landscape Painters			Police Officers			Deer		
	AS	MS	Total	AS	MS	Total	AS	MS	Total
Binary Rubric	83%	17%	100%	80%	20%	100%	----------		
	190	38	228	216	53	269	----------		
Multi-Category Rubric	6%	94%	100%	20%	80%	100%	34%	66%	100%
	13	215	228	53	216	269	91	174	265

We calculated the Kappa value between the test developers' hand-scored responses and responses scored by GE FRST for both the *Police Officers* item and the *Minor Dutch Landscape Painters* item, using binary and multi-category rubrics. The *Deer* item was scored with a multi-category rubric only due to time constraints. In order to ensure an accurate Kappa, unknowns were removed from the Kappa calculation. This was necessary so that hand-scored unknowns did not get erroneously compared to the test developers' hand-scored version of the responses.

It is quite possible that occurrences of the same response will be categorized manually for the first occurrence and automatically for later occurrences. To ensure that only automatically placed responses remained in the Kappa comparison, we checked for exact duplicates in the scored test set before the extraction program was run. If the number of duplicates found exceeded the number of hand-scored responses, we referred to our notes and removed by hand, only the hand-scored occurrences of the particular response

After necessary responses were removed from both response sets, we calculated the Kappa measure for the subset of responses scored by GE FRST, and the corresponding subset of responses hand-scored by the test development staff.

Kappa Results Using the Binary Rubric

The Kappa results for both the *Police Officers* and *Minor Dutch Landscape Painters* data scored with the binary rubric were not significant. The table below shows these results. Kappas from the FRST evaluation are also included for purposes of comparison. FRST's Kappas are noted in parentheses, next to the GE FRST Kappas.

	Police Officers	Minor Dutch Landscape Painters
Kappa:	0.208 (FRST= 0.000)	0.000 (.FRST=0.26)
Standard Error	0.064	0.000
Total Mismatches	7	7
False Positives	5/7	7/7
False Negatives	2/7	0/7

Table 2: Kappa Comparison for GE FRST and Human Raters Using a Binary Rubric

The reason for GE FRST's apparently poor performance can be explained by the *sparse data problem*. For the sets of training and test responses for both of these items there is a profound lack of responses which are *incorrect*. No more than five to ten per cent of the responses are *incorrect*. Also, there were approximately 5 - 7 categories for classifying incorrect responses. Each category contained only 3 - 6 responses. This large number incorrect Specific Categories over a relatively small set of responses contributed to the sparse data problem. The system does not have sufficient training data even as scoring progresses to appropriately place the small number of responses which were scored *incorrect* by the test development staff. The system's scoring decisions about *incorrects* resulted in a number of *false positives*. Like FRST, GE FRST had difficulty categorizing responses classified as incorrect by a human rater.

Kappa Results for Minor Dutch Landscape Painters and Police Officers Using the Multi-Category Rubric

We scored the Minor Dutch Landscape Painters and Police Officers data using the multi-category rubric for two reasons. First, we wanted to test GE FRST's performance using a the rubric formulated by the test development staff. Second, we wanted to see if Kappas would improve using a multiple category rubric. We expected to improve the Kappa for a binary rubric by scoring Minor Dutch Landscape Painters using the multiple category rubric, and then collapsing the categories of the scored test data into *correct* and *incorrect* categories to calculate Kappa. We anticipated that GE FRST would have less trouble scoring test responses in a multiple category rubric, since no one category would contain 90 per cent of the responses as was the case with the binary rubric.

For Minor Dutch Landscape Painters, only 13 of 228 responses were scored automatically by GE FRST when using the multiple category rubric. This contained 44 Specific Categories GE FRST scored 53 responses out of 269 responses automatically for Police Officers for which the multiple category rubric contained 73 Specific Categories. For the Deer data, GE FRST scored 91 of 265 responses. The multiple category rubric contained 87 Specific Categories. The results are summarized in Table 3 below.

Table 3: Kappa Comparison for GE FRST and Human Raters Using a Multiple Category Rubric

	Minor Dutch Landscape Painters	Police Officers	Deer
Kappa	0.000	not computable	1.000
Standard Error	0.000	not computable	0.105
Mismatches	1	not computable	0
False Positives	1/1	not computable	0
False Negatives	0/1	not computable	0

For the *Minor Landscape Artists Painters* and *Police Officers* data, when we collapsed the categories into corresponding *correct* and *incorrect* categories we found that there were no occurrences of *incorrects* in the set of responses automatically scored by GE FRST. Again, this is certainly due to the fact that there is not a large enough number of incorrect responses in the training set or over the entire set of test responses in order for the system to be sufficiently trained. We expect that with more data and a proportionate number of correct and incorrect responses that we would be able to see improved results. For the *Deer* data, on the other hand, approximately 17.6% (16 of 91) of the data was categorized as *incorrect*. There were only two rubric categories for classifying incorrect responses. Of these incorrect responses, 14 of 16 were placed into a single category (i.e., Clever). A reduced, more generalized rubric seemed to contribute to GE FRST's performance in scoring the Deer data.

Evidence From F-H Items Illustrates That GE FRST Does Learn

One of the useful features of GE FRST, which demonstrates its ability to learn as the scoring process progresses, is that when a user is placing responses GE FRST will display its best guesses for categorizing a response when it cannot be placed. Appendix 1 illustrates GE FRST's incremental learning ability as it occurred during the scoring of the Minor Dutch Landscape Painters data using the multiple category rubric. Since the Confidence Threshold was set to .65 for this particular scoring session, GE FRST placed responses whose confidence exceeded this threshold. If GE FRST had an idea about where a response should be placed, but its level of confidence did not exceed the Confidence Threshold specified, it told the user where it guessed that it should be placed along with its level of confidence. For example, the numbers below each response listed in Appendix 1 illustrate how confident GE FRST was about placing that response into the category which appears next to the number. For example, in (1), below, GE FRST is 22 per cent confident that the response should be placed into the category, *Dealers, others, faked documents, works*.

(1).

Response: (A FEW GOOD ARTISTS THAT CHOSE NOT TO PUT THEIR NAMES ON THE PAINTING)

Guess: .22 Dealers, others, faked documents, works (BEST)[7]

As was mentioned earlier in this report, notes were taken during automatic scoring to document the following: (a) Responses which were not placed automatically by GE FRST; (b) For responses not placed automatically, when did GE FRST's guess match the test developer's categorization; and, (c) For responses not placed automatically, when did GE FRST's *most confident guess* match the test developer categorization. Recall, that although GE FRST does not place a response whose confidence decision does not exceed the Confidence Threshold, it gives information about where it would have placed the response given a lower Confidence Threshold. For example, if the Confidence Threshold parameter set by the user requires GE FRST to be more than 65% confident about the categorization of a response, but it is only 45 % confident, this information can be accessed by the user. Thus, GE FRST's *learning curve* can be observed. GE FRST's learning performance is summarized in Table 4.

Table 4

Responses Automatically Placed (AUTO) by GE FRST; A GE FRST Guess Matched Test Developers' Categorization (TD-GUESS); GE FRST's Best Guess Matched Test Developers' Categorization (TD-MATCH)[8]

	Minor Dutch Landscape Painters		Police Officers	
Total Responses	228		269	
	No.	%	No.	%
Total Guesses	168/228	74	135/269	84
AUTO	13/228	5	53/269	20
TD-GUESS	120/168	71	46/135	34
TD-MATCH	21/168	8	42/135	31

[7] BEST refers to GE FRST's highest (or best) confidence level for categorizing a response.

[8] Due to time constraints and limited resources, this information was not collected for the Deer data

The Paraphrase Item

The Paraphrase item is a language proficiency which tests a non-native English speaker's ability to understand information in a short speech. In this item, test-takers are required to listen to about 30 seconds of a speech on a specific topic. They are then asked to respond to questions about the speech. The topics of the Paraphrase items reported in this study were morel mushrooms and Charlie Parker, hence the names the *Morels item* and the *Bebop item*.

GE FRST and the Paraphrase Item

Approximately 185 responses were collected for each of the Morel and Bebop items. One hundred of the 185 responses were selected for training data from each response set. These responses were selected by a program which selects the 100 responses whose lexical items contribute the most unique lexical information. All responses in the training set were hand-scored based on a rubric designed by test development on the GE FRST system. Rubric categories were manually assigned to training responses according to test development staff specifications. The remaining, approximately 85 responses were used as input to GE FRST to be scored automatically. Parameters for automatic scoring were set according to results of GE FRST's Learn option discussed earlier in this report. Any responses which could not be automatically scored by GE FRST were scored manually according to test development specifications (i.e., the test development scoring key). Since we are primarily interested in GE FRST's automatic scoring decisions, as compared to human scoring decisions, manually scored responses were not included in the final analysis.

GE FRST Analyses Results for the Morel Item

The Morels Paraphrase item has three parts which we refer to as: a) *Study morels*; b) *Morel appearance*; and, c) *Find morels.* In each part of the item, a different question is asked which requires the examinee to paraphrase information that s/he has just heard in a short recorded speech.

The three Morel items are illustrated in Tables 5, 6, and 7, below, along with the acceptable responses provided by test development staff.

Table 5: Study Morels Item

Study Morels item
Why is it a good idea for the beginning mycologist to start by studying morels?
• easy to find
• easy to identify
• good to eat

Table 6: Morel Appearance Item

Morel Appearance Item
What do morels look like?
• 4 to 5 cm high
• 6 cm wide
• don't look like supermarket mushrooms
• cap is pitted, grooved or has holes
• conical cap

Table 7: Find Morels Item

Find Morels Item
Why are morels sometimes hard to find?
• people won't tell you where to look
• they're the same color as leaves

Analyses for the Morel Item

When all of the responses had been scored by GE FRST (either automatically or manually), Kappas were calculated to measure agreement between GE FRST and human graders. Table 8, below, shows the Kappas derived for all three Morel items.[9]

Table 8: Kappas for Morel Items

	Total Scored	Kappa	Standard Error	False Corrects	False Incorrects
Study Morels	11/85 (13%)	-0.138	0.279	1	2
*Morel Appearance	15/83 (18%)	0.762	0.251	1	0
*Find Morels	50/85 (59%)	0.593	0.141	3	3

*Kappa shows statistically significant agreement between human graders and GE FRST

Discussion

The results in Table 8 show that the Kappas for *Find Morels* and *Morel Appearance* are statistically significant, where K > .4. The Kappa for *Study Morels* was not significant, where K < .4. What we understand from these results is that there was a significant

[9] To derive Kappas, rubric categories which denote acceptable answers are collapsed into a general category called *correct* and all those categories denotes unacceptable answers are collapsed into a general category called *incorrect*. This replicates how the Kappas were calculated for the FRST study (Kaplan and Bennett, 1994).

amount of agreement between human grader scores and GE FRST scores for *Find Morels*, but this was not the case for *Study Morels* and *Morel Appearance*. We are not discouraged by the small number of responses which GE FRST scored for *Study Morels* and *Morel Appearance*, nor by the non-significant Kappa for *Study Morels*. We believe that these results can be explained by overly specified rubrics, in which General Categories contain multiple Specific Categories with overlapping content. The overly specified rubric is a function of the item in its current state. We believe that a different administration of this item designed to make rubrics more concise would improve GE FRST's performance.

In GE FRST, the content of responses is used to build the rubric. Presumably, the primary content in a response should be represented by a rubric category. Accordingly, the more constrained an item response set is, the smaller the rubric will be, and conversely, the wider the scope of the content in a response set, the larger the rubric will be.

As is illustrated in Tables 5, 6, and 7, the scope of the content over these items varies. *Find Morels* has 2 possible responses, *Study Morels* has 3 possible responses, and *Morel Appearance* has 5 possible responses. However, the format of the response sheet used in the pilot did not require the examinee to put each response on a separate line, so many of the responses contained multiple responses which could not be separated for scoring. GE FRST does not have a mechanism to extract response parts. For *Morel Appearance*, in particular, the rubric contained 31 rubric categories representing correct responses, instead of only 5 rubric categories.[10] This is due to the increased number of combinations of responses that can occur over 5 possibilities, for example, some rubric categories were the following. Perhaps, GE FRST could show improved performance (that is, score a higher percentage accurately) if rubric categories are not overly specified, so that General Categories did not contain Specific Categories with overlapping content.

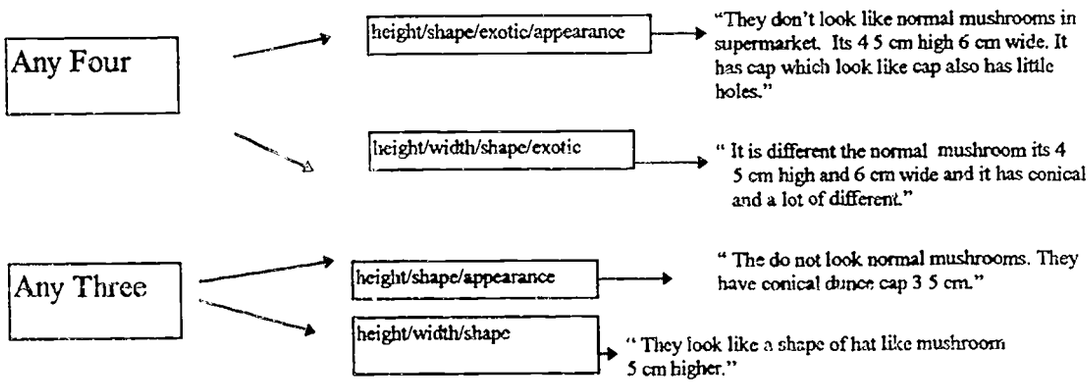

Figure 5: Excerpt from Morels Appearance Rubric and Sample Responses

[10] Had examinees been required to write one response per line, only 5 rubric categories would have been necessary, since combined responses would not have occurred. Multiple responses in a single response, however, could not have been forseen.

When rubrics are less generalized in GE FRST, it has been our experience that the system performance is degraded. For the three Morel items, we found that in the case of *Study Morels*, which had a insignificant Kappa, the largest clustering of correct responses occurred in two Specific Categories within a single General Category. These Specific Categories overlapped with regard to content. It appears that the overlap degraded GE FRST's performance.

Recall that GE FRST is a statistically-based system. To score a new response, it reviews all previously scored responses and their rubric categorizations. GE FRST looks for the closest similarity between previously scored responses and the current response. The current response is categorized the same way as previously scored similar responses. Therefore, a significant amount of overlap between responses over numerous rubric categories will confound GE FRST. The result is that either the system will be unable to score the current response automatically, or it will misplace the current response. It seems that in the case of *Study Morels,* that content overlap was problematic for GE FRST. For both *Find Morels* and *Morel Appearance*, there was little or no overlap between the most populated Specific Categories. Furthermore, the most populated Specific Categories in both *Find Morels* and *Morel Appearance* never occurred in the same General Category.

What is important to note with regard to the positive result obtained for *Find Morels* and *Morel Appearance*, is that the non-native grammar structures which occurred in these responses did not contribute to GE FRST's scoring performance. Another point to note is that although we did a significant amount of spelling correction before running the GE FRST application, not all spelling errors were caught. So, a small number of spelling errors did not affect GE FRST's performance either.

GE FRST Analyses Results for the Jazz (Bebop) Paraphrase Item

The same rubric building procedures, and scoring procedures were followed in order to score the Bebop item. Also, approximately the same number of responses, 85, were used as test data for GE FRST to score, and there were 100 training responses for each item. Tables 9,10,11,12, and 13 illustrate the item question and rubric.

Table 9: Bebop Characteristics Item

Bebop Characteristics Item
What are the musical characteristics of bebop?
• complexity
• vitality
• fast tempo
• rich sound

Table 10: Bebop Bird Item

Bebop Bird Item
Why was "Bird" a good nickname for Charlie Parker?
◦ The music he played suggested {flight, or grace, or freedom}

Table 11: Bebop Protest Item

Bebop Protest Item
The speaker states the "In some ways bop was a protest." What was the protest about ?
• World War II
• The commercialization of jazz
◦ Unhappiness with society

Table 12: Bebop Charlie Item

Bebop Charlie Item
The speaker claims that "Bop and Charlie Parker were made for each other." What does he mean?
• Charlie Parker was a soloist
• "Bop" was based on solo performances

Table 13: Bebop Soloist Item

Bebop Soloist Item
What does the soloist do during a be-bop performance?
• Improvises
• Leads the group/group follows him

The results of the Kappa calculation comparing agreement between GE FRST and human grader scores, are in Figure 6.

	Total Scored	Kappa	Standard Error	False Corrects	False Incorrects
Bebop Bird	60/81 (74%)	-0.027	0.102	4	1
*Bebop Protest	39/83 (47%)	0.473	0.160	1	1
*Bebop Charlie	13/79 (16%)	0.435	0.229	0	4
Bebop-Soloist	21/83 (25%)	0.000	0.000	1	0
*Bebop Characteristics	16/87 (18%)	1.00	0.25	0	0

Figure 6: Kappas for Bebop Items

* = Significant Kappa

Discussion

In reviewing response clustering for the Bebop items, we found the same trend for this group of items as for the Morel items. We found that for items which received a significant Kappa, the most populated Specific Categories in the rubrics did not occur in a single General Category. There was only one instance where two highly populated Specific Categories occurred in the same General Category. But, for this item, there were three other highly populated Specific Categories which occurred in unique General Categories. For *Bebop Bird* and *Bebop Soloist*, there was only a single correct Specific Category which was significantly more populated than either of the two incorrect Specific Categories. For these items, it appears that GE FRST did not have sufficient data to score incorrect responses accurately, due to a lack of incorrect responses. Thus, GE FRST's performance was degraded in this case due to a *sparse data problem*. We found that the sparse data problem occurred for the F-H data discussed earlier in the paper.

Overall, the data supports the initial conclusion (with regard to the Morel items) that the results would be improved with more clear-cut rubric categories where content overlap within rubric categories was either reduced or eliminated.

GUIDES Data and GE FRST

Recall that GUIDES is a program of instruction and assessment with writing, reading and study skills for remedial and developmental skills programs. The item used for this report was a reading comprehension item. The GUIDES data collected for this item are short-answer responses of up to 17 words. The item passage is about *scientists who were doing research about the possibility of growing food in salt water*. This data was originally analyzed using the FRST prototype. In the FRST study, 635 test-taker responses were used to test the system, and 100 test-taker responses were used to train the system. The same groupings of training and test data were used to train and to evaluate GE FRST's scoring capability. The multi-category rubric used in FRST was also used in GE FRST to categorize (i.e., score) the 635 responses in the set of test data. This multi-category rubric was collapsed into a binary rubric, so that it could be compared with the results of the FRST study.

Since the GUIDES data was elicited from a reading comprehension item, the data contains a significant amount of lexical overlap which seems to be common when examinees are asked to draw their responses from a single text. They often seem to just extract sentences or parts of sentences from the passage verbatim. The Kappa was insignificant for the GUIDES data scored by GE FRST as is illustrated in Figure 5 below. Most of the scoring errors were false corrects. This is most likely due to the fact that correct and incorrect answers had a large amount of lexical overlap, which can easily confound GE FRST. For instance, the response *Can we grow tomorrows food in today's saltwater?* is a correct response. The response *Can we grow tomorrows food in today's climate?*, however, is an incorrect response. But, GE FRST categorized the latter response as

correct since it had a significant amount of lexical overlap with the former response, as is illustrated from the underlined portions of the responses. It appears that a primary reason for the large number of false corrects and incorrects can be attributed to lexical overlap between correct and incorrect responses. Figure 7, below, shows the Kappa calculated for the GUIDES data.

	Total Scored	Kappa	Standard Error	False Corrects	False Incorrects
GUIDES	482 (76%)	0.037	0.016	245	4

Figure 7: **Kappa for GUIDES Data**

It appears that again, there is a *sparse data problem*, so the statistical approach used by the system to categorize responses is degraded. Perhaps an approach involving more linguistic information (e.g., syntactic and semantic) would be better able to handle the lexical overlap and to accommodate the sparse data problem.

Conclusion

From a research perspective, GE FRST has proven to be a useful research tool, from which we have gained a considerable amount of practical knowledge regarding what natural language processing components are relevant for evaluating content in F-H items. More generally, we have also made a large jump in our understanding of the usefulness of statistical methods for purposes of natural language analysis on the data used in this study. The knowledge which we have gained from the results of this evaluation will facilitate our on-going research in natural language understanding, specifically for the purpose of scoring natural language constructed-responses on exams.

From a practical point of view, GE FRST's graphical interface mechanisms for building rubrics has clearly facilitated certain aspects of the complex and continuous task of rubric development for test developers. The GE FRST interface also broadened the test developers' understanding of their needs with regard to rubric creation and hand-scoring. By using GE FRST, the test development staff discovered that although GE FRST had many useful features, it also lacked some features which they believed would facilitate the processes of rubric creation and hand-scoring. Based on what they learned about rubric creation and hand-scoring by using GE FRST, the test development staff who worked on the F-H item developed FRST AID, a PC-based tool which they developed to facilitate the processes of rubric creation and hand-scoring for F-H responses. The development of FRST AID points out that much can be learned from prototype tools.

In terms of GE FRST's actual performance, the system does not appear to exceed the performance of FRST, at least for the F-H items. Since GE FRST's methods of analysis are statistically based, with regard to the F-H item, we would see improvement in its overall performance if it were provided with more training data, as well as a more balanced set of training data. That is, its training set should contain a proportional number

of incorrect and correct responses, so that the system could learn the difference between a correct and incorrect response. We certainly see a trend in this direction with regard to the Paraphrase item type, for which GE FRST performs well if the rubric is generalized and the number of correct and incorrect responses is reasonably balanced. In addition, GE FRST appears to be able to handle non-native and native speaker data equally well. This is due to the fact that it does not consider the syntactic structure of a response.

If further evaluation of GE FRST is done on F-H data, we would build the rubrics based on the training data, that is, the set of data which GE FRST uses to make categorization decisions during automatic scoring. Previously, the test data for F-H items was used to build the rubric, and when the training data was actually input into GE FRST, many of the rubric categories were not represented. So, the system had no way to learn about what should be placed into these unrepresented categories until some test responses had been manually placed in those categories during the automatic scoring process.

With regard to the GUIDES data, perhaps a more linguistically-based approach will prove to be more efficient. We are currently developing such an approach using the Microsoft Natural Language Processing tool (MSNLP) which produces syntactic and semantic representations for text. We plan to test this new approach on the F-H data and the Morel and Bebop data used in the GE FRST evaluation.

It is clear, overall, from this preliminary evaluation, that significantly larger data sets would be required, in general, to make decisions about GE FRST's capabilities in an operational setting.

References

Fleiss, Joseph L. (1981). Statistical Methods for Rates and Proportions New York: John Wiley and Sons.

Kaplan, Randy M. and Randy Elliot Bennett. (1994). Using the Free-Response Scoring Tool To Automatically Score the Formulating-Hypotheses Item. (RR-94-08). Princeton, NJ: Educational Testing Service.

Kud, Jacquelynne M., George R. Krupka, and Lisa F. Rau. (1994). *Methods for Categorizing Short Answer Responses* in Proceedings of the Educational Testing Service Conference on Natural Language Processing Techniques and Technology in Education and Assessment, Princeton, New Jersey.

Appendix 1: Learning Information about Responses which GE FRST Recognized[11] for a Sample of 10 Examinees for the Minor Dutch Landscape Painters Item

This Appendix shows the *examinee response* in parentheses above a *number* which represents the confidence threshold and the *category* with which the confidence threshold is matched.

(BEST/TD) = GE FRST's best guess matched the TD categorization.
(BEST) = GE FRST's best guess
(TD) = GE FRST's guess matched the TD categorization.

Examinee ID#: 9991001:

(SO FEW MINOR DUTCH LANDSCAPE PAINTERS THAT PAINTED IN THIS STYLE)
.33 Have distinct styles, easier to identify (minor styles less familiar)
.33 Worth more money, more valuable, in safer environment
.33 Had commissions, money to paint

(A FEW GOOD ARTISTS THAT CHOSE NOT TO PUT THEIR NAMES ON THE PAINTINGS)
.22 Dealers, others, faked documents, works (BEST)

(THE INCREASED NUMBER OF GOOD ARTISTS THAT CHOSE TO PUT THEIR NAME ON EVERY WORK)
.22 Dealers, others, faked documents, works (BEST)

(ONE MINOR ARTISTS DID NOT PAINT IN THIS GENRE AS MUCH AS MAJOR)
.15 Had commissions, money to paint (BEST)

Examinee ID#: 9999003:

(MAJOR ARTISTS WERE COMMISSIONED TO PAINT MORE DURING THE TIME)

[11] Responses not noted here were either automatically placed by GE FRST or not recognized by GE FRST.

.18 Had commissions, money to paint (BEST/TD)

(SOME LESSER ARTISTS PAINTED FOR THE MAJOR ARTISTS WHO SIGNED THEIR NAMES)
.16 Dealers, others, faked documents, works (BEST)

Examinee ID#: 9999043:

(PEOPLE DOING ATTRIBUTIONS ARE BIASED TOWARD MAJOR ARTISTS BECAUSE THESE ARE MORE EXCITING AND VALUABLE)
.15 Had commissions, money to paint (BEST)
.09 Scholars more interested in attributing major, not minor (TD)

(PAINTINGS BY MAJOR ARTISTS MORE FREQUENTLY SURVIVED OVER TIME BECAUSE OF THEIR GREATER VALUE)
.15 Had commissions, money to paint (BEST)

(MINOR ARTISTS FREQUENTLY COPIED THE STYLE OF MAJOR ARTISTS SO THAT THEY ARE FREQUENTLY MISATTRIBUTED)
.33 Non-explanatory response
.33 Explains the reverse situation
.33 Had commissions, money to paint

(SEVENTEENTH CENTURY ART DEALERS DECEIVED CUSTOMERS MISATTRIBUTING MINOR WORKS TO MAJOR ARTISTS ON BILLS OF SALE)
.33 Dealers, others, faked documents, works (BEST/TD)

Examinee ID#: 9999059:

(COLLECTORS ATTRIBUTE PAINTINGS TO MAJOR ARTISTS WITHOUT GOOD EVIDENCE TO MAKE THE PAINTINGS MORE VALUABLE)
.17 Dealers, others, faked documents, works (BEST/TD)
(MAJOR ARTISTS PAINTED MORE BECAUSE OF A HIGHER DEMAND FOR THEIR WORK)
.13 Dealers, others, faked documents, works (BEST)
.12 Had commissions, money to paint (TD)

(MINOR PAINTERS WOULD SIGN THEIR TEACHERS NAMES TO PAINTINGS TO MAKE MONEY)
.25 Dealers, others, faked documents, works (BEST)

Examinee ID#: 9999085:

(MAJOR ARTISTS RECEIVED MORE COMMISSIONS FOR NEW PAINTINGS)
.15 Had commissions, money to paint (BEST/TD)

(SCHOLARS ARE MORE FAMILIAR WITH MAJOR ARTISTS SO EASIER TO IDENTIFY THEIR PAINTINGS)
.16 Dealers, others, faked documents, works (BEST)
.07 Have distinct styles, easier to identify (minor styles less familiar) (TD)

(MINOR ARTISTS EMULATED MAJOR ARTISTS SO THEIR WORK IS EASILY MISTAKEN FOR MAJOR ARTISTS)
.25 Had commissions, money to paint (BEST)
.12 Minor Artists copied or had styles that were similar to those of major Artists (TD)

Examinee ID#: 9999086:

(MINOR ARTISTS HAVE LESS WORKS TO BE COMPARED TO)
.20 Had commissions, money to paint (BEST)

(MINOR ARTISTS ATTEMPTED TO COPY THE WORKS OF THE BETTER KNOWN OR MAJOR ARTISTS)
.15 Dealers, others, faked documents, works (BEST)
.05 Minor Artists copied or had styles that were similar to those of major Artists (TD)

(MAJOR ARTISTS WERE COMMISSIONED BY NOBILITY WHO COULD AFFORD THE PAYMENT FOR OUTSTANDING WORKS)
.60 Had commissions, money to paint (BEST/TD)

(MAJOR ARTISTS WERE MORE LIKELY TO SIGN THERE PAINTINGS IN THE HOPES OF PRODUCING MORE)
.16 Had commissions, money to paint (BEST)
.05 Major Artists more likely to sign (minors didn't sign or used pseudonyms) (TD)