Submitted for publication to the: *Proceedings of the 28th Annual Cognitive Science Society Conference.*

An Empirical Feasibility Study of the ARCADE System

Richard M. Golden (golden@utdallas.edu)
School of Behavioral and Brain Sciences (GR4.1), UTD, Box 830688, Richardson, TX 75083-0688

Susan R. Goldman (sgoldman@uic.edu)
Department of Psychology (MC 285), 1007 W. Harrison Street, University of Illinois, Chicago, IL 60607-7137

Abstract

This paper introduces the ARCADE (Automated Reading Comprehension Assessment and Diagnostic Evaluation) system which is an automated psychometric diagnostic reading comprehension assessment tool based upon contempory theories of reading comprehension. ARCADE attempts to identify the presence of particular components of a reader's understanding of a text from open-ended free response data. An empirical evaluation of the ARCADE system showed that ARCADE could use student free response data to cluster students along meaningful dimensions of comprehension. In addition, directions for future research on the ARCADE project were clearly identified.

Introduction

There are a number of ways to define *reading comprehension assessment*. A strength of standardized assessment tests is that they provide reliable assessments of reading achievement through the use of psychometric modeling methodologies for equating items and estimating subject-specific ability parameters. However, standardized assessments of reading comprehension have limited validity because they are based on a one-dimensional ability scale of measurement for the purposes of quantitative analysis. That is, such tests focus upon whether an examinee's reading comprehension answer is correct or incorrect and report a single score.

In contrast, cognitive, discourse, and educational research indicates the importance of distinguishing among different levels of comprehension. At the basic level, comprehension focuses on what the text actually says (the literal meaning or textbase). At more complex levels, comprehension focuses on thinking and reasoning that integrate text information with information in other texts and with appropriate prior knowledge (Coté, Goldman, & Saul, 1998). When readers understand texts at complex levels, they have understood the meaning - what the text said and its relation to referents in the world, and have constructed one (or more) interpretations of the text. Together, meaning and interpretation constitute the reader's situation model. Especially for diagnostic purposes, it would be very desirable if reading comprehension assessments captured these multiple dimensions of understanding. By providing profiles of readers that reveal both meaning and interpretive understanding, such assessments would provide valuable information that classroom teachers could use to inform differentiated instruction and improve student learning.

The goal of ARCADE (Automated Reading Comprehension Assessment and Diagnostic Evaluation) is to instantiate a computationally *automated* and psychometrically valid multidimensional diagnostic reading comprehension assessment that can create profiles of readers based on the quality of their understanding. ARCADE assesses complex comprehension by identifying the presence of meaning (textbase elements) and interpretive (integrated knowledge elements) components of a reader's situation model. It does so by drawing on discourse analytic and computational modeling techniques to infer these components from readers free responses to questions about what they have read.

ARCADE System Methodology

Data-Informed Situation Model Specification

There are a number of challenges associated with the analysis of free response data, especially that generated by children and adolescents. The first is a computational one: existing natural language understanding systems (without substantial modifications) will have considerable difficulty processing the *raw* text of children's free responses which often contain misspellings, ungrammatical sentences, odd referential relationships, and ill-formed ideas. A second challenge concerns the "standard" against which children's responses are compared. It is common practice in discourse and educational research to compare the semantic content of the text input – of what the text *said*, to that in the free responses (Goldman & Wiley, 2004). In doing so, human coders are faced with complex semantic decisions about statements in free responses that do not appear to "match" text input. Many of these "nonmatching" statements reflect inferences based on what was in the text and many reflect inferences that integrate readers' prior knowledge. Still other "nonmatching" statements, may, in fact be entirely consistent with the explicit semantic content of the text but have been expressed in a novel manner by the children. Thus, "nonmatching" statements are particularly challenging when the text is lengthy or leaves open a number of interpretive possibilities for several reasons.

First, readers frequently summarize the meaning of multiple sentences from the input text in *summarizing* sentences that are not good matches to any of the sentences from the input text. Second, there is a wide range of prior knowledge inferences that readers could make for any given text. The challenge is specifying which of these is warranted by the text based upon personal experiences outside the text, and which are simply not consistent or plausible given the information in the text. Third, presented text information

accomplishes some particular function (or functional node) in the text (e.g, conveys setting information, establishes character(s)' goals, relates the consequence of a series of actions). In a free response a reader might accomplish these functions by including information that was in the text or by including inferred information that accomplishes the same function. In the latter case, it is redundant for the reader to also include the information that was presented; however, the function has been filled by the inference and a coherent situation model can be formed. (If the inferred information is not warranted by the text, one might say a distorted situation model results.) Inferences, especially knowledge-based inferences, introduce wide variation in the content of readers' free responses. Thus, it can be difficult to estimate the content and extent of readers' situation models.

In the face of these challenges and complexities, ARCADE relies on human analysis of the text semantics in conjunction with readers' free responses to construct a set of abstract nodes that reflect functional elements of the situation model. In this paper we describe the development and testing of this process on one narrative story for which fifth and seventh grade students provided free response data. Subsets of the behavioral data were used to "train" the computational model and other subsets were used to test the performance of the model.

Behavioral Data

In the study reported here, students from the 5th and 7th grades from three schools SD (63 students), JX (43 students), and PA (62 students) read a narrative text that was selected because it left a good bit of room for interpretation and dealt with issues and feelings that tend to interest adolescents. The text, "A Rice Sandwich" by Sandra Cisneros (1984), is about a girl named Esperanza wants to be like the children at school who do not have to go home for lunch. Esperanza begs her mother to let her eat at school, and her mother finally agrees. However, the principal of the school still will not permit Esperanza to eat in the cafeteria on a regular basis because she lives in the wrong part of town, too close to the school. At the end of the story, Esperanza does not want to eat in the cafeteria. The text is not explicit about why Esperanza changed her mind about eating in the cafeteria and there are several other places where there is room for interpretation, increasing the likelihood that readers would make knowledge-based inferences. The actual text passage consisted of 53 sentences, 719 words, and had a Flesch-Kincaid Grade Level readability index of 4.5 (approximately a 4th or 5th grade reading level).

After reading the text, the students were asked two questions. The first question was: "E*xplain Esperanza's feelings about eating at school at the beginning and at the end of the story."* The second question was: *"Explain Esperanza's mother's reaction when Esperanza tells her she wants to eat at school."* Students were allowed to refer to the text while composing their responses.

Text and Free Response Analyses

An *abstract story grammar analysis* based upon the text was done to identify the major functional plot elements of the story: Episodes, Initiating Events, Internal Responses (including goals), Attempts, and Consequences. These plot elements are consistent with a number of story grammar analyses of stories (e.g., Mandler & Johnson, 1977; Stein & Glenn, 1979).

These *Abstract Story Grammar Categories (ASGC)* were instantiated by 12 different classes of semantic information (e.g., emotions, cognitions, events), which we labeled as abstract story grammar (ASG) nodes. Each of these nodes might be manifest in students' responses by specific statements that were (i) very close matches to the presented text or by logical connections or summaries of what was presented, called Text-Based Inference (TBI) in this feasibility study; and/or (ii) inferences based on prior knowledge, called Knowledge-Based Inference (KBI).

GRADE 7 SUBJECT #3 Q2
KBI[5.2] *Esperanza's mother's reaction was that she was shocked .*
TBI[6.1] *She didn't want more work at first*
RN *but*
TBI[7.1] *she din't so she reluctantly gave in .*
KBI[4.1] *She din't know why her daughter wanted to eat at school*
RN *but*
KBI[7.2] *she could tell that she really wanted to*
KBI[7.2] *and a mother can't always say no .*
KBI[7.2] *Sometimes they just have ti give in*

Figure 1: Each student's free response data was modeled as an ordered sequence of complex proposition nodes. The notation KBI[5.2] means the second type of complex proposition in the fifth ASGC category of type KBI.

The range of ASG nodes included in the situation model was constrained by the behavioral data: If more than one student response included a KBI that fulfilled one of the ASG nodes, then it was included in the analytic template for the story; otherwise, the ASG node was manifest only in TBI nodes. Specific statements in the students' free responses were coded into complex propositions determined to semantically fill either a TBI or KBI ASG node and indexed accordingly. Figure 1 illustrates a typical analysis of a student's free response data. In addition, Figure 1 also illustrates the complexity of this data set which contains numerous ungrammatical sentences, misspelled words, and novel ways of expressing the same idea. There were 55 complex propositions which could be assigned to a clause in the student free response data.

Figure 2 shows the ASGCs and the TBI and KBI ASG nodes assigned to each ASGC which were obtained as a result of semantic analyses of the text and student response data. As shown in Figure 2, the human coding analysis of the behavioral data yielded 12 ASGCs, 12 TBI ASG nodes associated with each of the 12 ASGCs, and 9 KBI ASG

nodes associated with 9 of the ASGCs. Note that three of the ASGCs were not assigned KBI ASG nodes since examples of such KBI ASG nodes were not present in the student free response data. In addition, Figure 2 illustrates a representative data analysis regarding how the complex propositions in Figure 1 are represented as ASG nodes. For example, the complex propositions KBI[7.1] and KBI[7.2] are treated as members of an equivalence class of complex propositions which is labeled KBI[7]. The KBI[7] equivalence class corresponds to a particular KBI ASG node. Figure 2 also illustrates how the presence and ordering of the ASG nodes in Figure 1 is identified by an ASCG analysis. Specifically, ASG nodes present in the student's response data in Figure 1 are drawn as circles composed of dots (e.g., KBI[4], KBI[5]) while ASG nodes not present in the student's response data are drawn as circles composed of solid lines (e.g., TBI[3], TBI[4]). Semantic connections between adjacent complex proposition nodes in the student response data (Figure 1) which involve KBI nodes are classified as KBI connections and are represented by thick solid arrows in Figure 2 (e.g., connection from KBI[4] to KBI[5]). Semantic connections between adjacent complex proposition nodes in the student response data (Figure 1) which only involve TBI nodes are classified as TBI connections and are represented by thin solid arrows (e.g., connection from TBI[6] to TBI[7]). This type of analysis allows the student response data to be assessed in terms of the degree to which TBI semantic structure and KBI semantic structure influence the organization of student response data.

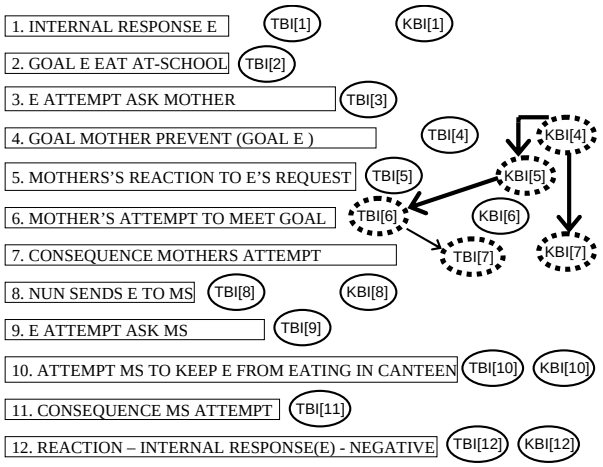

Figure 2: The Abstract Story Grammar Categories (ASGCs) shown here were derived from semantic analysis of the text and student response data. This figure also illustrates how the ASGCs are used to identify sequential structure in student response data presented in Figure 1.

ARCADE System

The ARCADE system is intended to automatically implement the process sketched in the previous section. Within the ARCADE framework, students would answer open-ended questions about a text which has been analyzed using an ASCG. The ARCADE system would then estimate for each student the relative impact of TBI and KBI influence factors based upon an analysis of the presence and ordering of the ASG nodes in the student's response data. The current implementation of ARCADE involves two stages. In the first stage, the ASMURF (Annotated Semantic Markov Utterance Random Field) system (Golden, 2006a) is used to identify a sequence of complex propositions for each student's response as in Figure 1. In the second stage of analysis, Golden's (1998, 2006b) KDC (Knowledge Digraph Contribution) analysis is used to compute the relative impact of TBI and KBI factors. Once these factors are assessed for each student, this information is available to provide feedback to classroom teachers in the form of suggested teaching strategies for specific groupings of students whose response data has similar TBI and KBI characteristics.

Automatic Semantic Annotation of Response Data

The ASMURF system was used to identify complex proposition sequences in the free response data for the purposes of automatically implementing the analysis in Figure 1. The essential idea of the ASMURF methodology is easy to explain. Key words (and misspelled words) are annotated as particular word-senses or "word-concepts". Then subsequences of word-senses corresponding to exactly one mental or physical action are annotated as particular "simple propositions". Subsequences of "simple propositions" are annotated as particular "complex propositions". Finally, equivalence classes of complex propositions were defined and labeled as ASG nodes. After semantic annotation was completed, first-order, second-order, and third-order statistical correlations between the various semantic annotations and words are learned. These estimated correlations are then used to automatically parse and semantically annotate novel word sequences.

Identifying Situation Models

The KDC system implements the analysis in Figure 2 by taking the complex propositions identified by ASMURF, mapping them into ASG nodes, and then looking for the presence or absence of the ASG nodes and how they are ordered. This produces a mapping of the free response data into a TBI influence measure reflecting the structure of the original text and a KBI influence measure reflecting the integration of prior knowledge.

KDC analysis not only matches sequences to graph structures such as that depicted in Figure 2 but also computes the unique maximum likelihood estimates of the link strengths in these graphs under the specific probabilistic modeling assumptions of KDC analysis (Golden, 2006b). Briefly, KDC may be viewed as a type of constrained multinomial logistic regression where the "beta weights" of the regression model correspond to link strengths. Thus, statistical model selection test and hypothesis testing procedures are available for psychometric analysis purposes within the KDC framework.

Identifying Student-Specific Situation Models

The estimation of the group-specific situation model is analogous to the estimation of item parameters in item-response theory (IRT) from group data. Like IRT, student-specific parameters can be estimated as well. However, unlike IRT, the concept of "ability" is absent from the ARCADE comprehension theory. Rather, the latent student-specific parameters are called "contribution weights" which represent the influence of the TBI and KBI dimensions of comprehension. For example, a student whose production data consists entirely of TBI ASGC propositions would have his (or her) TBI contribution weight estimated to be equal to zero. Golden (2006b) shows using theorems developed by Golden (2003) that not only are these parameter estimates generally uniquely determinable from the data but these parameter estimates are also maximum likelihood estimates whose asymptotic distributions can be characterized.

Results and Discussion

ASMURF Proposition Detection Performance

In order to quantify the performance of the ASMURF system, the recall and false alarm performance of the ASMURF system was evaluated on both training and test data sets. The ASMURF system computes a confidence level indicating its belief in the correctness of its choice of complex proposition. If the confidence level for a particular complex proposition semantic annotation exceeds the system's identification threshold value θ, then the system reports the presence of that complex proposition. By systematically varying θ, a response operating characteristic (ROC) curve for the ASMURF classifier system can be constructed.

The ROC curve displays the probability of correct identification of a proposition in a student's response given the human semantic annotator says that proposition is actually present ("recall rate") for a particular value of θ and the probability of false identification of a proposition in the student's response given the human semantic annotator says that proposition is not present ("false alarm rate") for a particular value of θ. From the ROC curve, an optimal threshold value θ^* may be computed which simultaneously maximizes recall rate while minimizing false alarm rate. In addition, a commonly used statistic in characterizing information retrieval systems called the "precision" was computed. The "precision" is the probability that the ASMURF system correctly identifies a proposition in a student's response given the number of propositions the human semantic annotator says which are present in the student's response.

Both training and test data were parsed into clauses corresponding to complex propositions by the human semantic annotators for evaluating the system's performance at decomposing complex propositions into simple propositions and semantically annotating the resulting decomposition.

Given the ASGC developed using the entire data set, the ASMURF system was trained on the SD data set and the optimal threshold θ^* for the SD data set was computed,. Given θ^*, the recall and false alarm rate using this training-set derived optimal threshold could then be computed for the training data (SD) and the test data (PA, JX). This procedure was then repeated by training on the PA data and testing on the SD and JX data as well as training on the JX data and testing on the SD and PA data. These results were then averaged to obtain recall, false alarm, and precision rates with standard errors.

The recall rate on the training data (62% ± 2.2%) was comparable to the recall rate on the test data (60% ± 1%) . This means that when a human coder decided a particular complex proposition was present in a particular student's free response, ASMURF would correctly decide that complex proposition (out of a possible set of 55 complex propositions) was present in the student's free response data about 60% of the time. The false alarm rate on the training data (37% ± 2.6%) was comparable to the false alarm rate on the test data (37% ± 2.0%). This means that when a human coder decided a particular complex proposition was absent in a particular student's free response data, ASMURF would incorrectly decided that complex proposition was present about 37% of the time. The precision rate on the training data (69% ± 1.7%) was slightly greater than the precision rate on the test data (60% ± 1.3%). This means that the percentage of propositions correctly identified in a student's response by ASMURF (out of the set of complex propositions identified as presented by the human coder in that response) on a test data set was 60%. Note that the roughly comparable performance levels on the training and test data indicate that the system was not "over-fitting" the data.

These performance level statistics are promising but clearly indicate the need for additional development of the ASMURF system. Indeed, these statistics are consistent with a qualitative analysis of the system's processing results. Many of the semantic annotations generated by the system would not be considered sensible by a human judge.

KDC Models of ASG Node Presence and Order

The goal of the KDC analysis is to take the complex propositions generated by the ASMURF analysis and attempt to automatically identify ASGC connections as illustrated in Figure 2.

To achieve this objective, the connection weights among and between TBI ASGC proposition nodes and KBI ASGC proposition nodes were simultaneously estimated using maximum likelihood estimation under the KDC probability modeling assumptions (see Golden, 2006b, for additional details) using the SD data set with the regularization term set to 100. As a result of this estimation process, a connection weight matrix for the TBI dimension and a connection weight matrix for the KBI dimension were obtained.

Three variations of these connection weight matrices were then considered: (1) the node presence model, (2) the node order model, and (3) the node presence and order model. The *node presence model* effectively measures the presence or absence of TBI and KBI ASGC nodes in student free

response data. The *node order model* effectively measures the degree to which the order of TBI and KBI ASGC nodes in the student free response data conforms to the connections in the knowledge digraph specifications (see Figure 2). The *node presence and order model* is a hybrid model which incorporates both sources of node presence and order. All three of the models are two parameter models where one parameter (called the "TBI" contribution weight) indicates the predictiveness of the TBI connection weight matrix while the other parameter (called the "KBI" contribution weight) indicates the predictiveness of the KBI connection weight matrix.

Sophisticated model selection criteria were used for the purpose of comparing competing KDC probability models (see Golden, 2006b, for specific mathematical details). Differences between model selection criteria were tested using Golden's (2003) DRMST (Discrepancy Risk Model Selection Test). Using the Generalized Bayesian Information Criterion (GBIC) for model selection, the node presence and order model provided a better fit (GBIC fit = 2.13) than the node order model (GBIC fit = 2.37) ($p < 0.05$). In addition, the node presence and order model provided a better fit (GBIC fit = 2.13) than the node presence model (GBIC fit = 2.31) ($p < 0.05$). Similarly, using a Generalized Akaike Information Criterion (GAIC), the node presence and order model provided a better fit (GAIC fit = 2.13) than the node order model (GAIC fit = 2.38) ($p < 0.05$). In addition, the node presence and order model provided a better fit (GAIC fit = 2.13) than the node quantity model (GAIC fit = 2.31) ($p < 0.05$).

Thus these findings show that *both the presence and the ordering* of ASG nodes in the student production data could be predicted in part by the ASGC analysis. Moreover, these results are consistent with numerous studies from the text comprehension literature which demonstrate that the order of propositions mentioned by subjects is often reflective of the semantic organization of the subject's situation model.

KDC Clustering of Students with Similar TBI and KBI Comprehension Dimensions

The long-term goal of the ARCADE project is to develop a system which can automatically process student free response data and group students with similar situation models and suggest appropriate instructional strategies for each student group by understanding the type of situation model shared by students within a group. For example, optimized instructional strategies designed for students with low KBI situation model components will look quite different from optimized instructional strategies designed for students with low TBI situation model components.

In order to evaluate the performance of the system from an educational technology perspective, the node presence and ordering model developed from the SD school data was used to estimate a unique TBI and a unique KBI contribution weight for the ASMURF annotated data for each student from the PA and JX schools. The KDC analysis program then uses a customized agglomerative cluster analysis which works by merging subgroups to minimize between-cluster variance.

The results of the cluster analysis are presented in Figure 3. Each student is represented by a circle in this cluster analysis with a particular KBI and TBI contribution weight. The cluster with the smallest circles corresponds to a group of students with large KBI and relatively low TBI weights. The cluster with the largest circles corresponds to students with moderate KBI and TBI scores. The seven medium-sized circles corresponds to students with relatively low KBI scores but larger TBI scores.

In order to evaluate the validity of the cluster analysis results, the node presence and order model developed using the SD school data was used to compute KBI and TBI contribution weights for each student from the PA and JX schools using the human annotated data as well as the ASMURF annotated data. Thus, the effectiveness of the ASMURF system in generating semantic annotations which are quantitatively equivalent (in contribution weight space) to that of the human semantic annotators could be evaluated.

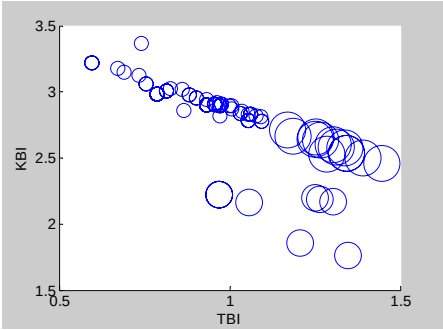

Figure 3: Three clusters of students identified by ARCADE. Students within a cluster are classified as having similar situation models and are associated with circles of the same radius.

Using the PA and JX student response data and the node presence and order model developed using SD data, the TBI contribution weights computed using ASMURF annotated data were positively correlated with TBI contribution weights using human annotated data ($r(103) = 0.96$, $p < 0.05$ for a no-intercept model). Similarly, the KBI contribution weights computed using ASMURF annotated data were positively correlated with KBI contribution weights using human annotated data ($r(103) = 0.98$, $p < 0.05$ for a no-intercept model). Moreover, visual inspection of scatter plots of the correlational data analyses showed that a significant percentage of students had TBI/KBI scores calculated using the ASMURF annotated response data which were quantitatively similar to the human expert annotated response data.

These results provide evidence that even though the semantic annotation performance of the ASMURF system in its current form needs additional work, the current version of the ASMURF system appears to be reasonably effective at assessing contribution weights similar to those calculated from expert human semantic annotators.

Summary and General Discussion

In this paper we introduced an entirely new methodology for complex reading comprehension assessment which is based upon established findings from the existing scientific text comprehension literature. Specifically, our methodology is based upon the idea that the organization of ideas in student free response data can provide important clues regarding how a student understands a text.

Within the ARCADE framework, students are asked open-ended questions about specific carefully chosen texts. A subsample of the student responses is then semantically annotated using an ASCG. This subsample of student responses is also used to train a natural language understanding system to identify TBI and KBI components of the ASCG in student response data. The natural language understanding system's output is then a sequence of ASCG propositions for each student. Statistical regularities in those proposition sequences are then analyzed using the KDC categorical time-series analysis in order to group students whose patterns of responses to the open-ended questions have similar structures.

It should be emphasized that our natural language understanding system had to deal with many challenges such as the ability to process misspelled words, ungrammatical sentences, and inferences driven by prior knowledge. In order to develop a system which could achieve these objectives, we developed the ASMURF system. Although the ASMURF system demonstrated the ability to semantically annotate novel free response data in a manner similar to human semantic annotators when using a TBI/KBI performance measure, our long-range goal is the development of a reading comprehension assessment system which is capable of complex comprehension assessment. Accordingly, further future research to improve the performance of the ASMURF system is planned since its semantic annotations are generally semantically implausible.

This unsatisfactory performance of the ASMURF system is probably due to two factors. First, the ASMURF system currently does not incorporate state-of-the-art or even standard natural language parsing mechanisms such as a part-of-speech tagger or a spell-checker. The incorporation of such mechanisms is expected to improve the performance of the system. Second, the process of semantically annotating the free response data was relatively tedious resulting in coding errors and thus corrupted training data. This problem could be addressed by improving the user-interface and the semantic annotation performance of the ASMURF system. If the ASMURF system can make better suggestions to the human semantic annotator during the coding process, this would reduce the coding errors.

Nevertheless, it was shown that when used in conjunction with KDC analysis the current version of ASMURF may be viewed as a version of other indirect methods for comprehension assessment which are based upon word co-occurence such as latent semantic analysis (Foltz, Kintsch, & Landauer, 1998). In particular, it was demonstrated that ASMURF appeared to pick up a sufficient number of statistical regularities in order to meaningfully cluster students along the TBI and KBI comprehension dimensions. We find this result very encouraging and expect that by incorporating state-of-the-art natural language machinery into the ARCADE/ASMURF/KDC methodology developed here that even further progress will be made towards the development of a reading comprehension assessment tool intended to assess complex comprehension processes for the purposes of enhancing classroom instruction experiences.

Acknowledgments

This research was supported by the National Science Foundation (NSF) Information Technology Research (ITR) Award Initiative through the Research On Learning and Education (ROLE) Program Award 0113369 within the REC Division.

We also express our appreciation to the teachers with whom we have collaborated in this work and our research team members. We are particularly grateful to Shaunna Macleod (UIC) and Bita Payesteh (UTD) for their contributions to the data analysis component of the research reported here.

The KDC, AUTOCODER, and ASMURF Software developed for this project were funded by this ROLE Program Award and may be downloaded for non-profit academic research purposes from the website: www.utdallas.edu/~golden/arcade.

References

Coté, N., Goldman, S. R., & Saul, E. U. (1998). Students making sense of informational text: Relations between processing and representation. *Discourse Processes, 25*,1-53.

Foltz, P., Kintsch, W., & Landauer, T. (1998). The measurement of textual coherence with latent semantic analysis. *Discourse Processes, 25*, 285-308.

Golden, R. M. (1998). Knowledge digraph contribution analysis of protocol data. *Discourse Processes, 25*, 179-210.

Golden, R. M. (2003). Discrepancy risk model selection test theory for comparing possibly misspecified or nonnested models. *Psychometrika, 68*, 229-249.

Golden, R. M. (2006a). *Annotated Semantic Markov Utterance Random Fields for Information Extraction*. BBS, University Texas at Dallas, Richardson,TX.

Golden, R. M. (2006b). *Knowledge Digraph Contribution Analysis*. BBS, University Texas Dallas, Richardson, TX.

Goldman, S. R., & Wiley, J. (2004). Discourse analysis: Written text. In N. K. Duke & M. Mallette (Eds.), *Literacy research methods* (pp. 62-91). NY: Guilford.

Mandler, J. and Johnson, N. (1977). Remembrance of things parsed: Story structure and recall. *Cognitive Psychology, 9*, 111-151.

Stein, N. L., & Glenn, C. G. (1979). An analysis of story comprehension in elementary school children. In R. O. Freedle (Ed.), *New directions in discourse processing: Vol. 2. Advances in discourse processing* (pp. 53-120). Norword, NJ: Ablex.