Reading in a Foreign Language
Volume 16, No. 2, October 2004
ISSN 1539-0578

Towards enhanced second language reading comprehension assessment: Computerized versus manual scoring of written recall protocols

Peter J. Heinz
Pikes Peak Community College

Abstract

Second language (L2) reading comprehension assessment has long relied upon classical quantitative, product-oriented measurement techniques (i.e., multiple-choice and cloze) in both research and classroom assessment. As Bernhardt (1991) clearly demonstrated, these traditionally employed assessment methods are unable to capture the complex processes that take place between learner and text. The present paper has as its central purpose to enhance and extend the efficiency, consistency, and validity of an alternative measure, the immediate free recall protocol. Unlike the multiple choice or cloze tests, the recall protocol is a truly integrative authentic-task measure, firmly grounded on a constructivist model of reading comprehension. Given an understanding of the model, the literature base on memorial representation, and the formulation of a "weak-rule" scoring system, the present study demonstrates a computerized recall protocol scoring system that has high correlation with traditional manual scoring methods. The results further demonstrate that the computerized procedure provides efficiency in delivery and scoring, enhances consistency, is practical for large-scale assessment, and can lead to improved diagnostic and placement testing. Using this system as part of a multiple-measures approach, valid and reliable quantitative score information is readily available and directly linked to a qualitative database ripe for additional examination to advance L2 reading comprehension research and model development.
Keywords: reading comprehension, assessment, recall protocol, computer based testing

Introduction

Significant attention in the second language (L2) reading comprehension assessment literature revolves around the development of a dynamic, learner-based, constructivist model of reading comprehension that demands new and equally dynamic paradigms of assessment (Alderson, 2000; Berkemeyer, 1989; Bernhardt 1983a, 1985, 1986b, 1990, 1991, 2000; Bernhardt and Berkemeyer, 1988; Bernhardt and DeVille, 1991; Brisbois, 1992; Maarof, 1998). L2 reading comprehension assessment, however, has long relied upon classical quantitative, product-oriented measurement techniques (i.e., multiple-choice and cloze) in both research and classroom assessment. As regards reading comprehension, much has been written about the shortcomings

of traditional assessment such as multiple choice and cloze tests (see for example, Alderson, 2000; Maarof, 1998) so prevalent in education today. These product-oriented assessment methods are traditionally employed to gain insight into the learners' L2 reading comprehension abilities, but as Bernhardt (1991) clearly demonstrated, this sort of assessment is not enough because it is unable to capture the complex processes that take place between learner and text. In order to remedy this problem, a new and welcome trend in reading comprehension assessment emerged that looked "more carefully at the authenticity of the assessment tasks and their alignment with current research, theory, and instructional practices" (Valencia, 1990: 60).

The present paper, firmly grounded in a constructivist view of L2 reading comprehension, embraces and builds upon the considerable body of research and extensive database developed primarily by Bernhardt and her many colleagues. In fact, it is the central purpose of this paper to provide a pathway to ultimately enhance and extend the efficiency, consistency, and validity of the primary measure used in many of the studies cited above, the immediate free recall protocol. This paper clearly demonstrates the beginnings of an efficient, automated recall protocol data collection and scoring system for large-scale studies that, when used as a part of a multiple-measures approach, significantly enhances quantitative and, more importantly, qualitative data collection to further inform research and instruction. The richness of the data may ultimately serve to provide greater insight into that which is currently unexplained or unknown, and thus help extend our understanding of the complexities of L2 reading comprehension development and subsequently further the development of existing and future research models.

It is beyond the scope of this article, however, to explicate detailed aspects of traditional recall protocol development, administration, and analysis using the weighted pausal unit scoring system (Johnson, 1970). A significant body of literature already exists to inform the unfamiliar reader with this measure (Berkemeyer, 1989, Bernhardt, 1991; Brisbois, 1992; Maarof, 1998, among others). Rather, this paper describes the development, testing, and validity of an automated recall assessment procedure designed to mirror the already widely used and accepted manual procedure and to facilitate efficient scoring and data analysis.

The recall protocol procedure in L2 reading assessment

Operating under a learner-based theory of reading comprehension, Bernhardt (1990, 1991) demonstrated the inability of traditional, quantitative reading comprehension assessment to examine the complex and dynamic processes involved in L2 reading comprehension. She found that "if a test is to adequately assess L2 reading ability it must acknowledge the status of the reader's knowledge base," and "a successful assessment mechanism must be integrative in nature" (Bernhardt, 1991: 193).

As a research-based, appropriate and suitable alternative, the recall protocol procedure is seen as a highly valid and effective L2 reading comprehension assessment measure that provides both qualitative and quantitative information (Berkemeyer, 1989; Bernhardt, 1983a, 1985, 1991; Bernhardt and Berkemeyer, 1988; Brisbois, 1992; Lee, 1986). According to Berkemeyer (1989: 131):

> It does not allow students to guess their way through the text…nor does it influence students' understanding of the text. In short, the immediate recall protocol demands that the reader comprehend the text well enough to be able to recall it in a coherent and logical manner…. This procedure allows misunderstandings and gaps in comprehension to surface; a feature that other methods of evaluation cannot offer.

Bernhardt and others (e.g., Berkemeyer 1989, 1991; Brisbois 1992; Lee, 1986, Maarof, 1998) have long championed the use of the immediate, free-response reading recall protocol as a particularly efficacious measure of a learner-based L2 reading comprehension paradigm. That paradigm promotes a multifaceted reading comprehension research and assessment approach combining quantitative and qualitative methods in order to arrive at a more complete picture of the comprehension process. Clearly, the recall protocol procedure can be an important part of that multifaceted approach. As Brisbois (1992: 168) noted:

> …testing methods, such as the recall protocol procedure, need to gain wider acceptance as a measure of reading comprehension. Not only is the recall protocol more sensitive than discrete-point tests, but its sensitivity becomes more pronounced as reading proficiency increases.

Recall as an authentic assessment task

The authenticity of the recall protocol procedure to assess L2 reading comprehension has come under question. Schmidt-Rinehart (1994), for example, reports on research that used the recall protocol procedure for analysis. She states, "although the advantages of this comprehension measure are well documented …one would rarely be asked to perform a similar task in real life" (Schmidt-Rinehart, 1994: 186). In fact, nothing could be further from the truth as can be argued on three levels. First, from a theoretical perspective the recall protocol procedure has been successfully used as a noninvasive technique for understanding how persons have put a message together in both L1 and L2 research (e.g., Bartlett, 1932; Berkemeyer, 1989, 1991; Bernhardt, 1983a, 1983b, 1985, 1986a, 1986b, 1990; Brisbois, 1992; Craik and Lockhart, 1972; and Rumelhart, 1980; Maarof, 1998, among many others). Second, from a practical level the recall task is encountered daily. Who has not experienced being asked by an acquaintance if he or she had read an article in some newspaper or magazine? The acquaintance then proceeds to relate his or her understanding of the article. Upon later inspection of the original piece we often find that items were omitted or embellished upon based on the understanding created by the reader. Children often cannot wait to relate to their parents stories they read or that were read to them in school. The retelling of what someone has read is, in fact, a daily, rather common occurrence--one in which human beings engage naturally and readily in the course of conversation. Third, for L2 reading comprehension the recall protocol procedure has been shown to be an important part of an integrated, multiple measures assessment approach (Berkemeyer, 1989; Bernhardt, 1983a, 1985, 1990; Bernhardt and Berkemeyer, 1988; Bernhardt and Deville, 1991; Brisbois, 1992; Lee, 1986; Maarof, 1998). Unlike discrete point measures, it is an integrative task where students write down everything they remember about what they read and thus provide a rich sample of their individual construction of the text.

Thus, the present study posits the recall protocol procedure as an authentic-task and clear alternative to use in conjunction with more traditional instruments such as the multiple choice

and cloze examinations. It is a procedure that has several distinct advantages (Bernhardt, 1983a: 31-32):

1) The recall procedure shows where a lack of grammatical skill interferes with the student/text communication.
2) The recall procedure does not influence the readers' understanding of the text.
3) The procedure stresses the importance of understanding. Students cannot simply guess at answers; they must attempt to form an understanding of the text.

Construct validity: An evolving model

The recall protocol procedure has construct validity firmly grounded in a reader-based, constructivist model of L2 reading comprehension that has continued to evolve. Bernhardt (1985, 1986a, 1990) describes an L2 reading comprehension model that is based in part on a psycholinguistic model developed by Coady (1979) and to a great extent on analyses of recall protocol data. This interactive model attempted to capture the complex comprehension processes that take place during the reader/text interaction. Bernhardt (1990: 25) describes the various components of the model as being either text based, or extra-text based. The text-based components include:

1. Word recognition (the attachment of a semantic value).
2. Phonemic/graphemic decoding (recognition of words based on sound or visual match).
3. Syntactic feature recognition (the relationship between words).

The extra-text based components of the model consist of:

1. Intratextual perception (the reconciliation of each part of the text with that which precedes and succeeds).
2. Prior knowledge (whether or not the discourse is sensible according to the reader's knowledge of the world).
3. Metacognition (the extent to which the reader thinks about what he or she is reading).

Using the model, Bernhardt (1985, 1990) was able to reconstruct the processes and mental models developed by the individual reader during comprehension. She found that "discovering the mental model can be done using the recall protocol procedure and thereby working from the students' reconstructions in order to make students actively attend to their process of model building" (Bernhardt, 1990: 41).

Based on an extensive synthesis of recall protocol data, Bernhardt further extended her model to account for data that indicated that "problems or inaccuracies in L2 text processing may be differentially linked to L2 literacy development" (Bernhardt, 1991: 168). The theoretical model she posited attempted to explain the development of L2 reading proficiency based on the following assumptions (Bernhardt, 1991: 169):

1. Text processing abilities develop over time.
2. Readers demonstrate the use of different facets of the features of the model over time.
3. Errors in understanding can reveal development in literacy.

4. Assumes commonality in L2 text processing between and among literate learners and languages.
5. No L2 reader would ever be 100% proficient with a 0% error rate nor would an L1 reader be 0% proficient with a 100% error rate.

Thus, based on recall protocol evidence, Bernhardt delineated a construct whereby exhibited learner errors in the use of the L2 reading comprehension factors and their interrelationship vary as proficiency increases. This model also recognizes that metacognition occurs at all levels of proficiency and must still be included in L2 comprehension theory, but acknowledges it as a characteristic that varies and is highly dependent on the individual.

Reading comprehension research in the 1990's further re-examined the relationship between L1 reading ability and the development of L2 reading ability. Alderson (2000) summarizes research that investigates or tries to "resolve the question of whether L2 reading is a language problem or a reading problem" (Alderson, 2000: 38). A literature survey conducted by Bernhardt and Kamil (1995) subscribed a strong 20% of the variability in L2 comprehension test scores to first language (L1) literacy. L2 linguistic knowledge accounted for more than 30% of the variability, while 50% of the variability in their research of L2 reading comprehension studies remained unexplained. Citing additional studies that demonstrated the importance of L2 knowledge, Alderson (2000) posited that while this may be so, L2 learners must also cross a "'linguistic threshold' that varies by the difficulty of the comprehension task before L1 reading ability transfers to the L2 reading task" (Alderson, 2000: 39).

Building on the understanding that L1 reading abilities both promote and hinder L2 reading comprehension, Bernhardt (2000) recognized that the relationships between factors in her 1991 model are also affected by the relationship of the "linguistic overlap" between the two languages (Bernhardt, 2000: 803). Noting that the features of her 1991 model provide a picture of L2 reading ability at a moment in time, Bernhardt determined that an articulation of the development of comprehension ***over*** time was required. Her revised alternative conceptualization (Figure 3) stems from the evolution of 1990's reading comprehension research and acknowledges the time required to learn and develop L2 reading ability as related to actual comprehension. In this model, as comprehension ability grows (Score 1, 2, or 3) general reading ability explains 20% of the result, L2 word and syntax knowledge explain approximately 30%, and almost 50% of the variability remains unexplained. While appreciating the important role that L1 reading ability plays in developing L2 ability, the model also recognizes that comprehension scores consist of different elements that the learner brings to bear on any given task. In addition, having no zero point on the y-axis, the model recognizes that learners will have some comprehension ability especially in cognate-rich languages. Finally, and most importantly, the model "promotes the consideration of unexplained variance in individual performance and after considerable time in instruction" (Bernhardt, 2000: 804).

It is precisely the need to further our understanding of the multi-faceted and complex processes involved in the development of L2 reading comprehension ability and the realization that much is yet unexplained or unexplored that necessitates the drive to enhance the effectiveness and efficiency of the recall protocol procedure as a part of a multiple measures approach. Alderson (2000) correctly notes that "how researchers operationalize their constructs crucially determines the results they will gather and thus the conclusions they can draw and the theories they develop." (Alderson, 2000: 356).

Disadvantages of the recall protocol

As with any assessment measure, limitations to the recall protocol exist and must be acknowledged. Alderson (2000) and Brisbois (1992) point to the major disadvantage of the recall protocol procedure, namely that traditional scoring is very time consuming. While Bernhardt (1991) notes that scoring can take up to 10 minutes per recall, Brisbois (1992), based on her research, found that "in order for this procedure to attain wider use, however, the scoring process needs to be rendered less time consuming. Research into the automatization of this process would open the way to increased use of this testing method" (Brisbois, 1992: 169). Thus, due to the enormous scoring time requirements and subsequent impact on rater consistency over time, traditional studies using the recall protocol were necessarily limited to small groups. Clearly, a streamlined, automated procedure would greatly enhance the utility of the recall protocol procedure.

Administration of the recall protocol task can also present problems and affect the resulting data. Alderson (2000) and Lee (1986) denote objections that the immediate recall protocol may be more of a test of memory rather than a measure of comprehension. These objections are minimized since in this procedure, the recall typically occurs immediately after reading. Riley and Lee (1996) found that the performance on the recall task varied by the instructions given to the subjects. The recalls provided by subjects told to summarize the main ideas of the text were found to contain significantly more main idea units than the recalls of subjects simply told to write down what they could remember. From their research, it is clear that the task may have an effect on what is recalled and must be clearly defined.

Another frequently listed disadvantage of the recall protocol is that of production difficulties in the L2 (Maarof, 1998). If subjects were required to produce their written recall in the L2 the results may be confounded by their production ability. To avoid this limitation, most studies have required subjects to recall in their native language so as not to interfere with their ability to demonstrate comprehension (e.g., Bernhardt, 1983a; Bernhardt and Berkemeyer, 1988; Maarof, 1998).

Computerizing the recall protocol procedure: A theoretical framework

As the goal of the present research was to produce a computer-based, valid, and highly efficient authentic-task assessment procedure, the theoretical underpinning for automating the recall protocol is founded on the extensive body of literature regarding memorial representation (e.g., Kintsch, 1974, 1988; Kintsch and van Dijk, 1978; Weaver and Kintsch, 1991). Kintsch (1974) developed an expository text analysis system based on "the notion of propositions as the basic unit of meaning" (Weaver and Kintsch, 1991: 233). Kintsch (1974: 62) formulated a general theory of episodic memory storage, organization, and retrieval centered around three assumptions:

1. Information is stored as sets of phonemic, semantic, and imagery elements.
2. The basic operation is one of pattern completion.
3. The contents of short-term memory bias encoding processes both in perception and memory.

He postulated a two-stage generation-recognition model that includes an encoding specificity principle. The encoding specificity principle, previously advanced by Tulving and Thomson (1973: 16), states that "a retrieval cue can provide access to information available about an event in the memory store if and only if it has been stored as part of the specific memory trace of the event." In Kintsch's model, the storage, retrieval, and organization of semantic elements depend greatly on pattern matching and completion. The elements of stimulus propositions match up with representations stored in memory and result in the retrieval of stored information.

Much of Kintsch's work on retrieval from memory stems from previous list-learning research. In the case of list retrieval, subjects retrieve from memory relatively unrelated words based on their individual ability to develop a retrieval scheme while learning the words. Kintsch (1974: 259) points out that textual retrieval is a process whereby "text bases are structured lists of propositions, with rich interconnections …the retrieval cue, usually the title of the story or the like, shares a sufficient number of elements with the episodic memory representation of the text."

Kintsch also notes research that proposes a multilevel perspective of memory for text. Research indicates, for example, that text memorized verbatim is subject to more rapid forgetting than is text stored for meaning. Items memorized verbatim appear to be stored in short-term memory and thus lack linkage with existing memory nodes. Kintsch (1974) demonstrates that the verbatim memory level consists of perceptual-linguistic information that is subject to rapid forgetting. The deeper-, or propositional-level representations in memory, having been linked to pre-existing propositions, remain available for retrieval far longer than the surface level structures. Thus, a hierarchical structure of memory is postulated, which Kintsch and van Dijk (1978) later divide into macro- and micropropositions.

Weaver and Kintsch (1991: 233) define macropropositions as "propositions that contain only top-level 'gist' information." They found that micropropositions are directly derived from the text and "refer to the smallest definable text units, and completely represent the microstructure of the text" (Weaver and Kintsch, 1991: 233). Macropropositions, in contrast, do not contain detailed text information, but only surface-level detail. A hierarchical textbase is constructed by linking the propositions together, with the important macrostructure information stored at the top. It is this top-level information that is recalled in more detail in long-term experiments (Weaver and Kintsch, 1991). Detailed, microstructure information is stored toward the bottom of the hierarchy, where it is more easily forgotten.

Guindon and Kintsch (1984) further demonstrate the importance of macropropositions in memory. Macropropositions have a "priming effect" and form strong memory units. Guindon and Kintsch found evidence that "word pairs from the macrostructure prime each other more strongly than word pairs from regular sentences, and within-sentence priming is greater than between-sentence priming" (Guindon and Kintsch, 1984: 516). So strong is the effect of macropropositions, that readers form them whether "they are stated explicitly in the text or not, and whether subjects are asked to do so or not" (Guindon and Kintsch, 1984: 517).

Continued research by Kintsch (1988) led to the formulation of the construction-integration model that "describes how texts are represented in memory in the process of understanding and how they are integrated into the comprehender's knowledge base" (Kintsch, Welsch, Schmalhofer, and Zimny, 1990: 136). The model postulates weak (dumb) rules that simulate comprehension as a production system.

Kintsch et al. (1990) found that "dumb" rules that do not always work are superior to so-called "smart" rules that attempt always to arrive at the correct interpretation of a text. According to Kintsch et al. (1990: 136):

> Weak rules…do not generate acceptable representations of the text. Irrelevant or contradictory items that have been generated by weak rules, however, can be eliminated, if we consider not just the items generated by the rules, but also the pattern of interrelationships among them.

They further found that irrelevant items in a text will only be related to one or to a few items, and that contradictory items will be negatively connected to other items. As Kintsch et al. (1990: 136) note:

> Relevant items, on the other hand, will tend to be strongly interrelated--be it because they are derived from the same phrase in the text, or because they are close together in the textbase, or because they are related semantically or experientially in the comprehender's knowledge base.

In this model, test sentences or words are compared to the entire text (knowledge) base. Using complex activation vector analysis techniques, Kintsch et al. (1990) found that test sentences that were highly related to original text bases had strong activation vectors, but if there were no connection at all, no activation vector existed. In their words, "the more similar it is to the original, the more connections there will be, and the more highly activated the test sentence will become" (Kintsch et al., 1990: 137).

Three components underpin the Kintsch (1988) construction-integration model: a) recognition based on list-learning research, b) the hierarchical representation of text, and c) the processing mechanism of the construction-integration model. The present research utilized elements of the Kintsch construction-integration model to develop an efficient, objective, and reliable L2 recall protocol scoring procedure.

Computerized recall protocol text analysis

Traditional computer text analysis procedures have proven cumbersome and tentative at best, with none being able to analyze texts accurately and completely. Artificial intelligence (AI) techniques only now coming into existence may eventually hold the key to true natural language processing. Unlike traditional text processing requirements, however, whereby computers process unknown texts, the present research was intended to analyze students' recall of a known textbase. Intuitively, this procedure requires a matching of the propositions recalled with the propositions in the original textbase. This is easier said than done as languages are rich with variation, and any proposition can be expressed in a number of ways. For efficient, simple processing to occur, algorithms must be found to match recall text with the original textbase as closely as possible and give credence to the strength of the relationship.

Perhaps the development of several computer-coded "weak" rules that tap into the fact that a person's recall directly reflects and is related to what has been comprehended will allow for an accurate picture of his or her level of comprehension. In line with Kintsch et al. (1990), these

"dumb" rules may not always reflect accurately what was comprehended but, by examining the patterns of interrelationship, the computer may be able to eliminate, or at least limit the inconsistencies. It may be possible that the use of computerized "weak" rules may result in an innovative algorithm to produce automated recall protocol scores that closely correlate with recalls scored by human raters.

The state of available technology

Rapid technological advancements make possible expeditious analyses of large amounts of numerical and textual data. Off-the-shelf technology was used in the present study to develop a practical and efficient scoring procedure by combining the power of the computer and the recall protocol. Integrated computer systems using high-speed processors and mass storage devices as described by Baker (1984) and Nitko and Hsu (1984) are now readily accessible. Computer concordancing software has revolutionized literary research (Pfaffenberger, 1988; Tang, 1985). What was once a lifetime task can now be accomplished in minutes. As a result, concordancing moved beyond the field of literary criticism and rapidly became a valuable research tool in fields such as political science (Pfaffenberger, 1988) and could also prove useful in qualitative recall protocol analysis.

In addition, state-of-the-art relational databases, spreadsheets, and interactive software authoring systems give the researcher the resources necessary to create complex programs and simulations. Relational databases cross-link bits of related data stored in separate data files for processing as a combined entity. Spreadsheets allow users to make alterations to quantitative data, perform complex calculations, and view and evaluate the results instantaneously. Interactive software authoring systems give even novice users the power to create complex programs and to manipulate electronic information between different data processing software packages.

Clearly, the advances in software technology make possible sophisticated textual analyses. Programs can be developed to carry out weighted pausal unit analysis as described by Johnson (1970), or explore the feasibility of the Kintsch text analysis system (Kintsch, 1974; Weaver and Kintsch, 1991). An integrated computer assessment system, based on the recall protocol, has the potential to provide researchers, teachers, and policy makers with an efficient, valuable, and powerful assessment tool. Innovative recall data processing algorithms and procedures, based on text analysis research (e.g., Johnson, 1970; Kintsch, 1974), need to be developed, integrated, demonstrated, and evaluated. In doing so, the present research will help answer the call to "embark on a large-scale effort directed toward the development of practical methods of text analysis that allow nonexperts to do with relatively little effort what only experts can do today with considerable effort" (Weaver and Kintsch, 1991: 242).

Development of computer weak-rule scoring algorithms

The researcher examined the use of a weighted means of calculating recall protocol scores as outlined in Bernhardt (1991) and Maarof (1998). In addition, various text processing and analysis perspectives (e.g., Kintsch, 1974; Kintsch and van Dijk, 1978) were examined, used, combined, and/or modified.

The first step was to determine if an interrelationship could be detected between the text of a student's recall and the propositions of the textbase as presented by Guindon and Kintsch (1984),

Kintsch (1988), and Kintsch et al. (1990). Several "findings" from Guindon and Kintsch (1984) are particularly relevant. These findings state that:

1. "Words belonging to the same sentential unit should prime each other more than words that do not belong to the same unit" (p. 509).
2. "Words from the same sentence unit, in turn, prime each other more strongly than words from different sentences" (p. 514).
3. "Within-sentence priming is greater than between-sentence priming" (p. 516).
4. "Words that belong to the macrostructure of a text are recognized faster than microstructure words, irrespective of priming effects" (p. 512).
5. "One content word from a macroproposition provides ready access to another word from the same macroproposition" (p. 514).

After careful examination of recalls from a previous study (Allen, Berry, Bernhardt, and Demel, 1988), the interrelationships that Kintsch and his colleagues described were indeed found to be evident--where recall of a passage was evident, closely related propositions from the original text appeared in proximity to each other in the subject recalls. They may not always have been worded the same way, but the relationship was unmistakable. For example, consider the following sentence recalled in English by a high school German student from an authentic L2 text on a visit by then President Reagan to Moscow: "The man from Spartanburg South Carolina said, 'If I'd wanted to see Russians, I'd have bought a Russian TV.'" Compare this to the original sentence, "'When I want to see a Russian, I will buy myself a Russian TV', said a citizen from Spartanburg, in the State of North Carolina." Examination of all student recalls revealed similar, consistent linkage patterns, the variety and quality of which greatly depended on the depth and amount of material recalled.

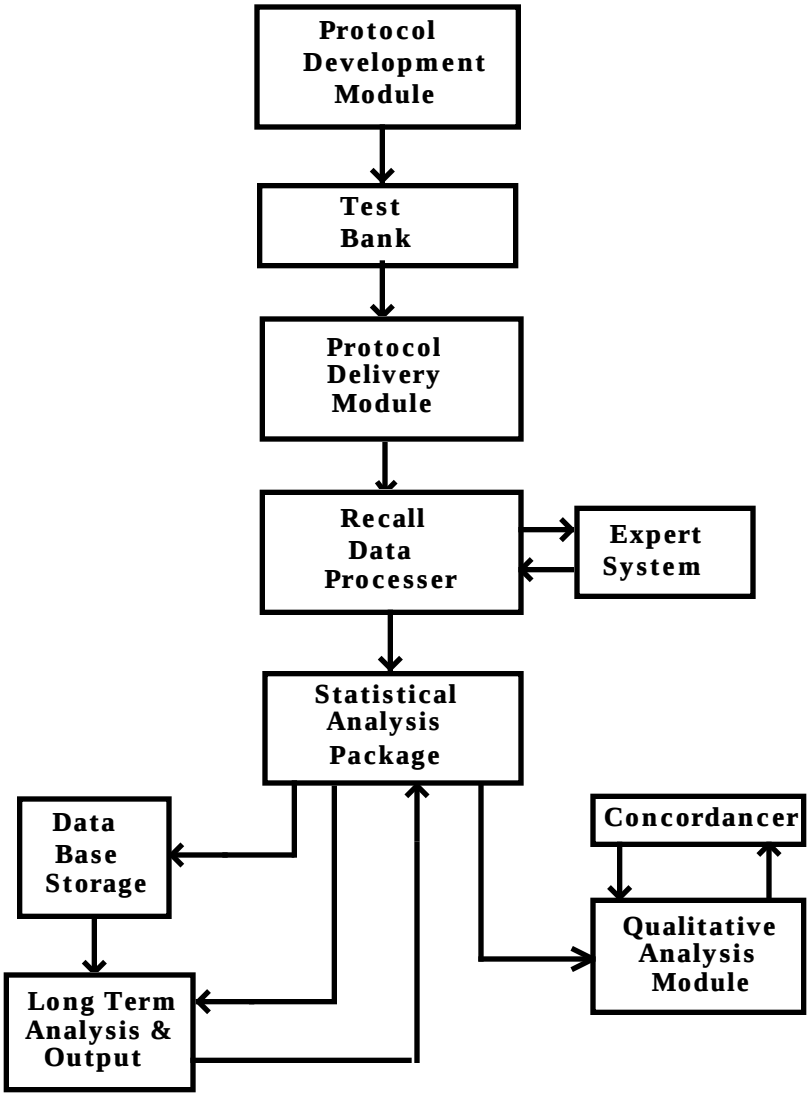
Weak rule scoring. Based on an analysis of these relationships, the researcher developed two weak-rule scoring systems. To summarize these rules, the first, designated "link-only" scoring, was designed to examine the recall word-by-word from the beginning and match these words to propositions in the original text. Link-only scoring captured and credited items in the recalled text based on their relationship to the original text. Under these rules, items that were collocated in the recall were ranked according to the following priority: a) the proposition represented by the current word corresponds +/- 1 to the proposition represented by the following word; b) the proposition represented by the current word is in the same sentence as a proposition represented by the following word; and c) the proposition represented by the current word is the same proposition represented by the following word. The computer link-only procedure, using these priorities, attempts to build as many unbroken linkages possible. The propositions contained within the linkages with the strongest relationship (i.e., proximity and propositional valuation) are then credited as having been remembered.

A second weak-rule scoring system, "word-only" scoring, was developed to capture certain words in the comprehender's textbase that could not be readily linked to other surrounding propositions, though their presence indicates that the reader indeed comprehended something from the text (i.e., the word did not simply appear by chance). For example, a recalled reference to the "Rose Bowl" mentioned in an original text about Ronald Reagan's visit to Moscow (Bernhardt, 1991) would be credited as having been recalled, even if it occurred in isolation with no linkages possible.

Computerized recall protocol assessment system

In order to test these rules and answer the research question, a computerized recall protocol assessment software package (Figure 1) was developed by the researcher using state-of-the-art, readily available, off-the-shelf technology that can be found on almost any desktop personal computer today (e.g., Microsoft Corporation's Word, Excel, and Visual Basic software). The first step was to develop a system to deliver the original texts and capture the student recalls. This was done using Microsoft Visual Basic on a Windows-based platform. In this system, students would read the original L2 texts for as long as they desired. When ready, they clicked on a button, the text permanently disappeared and they would then immediately input their recall into the computer. Once complete, a mouse click would then bring up the next text and the process would repeat until all three texts were delivered and the resulting recalls were captured into individual data files.

Figure 1: Automated recall protocol assessment system. (Heinz, 1993: 21)

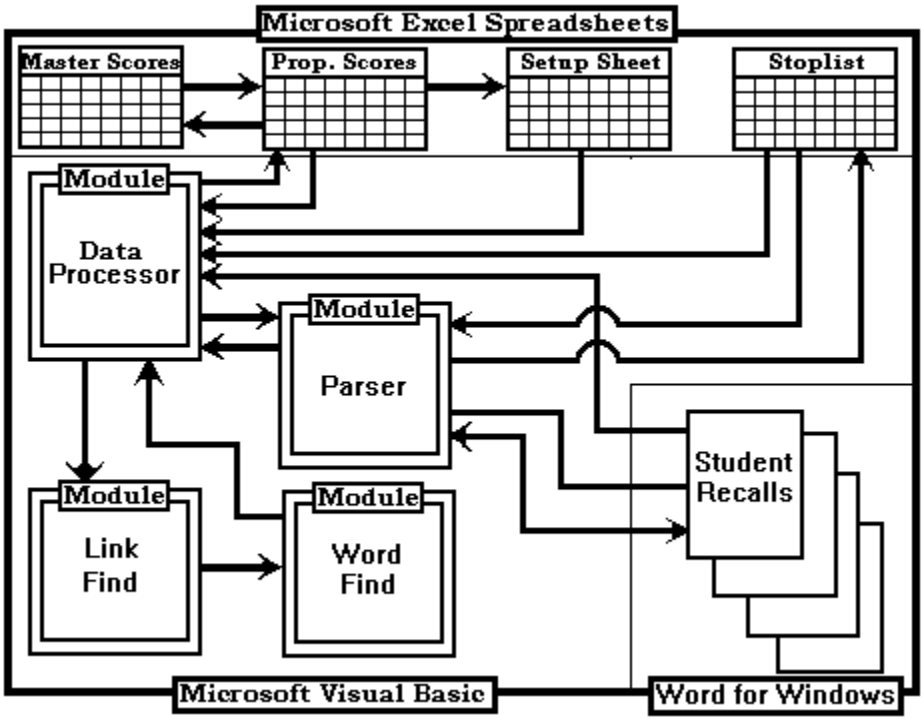


The second step was to develop the computer scoring system. In this system, Microsoft Excel spreadsheets were used to store data such as the generated student scores, the recall rubrics containing the propositional scores (weights), and additional data required to facilitate processing. The recall rubrics contained in the system mirrored those developed by the human raters and used in the manual scoring procedure.

Text processing. At the heart of the system is the recall data processor and "expert system" that contained the recall scoring program itself (Figure 2). The program was designed to assess student-generated recalls using the weak-rule scoring systems previously described. These rules were carefully converted into software algorithms and repeatedly tested for reliability using sample recalls from a previous study.

Prior to processing, student-generated recalls must be spell checked to ensure that the computer can recognize the words. Due to the limited availability of authoring systems and the level of computer expertise of the researcher at the time, spell checking was done one text at a time (with currently available software, this process can now be easily automated and integrated). The researcher was careful to simply correct misspellings and not change words. In addition, a parsing routine, based on algorithms described by Smith (1991), was designed to ensure that variations of words (i.e., past tense, plurals, etc.) would be recognized by the computer as being equivalent to their root words (i.e., "played" equals "play"; "carries" equals "carry").

Figure 2: Recall scoring program. (Heinz, 1993: 78)



Once spell-checked and parsed, the computer program was designed to process the entire set of student recalls one at a time through the Recall Data Processor and submit them to the "Link Find" module. At this point, the computer examined each student-generated recall, word-for-word, and attempted to credit propositional matches between the recall and the original text in accordance with the algorithms outlined above. As a part of the process, every student-generated word was compared to a list of logical and acceptable synonyms (as determined *a priori* by a panel of trained raters). This procedure mirrors the process that human raters routinely make as they score the recalls and attempt to match propositions. Once all possible linkages were formed, evaluated for their strength of relationship and credited, the recall was passed to the "Word Find" routine that then attempted to make propositional matches based on the remaining unmatched unique words. Finally, the computer recorded the matched propositions and proceeded to score the next subject's recall. This process continued until all recalls in the study were scored.

Research question

Given an understanding of the recall protocol, the literature base on memorial representation, and the formulation of a "weak-rule" scoring system, the major research question was to determine if there is a significant correlation between recall protocol scores that are manually assessed using a pausal unit analysis procedure (Bernhardt, 1991) and computer-generated recall scores using the algorithms described. Is it possible, in other words, to use commonly available technology to mirror the processes used by human raters to come to an acceptable and equally valid quantitative score that might help guide and facilitate qualitative examination of recall data on a large number of subjects?

Research design

The present study used a series of Pearson product-moment correlation coefficients between scores manually generated by trained raters using the pausal unit analysis system and the computer-generated recall scores. In addition, the researcher conducted a qualitative analysis on the scoring procedure by analyzing how the computer-scored propositions compared with those scored by the human raters for each student-generated text. The researcher also collected scoring time intervals and analyzed them in order to measure efficiency.

Independent variables

The two independent variables consisted of the various methods of recall protocol assessment, specifically: a) the manually scored pausal unit analysis procedure (Bernhardt, 1991); and b) the automated computer analysis procedure developed in the present study.

The recall texts used in the present study were administered by the researcher-developed computer program. Each subject received three authentic German passages (see Appendix) of approximately 200 words in length delivered in random order. Subjects were allowed to read the L2 text and then to write in their L1 (English) all that they could remember about that text. The subjects typed their recalls directly into the computer. As each subject completed his or her recall, the computer automatically stored the recall for further processing.

Dependent variable

The dependent variable, reading comprehension, was measured using the scores obtained on each of the written recall protocols, and the combined mean score obtained on all three. The computer instructed the subjects to read each passage as many times as desired and then to write everything they could remember about the text, in English.

Reading passages

For the research, the two authentic German articles and a personal letter used in a previous study (Allen et al., 1988), each approximately 200 words in length, were selected from newspapers and nationally read magazines. The articles were then divided into pausal-breath unit propositions and leveled based on hierarchical importance (using a scale of 1 to 4) according to the procedures outlined by Bernhardt (1991) by native-speaking raters (all raters were German

professors at the United States Air Force Academy where the study took place). Manual scoring templates (rubrics) were also developed as described in Bernhardt (1991). From these rubrics, identical automated scoring masters (spreadsheets) were prepared and checked by the same raters. Finally, for the words contained in every proposition, "logical synonyms" (those that indeed fit within the context of the story as judged by the raters) were gathered using the thesaurus function of Microsoft Word and agreed upon. Other appropriate synonyms, not in the thesaurus but deemed acceptable by the raters were also added to the scoring masters (i.e., "Soviet leader" as a reference to Soviet President Gorbachev) as deemed appropriate by the raters.

Subjects

Two hundred and forty students studying German at the United States Air Force Academy in Colorado Springs, Colorado were preliminarily selected at random from the first-, second-, and third-year course levels (henceforth noted as levels 1, 2, or 3 of ability). The university's selection process (students must graduate high school from within the top 10% academically and be U.S. citizens) ensured that the subjects were all highly literate L1 English speakers. In addition, they routinely used computers and were familiar and comfortable with using them. All students were required to take a foreign language placement test upon admission to the university. Their initial course-level placement, however, was not only based on the placement test scores but also previous (high school and personal) experience in the L2.

Procedures

Subjects received three practice, paper and pencil recall protocol assessments during their normal class periods before data collection took place in order to familiarize them with the procedure. Additionally, all subjects received a familiarization session on the computer. The actual assessment was conducted during the 15th class meeting after the beginning of the fall term in normal class periods. The assessment measure was administered to all 240 subjects, but 100 were randomly selected (in proportion to the number of subjects at each level of German ability) for inclusion in the actual study. This allowed for ample additional data sets should errors or computer glitches occur in the processing or capture of data while students were on-line. Subjects were not told whether or not they were selected to be included in the final sample. Subjects worked independently on individual computers located in a central testing facility and the three original L2 texts were presented to the students on the computer screen in a random order. Subjects were given as much time as they needed to read the text. As per the program design, once they indicated that they were ready to write their recalls, the text disappeared and they were prompted to enter their recall into the computer in their native English. This procedure was repeated until all three texts were administered.

Data analysis

All student recalls were spell checked and then submitted to the parsing routine developed by the researcher as previously outlined. Again it must be emphasized that with currently available software this process is now easily automated so that students conceivably could accomplish the spell check as they complete their recall.

Once parsing and spell checking were completed, the recalls were submitted to the scoring program and automatically scored. Concurrently, the recalls were manually scored in a traditional manner (as specified in Bernhardt, 1991) by the previously trained raters working independently. Each rater scored the entire data set for only one of the three texts and worked during one, day-long session to score the entire set of 100 recalls. The researcher also scored the first 20 randomly selected recalls to establish interrater reliability, and randomly scored 10 of the remaining protocols of each text. Interrater reliability averaged 0.98 across all recalls and raters. The automated and manually scored recalls were submitted to correlational analysis (Pearson product-moment), and the automated recalls were further submitted to item and qualitative analyses.

Table 1: Recall protocol analysis summary of correlations (Manual versus Computer Analysis)

Text	Correlation
Batman Article	0.88
Bernhardt Letter	0.89
Travel Article	0.87
Average "T" Scores	0.94

Results

The results of the correlational analysis for the three articles and an overall, combined score condition is depicted in Table 1. Consistent with previous research (Bernhardt, 1991), recall protocol scores are ultimately combined in order to compensate for varied background/topic knowledge. In the present analysis, scores have been standardized through the T-scoring method. This method was used to ease comparison since the three recall protocols each had different maximum scores. The highly positive correlations for the computer-scored versus manually-scored analyses of the three recall protocol texts, and the combined average "T"-scores obtained from the data were found to be statistically significant ($p<.001$, $df = 98$).

Table 2: Summary of recall protocol computer analysis-combined scores

	Manual	Computer
Pearson product-moment "r":	N/A	N/A
Standard Deviation:	9.41	8.93
Mean Score:	50.00	50.00
Minimum score:	35.56	34.97
Maximum Score:	75.65	69.83
Mean Number of Propositions Recalled:	13.23	15.50

Note: Raw scores converted to T-scores. Sample size = 100. *$p < .001$, $df = 98$

For the combined analysis (Table 2), the correlation between the two scoring procedures was 0.94, meaning that the computer procedure accounted for approximately 88.3% of the variance. In addition, Table 3 shows that overall, both the manual and automated scoring procedures

display clear and similar effects by both the mean number of propositions recalled and level of instruction. The computer order of recall tracks with previous research (Bernhardt, 1991) with level 4 propositions being recalled more often than level 3, followed by level 1, and finally by level 2. In contrast, the manual results in the combined condition differed slightly from previous research in that, overall, the propositions most frequently recalled were level 4, followed by level 1, followed by level 3, followed by level 2.

Table 3: Table of selected means-computer analysis versus manual scoring-combined scores

By Proposition Level

		Mean Number of Propositions Recalled	
Proposition Level	**Number of Subjects**	**Manual**	**Computer**
1	100	4.12	3.88
2	100	2.71	3.46
3	100	4.04	4.69
4	100	4.54	5.22

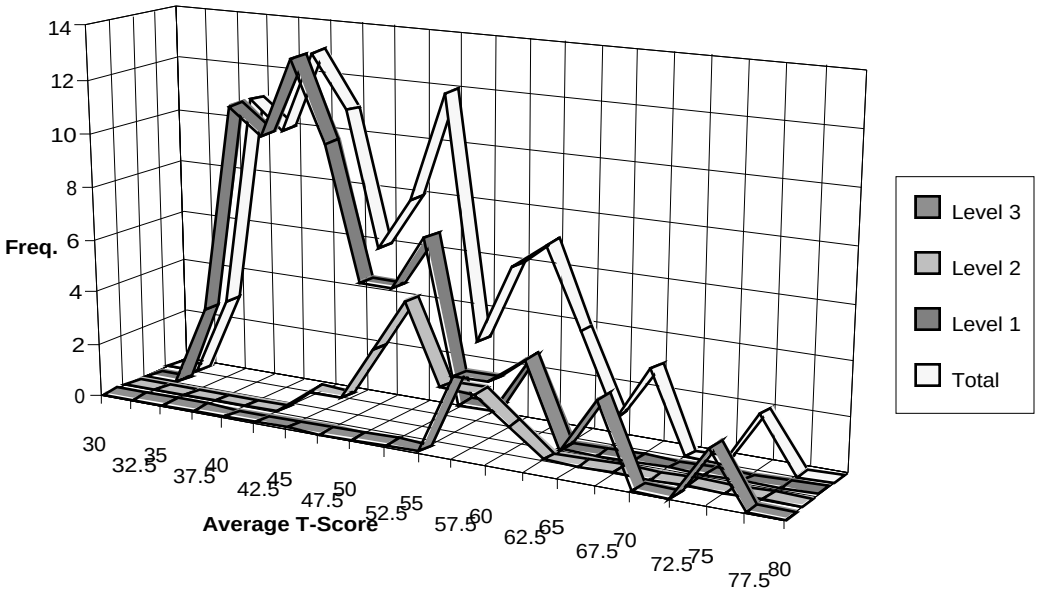
By Level of Instruction

		Mean Score (T)	
Level	**Number of Subjects**	**Manual**	**Computer**
1	69	45.70	46.10
2	16	56.00	55.90
3	15	65.90	63.80

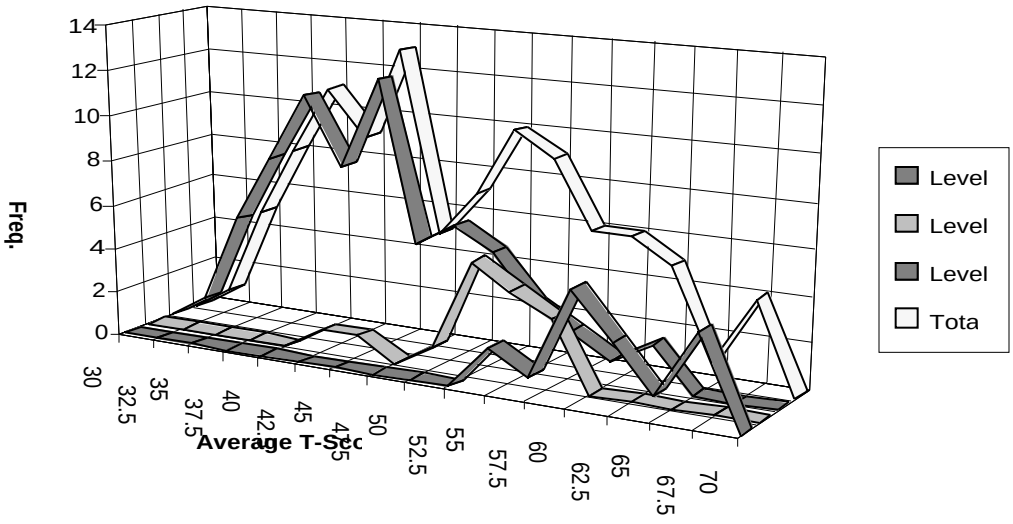
As regards level of instruction, both the computer and manually scored procedures also show that the advanced (level 3) subjects outscored both the level 2 and level 1 subjects. Figure 3 graphically depicts the results by level of instruction for both the combined manual and automated scoring procedures. In the figure, the similarity of the results between the two procedures becomes evident. Thus, both scoring systems provide valid and quantifiable data that generate similar information on the ability level of a subject.

Figure 3: Distribution of Average Total T-Scores--Combined Manual and Computer Analyses (Heinz, 1993: 148).

Manual Analysis



Computer Analysis



Note: Levels 1, 2, 3 correspond to the 1st-year, 2nd-year, and 3rd-year German course levels (proficiency)of the subjects.

One advantage of using the automated system to generate scores clearly becomes evident when scoring times are examined. As Table 4 indicates, manual scoring required an average of 2.8 minutes per recall. In fact, the raters completed their 100 assigned recalls in an average of four and one-half hours of dedicated, on-task time. These times do not include rest breaks, lunch, or

consultation time (with the researcher), which added almost another three hours for a total time of approximately eight hours required to complete the assigned task.

Table 4: Recall scoring time comparisons (N=100)

	Manual Scoring		**Computer Scoring**	
	Avg./Recall (min.)	Total (hh:mm)	Avg./Recall (min.)	Total (hh:mm)
Bernhardt Letter	2.41	4:00	0.23	0:28
Batman Article	3.02	5:02	0.38	0:38
Travel Article	2.42	4:02	0.34	0:33
Average Time	2.62	4:21	0.32	0:33

Note: Manual rater total scoring time represents time actually spent scoring and does not include breaks, lunch, or any time spent between recalls. Computer total scoring time represents processing time spent on actual recall computation, and does not include pre-computation data preparation time (i.e., spell check, etc.).

It was also noted that the human raters, though working diligently at the task, required far more frequent breaks as the scoring session continued. The raters were asked to remain at the task until completely finished. In order to help control for fatigue, raters were allowed to spend as much time as they needed to do all 100 recalls. It was felt that although fatigue would be a minimal factor, completing the task the same day was better than allowing raters to take the protocols home where they could forget how they rated previously, resulting in additional inconsistencies in the scoring process.

What is clearly evident is the overwhelming time saving superiority of the computer scoring method. Processed on an early 1990's home computer operating at a clock speed of only 33MHz (newer machines now operate in the 2 to 3 GHz range), each recall was scored in an average time of 32 seconds, or only 33 minutes for all 100 recalls. Additional time was required for spell checking and parsing, as previously noted, but the development of an automated facility now is definitely feasible. Once incorporated into the overall protocol analysis system, an automated parser would streamline the entire process with a minimal increase in processing time, and spell-checking would be accomplished on-line with the subject.

The bottom line, however, is the undeniable fact that the automated system enjoys a great advantage in processing time over traditional manual scoring, especially when the costs associated with time and labor are added into the equation. The system described here for the first time makes large-scale L2 reading comprehension studies using the recall protocol procedure a feasible, time-efficient and cost-effective reality.

Additional analyses

Using a standard word processor, the student generated recalls were individually examined. This qualitative analysis revealed that "errors" by the computer scoring procedure and those made by the human raters could be examined, explained, and accounted for. Computer errors were mainly caused by errors in the preparation of the master recall scoring template (incomplete synonym lists, incorrect data entry, or inappropriate propositions being credited by the weak-rule

scoring systems programmed into the procedure). Overall, the computer scoring rules overestimated the propositions recalled, but apparently did so in a ***consistent*** manner across all recalls. Human errors were found to result from instances where a rater may have misjudged or missed a valid scoring opportunity, or may have been caused by rater fatigue, but in general, occurred ***inconsistently***.

The data also indicate that both the manual and computer recall scoring procedures used in the present study tapped similar text comprehension indicators that became evident in the qualitative examinations. Both procedures showed similar recalled propositional pattern groupings across the range of subjects. It was extremely rare, for example, for a person to have comprehended a top level (level 4 or 3) proposition and for that proposition not to subsequently have a strong relationship with several other propositions in the text.

Further qualitative analysis also revealed that in comparison to the results of the manually scored recalls, the automated recall protocol procedure developed for the present study indeed reflects the complexity of L2 reading comprehension. Although by no means perfect, the theory-based weak-rule scoring system was able to capture what was comprehended in an organized and coherent manner that is analyzable quantitatively and subsequently highlights areas for further qualitative analysis. The qualitative analysis is further enhanced by the fact that the data are aggregated by the computer and are readily available for additional processing and analysis using a variety of software packages (e.g., concordancing), something not easily accomplished with hand written recalls. Thus, the information provided by the automated scoring process not only serves to provide quantifiable recall scoring, but is also a valid indicator of what is comprehended, how it is comprehended, and by whom it is comprehended.

In the qualitative analysis, the researcher also found evidence to support Cummins' (1979) threshold hypothesis, as was confirmatory evidence of Bernhardt's (1991) theoretical distribution of reading factors. The L2 learners in the present study, a highly select group attending the United States Air Force Academy, typically rank in the top 10% of all high school graduates nationally. Thus, they should be highly literate in their L1 English. Although they clearly possess good L1 reading comprehension skills, it is evident from analysis of the recalls that these skills did not transfer to the L2, at least not for the beginning students. The advanced-level students, however, appear to make use of higher-level processes because their recalls were indeed richer and showed evidence of transfer of L1 reading comprehension skills.

Advantages of the automated recall protocol

The automated reading recall protocol procedure demonstrated in the present research was found to have several distinct advantages. Among them:

1. It is an authentic, integrative task that is firmly construct referenced.
2. It is highly efficient and consistent.
3. It allows for large "N" research and testing.
4. It greatly reduces scoring subjectivity.
5. It provides a window into the L2 reading comprehension process.
6. It is low cost, and uses readily available, state-of-the-art technology.

Finally, the procedure addresses Calfee and Hiebert's (1991: 282) six key assessment design issues:

1. It is valid to the extent that it is firmly grounded on a sound reader-based theory of L2 reading comprehension (Bernhardt, 1991, 2000), uses authentic texts (Bernhardt and Berkemeyer, 1988), and requires the subjects to respond in their L1 (Alderson, 2000; Lee, 1986).
2. The procedure uses the most direct outward manifestation of the reading comprehension process (Lee, 1986), and is thus the most suitable method to obtain appropriate and meaningful measures of L2 reading comprehension ability.
3. The data provided by the procedure can easily be subjected to analysis and are readily available to the teacher or researcher.
4. The procedure has an inherent consistency in scoring, far superior to what human raters can achieve. Thus, reliability is significantly enhanced.
5. The procedure is highly efficient in terms of cost (time, money, and degree of effort required to gather the information).
6. The data collected have been shown to be readily aggregable because a vast database of subject-generated information can be collected for analysis.

Recommendations for further research

For the automated recall protocol assessment system to become a truly valuable tool for educators and researchers, however, the advantages mentioned are clearly not enough. In order to be more than just useful, it must also provide insight into the recall process, because "a successful assessment mechanism for L2 reading comprehension must provide in-depth information on *how* readers cope with text while, at the same time, providing quantifiable data for large-scale comparison and contrast" (Bernhardt, 1991: 194).

Continuing research using automated recall protocol scoring must take several paths. First, although high correlations have been achieved here between manually and computer scored recalls, research is needed to improve the correlations. The weak-scoring rule must be refined, and this can be accomplished by looking more closely at propositional location in the recall versus the original text. Time on both the reading and writing tasks must also be examined as potential indicators of comprehension ability.

The assigning of pausal unit hierarchies needs further exploration. Currently they are determined by a team of native (or near native) speaking, trained raters working independently. The analysis of recall across many subjects, however, may reveal that L2 learners at various levels "rate" propositional importance differently than the native speakers. In addition, like Schmidt-Rinehart (1994: 186), during the present analysis, the author found evidence that the weighted four-point scoring was equivalent to an unweighted system. To simply throw away hierarchies, however, would be to deny that comprehenders unknowingly rate some propositions within a text as being more important than others and tend to build comprehension around these structures. While additional research is needed, data from the present study indicate that this denial is not valid.

Finally, another avenue of exploration is the use of concordancing programs to analyze the student-generated data. While the quantitative scores generated by the program may provide some insight into the level of ability of the individual subjects, concordancers may provide us the vehicle for further qualitative inquiry into L2 reading development to help clarify the unexplained and unexplored regions of the Bernhardt (2000) model. We may be able to locate verb tenses and determine through correlation whether or not readers are making links between and among tense versions in the L2 and their L1 reconstructions. In addition, such programs may help us look for vocabulary development over time and examine recency and latency effects in recalls. For example, are there consistencies in recalled items at the beginning, middle, and end of recalls?

Summary and conclusion

The automated recall protocol procedure has the potential to greatly enhance L2 reading comprehension assessment. The present study demonstrates that the procedure can be used to perform large-scale assessment, can lead to enhanced reading comprehension test development, and can improve diagnostic and placement testing. No longer will researchers need to rely strictly on discrete-point examinations that provide quantitative information but little else. Furthermore, unlike the multiple choice or cloze tests, the present procedure is a truly integrative, authentic-task firmly grounded on a reader-based construct of reading comprehension. In addition, compared to development of valid and reliable multiple-choice instruments, an often lengthy and expensive process (Maarof, 1998), development of a valid recall protocol assessment rubric is relatively easy. Using the present computer scoring procedure, quantitative score information is readily available and is directly linked to a qualitative database ripe for additional examination.

Diagnostic and placement testing

Further analysis of student-generated data enabled by the automated system will greatly enhance diagnostic testing as target language reading comprehension difficulties become evident, and thus will provide information to empower dynamic and active instruction. Such analyses can be streamlined through the use of word processors, concordancing programs, or statistical significant probability analysis. The automated recall protocol procedure also provides important data that can be used as the basis of comparison to make placement decisions. In sum, as an integrative measure, the automated recall protocol has the potential to become a powerful tool especially when used in a multiple measures assessment approach for research, and may help explain the unexplainable and unknown and further enhance L2 reading comprehension model development.

Pedagogy

Moreover, the automated procedure used in the present study has the potential to inform L2 reading comprehension pedagogy. Understanding the state of a student's L2 reading comprehension ability can be an extremely important factor in deciding appropriate reading strategies to teach and develop. With this knowledge, teachers can guide and possibly accelerate students through development of enhanced L2 reading strategies and skills. Furthermore, once the scoring templates are developed, the automated recall protocol procedure provides ready

access to student-generated reading comprehension data. These data can help classroom teachers address comprehension problems students ***are*** having, not just those that someone thinks they might be having *a priori*.

Teachers will no longer be reliant on expensive, development-intensive multiple-choice examinations that provide limited data. Nor will they need to rely on cloze examinations that are not truly integrative and thus fail to tap constructivist L2 reading comprehension constructs, as used here. For the first time, the L2 teacher can have a powerful assessment device that is easy to construct, is immediately available, uses authentic material that he or she chooses, and can be used to provide valid, reliable, and efficient reading comprehension assessments throughout the course. Used in a multiple-measures approach, the automated reading recall protocol can provide information that teachers can use to tailor instruction to student needs and thus promote active, dynamic, learner-centered instruction. Thus, ultimately, the use of this assessment procedure will not only help provide an accurate picture of the state of the student's development but may also help accelerate that development. Making the recall protocol procedure a usable reality through fully automated delivery and scoring will provide teachers with the ability to use the results (both quantitative and qualitative) for placement, diagnostic and classroom testing, and instruction.

References

Alderson, J. C. (2000). *Assessing Reading*. Cambridge, MA: Cambridge University Press.

Allen, E. D., Bernhardt, E. B., Berry, M. T., & Demel, M. (1988). Comprehension and text genre: Analysis of secondary school foreign language readers. *Modern Language Journal*, 72(2), 163-172.

Baker, F. B. (1984). Technology and testing: State of the art and trends for the future. *Journal of Educational Measurement*, 21(4), 399-406.

Bartlett, F. C. (1932). *Remembering*. Cambridge: Cambridge MA.

Berkemeyer, V. C. (1989). Recall protocol data: Some classroom implications. *Die Unterrichtspraxis*, 21(3), 131-137.

Berkemeyer, V. C. (1991). *The effect of anaphora on the cognitive processing and comprehension of readers of German at various levels of baseline German language ability*. Unpublished doctoral dissertation, The Ohio State University, Columbus, Ohio.

Bernhardt, E. B. (1983a). Testing foreign language reading comprehension: The immediate recall protocol. *Die Unterrichtspraxis,* 16, 27-33.

Bernhardt, E. B. (1983b). Three approaches to reading comprehension in intermediate German. *Modern Language Journal*, 67, 111-115.

Bernhardt, E. B. (1985). Reconstruction of literary texts by learners of German. In M. Heid (Ed.), *New Yorker Werkstattgespräch 1984: Literarische texte im Fremdsprachenunterricht* (pp. 255-289). München: Kemmler and Hoch.

Bernhardt, E. B. (1986a). Cognitive processes in L2: An examination of reading behaviors. In J. Lantolf & A. Labarca (Eds.), *Delaware symposium on language studies: Research on second-language acquisition in a classroom setting* (pp. 35-51). Norwood, NJ: Ablex.

Bernhardt, E. B. (1986b). Reading in the foreign language. In B. H. Wing (Ed.), *Listening, reading and writing: Analysis and application* (pp. 93-115). Middlebury, VT: Northeast Conference on the Teaching of Foreign Languages.

Bernhardt, E. B. (1990). A model of L2 text reconstruction: The recall of literary text by learners of German. In A. Labarca & L. M. Bailey (Eds.), *Issues in L2: Theory as practice/practice as theory* (pp. 21-43). Norwood, NJ: Ablex.

Bernhardt, E. B. (1991). *Reading development in a second-language*. Norwood, NJ: Ablex.

Bernhardt, E. B. (2000). Second-language reading as a case study of reading scholarship in the 20th century. In M. L. Kamil, P. D. Pearson, & R. Barr (Eds.), *Handbook of reading research, Vol. III* (pp. 791-811). New Jersey: Lawrence Erlbaum Associates.

Bernhardt, E. B. & Berkemeyer, V. C. (1988). Authentic texts and the high school German learner. *Die Unterrichtspraxis*, 21(1), 6-28.

Bernhardt, E. B. & Deville, C. (1991). Testing in foreign language programs and testing programs in foreign language departments: Reflections and recommendations. In R. V. Teschner (Ed.), Issues in language program direction: *Assessing foreign language proficiency of undergraduates* (pp. 43-59). Boston, MA: Heinle & Heinle Publishers, Inc.

Bernhardt, E. B. & Kamil, M. L. (1995). Interpreting relationships between L1 and L2 reading: Consolidating the linguistic threshold and the linguistic interdependence hypothesis. *Applied Linguistics*, 16, 15-34.

Brisbois, J. E. (1992). *Do first language writing and second-language reading equal second-language reading comprehension? An assessment dilemma.* Unpublished doctoral dissertation, The Ohio State University, Columbus, Ohio.

Calfee, R. & Hiebert, E. (1991). Classroom assessment of reading. In R. Barr, M. L. Kamil, P. B. Mosenthal, & P. D. Pearson (Eds.), *Handbook of reading research: Vol. 2* (pp. 281-309). New York: Longman.

Coady, J. A. (1979). Psycholinguistic model of the ESL reader. In R. Mackay, B. Barkman, & R. R. Jordan (Eds.), *Reading in a second-language* (pp. 88-96). Rowley, MA: Newbury House.

Craik, F. & Lockhart, R. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior*, 11, 671-684.

Cummins, J. (1979). Cognitive/academic language proficiency, linguistic interdependence, the optimum age question and some other matters. *Working Papers on Bilingualism*, 19, 197-205.

Guindon, R. & Kintsch, W. (1984). Priming Macropropositions: Evidence for the primacy of macropropositions in the memory for text. *Journal of Verbal Learning and Verbal Behavior*, 23, 508-518.

Heinz, P. J. (1993). *Towards enhanced, authentic second-language reading comprehension assessment, research, and theory building: The development and analysis of an automated recall protocol scoring system.* Unpublished doctoral dissertation, The Ohio State University, Columbus, Ohio.

Johnson, R. (1970). Recall of prose as a function of the structural importance of linguistic units. *Journal of Verbal Learning and Linguistic Behavior*, 9, 12-20.

Kintsch, W. (1974). *The representation of meaning in memory*. New York: John Wiley & Sons.

Kintsch, W. (1988). The use of knowledge in discourse processing: A construction-integration model. *Psychological Review*, 95, s163-182.

Kintsch, W. & van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological Review*, 85, 363-394.

Kintsch, W., Welsch, D., Schmalhofer, F., & Zimny, S. (1990). Sentence memory: A theoretical analysis. *Journal of Memory and Language*, 29, 133-159.

Lee, J. F. (1986). On the use of the recall task to measure L2 reading comprehension. *Studies in Second-language Acquisition*, 8, 83-93.

Maarof, N. (1998). *Assessing Second-language Reading.* Selangor, Malaysia: Faculty of Language Studies, Universiti Kebangsaan Malaysia.

Matussek, M. (1992). Aufstand im Kinderzimmer. *Der Spiegel*, 46(27), 193-198.

Nitko, A. J. & Hsu, T. (1984). A comprehensive microcomputer system for classroom testing. *Journal of Educational Measurement*, 21(4), 377-390.

Pfaffenberger, B. (1988). *Microcomputer applications in qualitative research*. Newbury Park, CA: SAGE Publications, Inc.

Riley, G. L. & Lee, J. F. (1996). A comparison of recall and summary protocols as measures of second language reading comprehension. *Language Testing*, 13(2), 173-189.

Rodrian, H. W. (1990). Auf die Schnelle in die Ferne. Erst packen, dann buchen. *AZ-Journal*, 2, 103-104.

Rumelhart, D. (1980). Schemata: The building blocks of cognition. In R. J. Sprio, B. C. Bruce, & W. F. Brewer (Eds.), *Theoretical issues in reading comprehension* (pp. 33-58). New Jersey: Lawrence Erlbaum.

Schmidt-Rinehart, B. C. (1994). The effects of topic familiarity on second-language listening comprehension. *Modern Language Journal*, 78, 179-189.

Smith, C. (1989). Text analysis: The state of the art. *The Computer-Assisted Composition Journal*, 3, 68-77.

Smith, G. W. (1991). *Computers and human language*. New York: Oxford University Press.

Tang, M. S. (1985). Microcomputers, software and foreign languages for special purposes: An analysis of TXTPRO. *Proceedings of the Eastern Michigan University Conference on Language for Business and the Professions*, 4, 172-260.

Tulving, E. & Thomson, D. M. (1973). Encoding specificity and retrieval processes in episodic memory. *Psychological Review*, 80, 352-373.

Valencia, S. W. (1990). Alternative assessment: Separating the wheat from the chaff. *The Reading Teacher*, 44(1), 60-61.

Weaver III, C. A. & Kintsch, W. (1991). Expository text. In R. Barr, M. L. Kamil, P. B. Mosenthal, & P. D. Pearson (Eds.), *Handbook of reading research: Vol. 2* (pp. 230-245). New York: Longman.

Appendix: Recall protocol texts

Batman -- Aufstand im Kinderzimmer (Matussek, 1992: 193)

"Er soll ja besser sein als der erste", sagt die 14jährige Maria, die mit ihrer Freundin vor dem Loews-Kino am Broadway ansteht und nur millimeterweise vorrückt in der Schlange, die sich vor der Kasse gebildet hat. Obwohl sie an diesem Wochenende mithelfen wird, einen Rekord zu brechen, klingt sie nicht gerade begeistert. Es klingt wie: mitmachen und absitzen. Hier wird kein Fest angesteuert, sondern eine Hypnose.

Für Maria steht der Termin seit Wochen fest, auf einem Plakat, drei Stockwerke hoch über Times Square, schwarz auf gelb: "Batman kehrt zurück". Der Film.

Batman, die Geldmaschine, spuckt wieder. Bereits im ersten Anlauf vor zwei Jahren hatte der Mann mit der Fledermausmaske Platz sechs in der Liste der besten Filme aller Zeit geschafft. Nun spielte die Fortsetzung schon am ersten Wochenende 46,5 Millionen Dollar ein. Weltrekord.

Alle amerikanische Kinder seit 1939 sind mit Batman groß geworden. Der Fledermaustyp mit der tragischen Kindheit ist ein schüchterner einsamer Mensch, der sich verwandelt, wenn er sich die Maske überstülpt. Batman, tagsüber braver Bürger, ist der Lotse durch die Schattenwelt.

Bernhardt Letter (Bernhardt 1991: 124)

Prof. Dr. E. Buchter-Bernhardt
227 Arps Hall
1945 N. High Street
The Ohio State University
Columbus, OH 43210
U S A

Liebe Frau Buchter-Bernhardt,

in der Anlage finden Sie die Dinge, die ich Ihnen in Newark versprochen habe. Wenn Sie an dem einen oder andern von uns interessiert sein sollten, können wir dies gerne kopieren.

Unnötig zu sagen, daß es großen Spaß gemacht hat, Sie kennenzulernen, mit Ihnen zu plaudern und gemeinsame Interessen und Bekannte zu entdecken.

Ob Sie so nett sein könnten, mir bei Gelegenheit den Namen und die Adresse Ihres Mitarbeiters, der jetzt in Virginia ist, mitzuteilen, damit ich auch ihm die versprochenen Materialien schicken kann. Ich vergaß, mir seine Adresse aufzuschreiben.

Mit den besten Grüßen und allen guten Wünschen bin ich

Ihr

Auf Die Schnelle In Die Ferne -- Erst Packen Dann Buchen (Rodrian, 1990: 103)

So schnell kann es gehen: Am Dienstag letzter Woche dachte Peter Frisch noch darüber nach, ob er sich einen Trip nach Spanien leisten könnte. Am Donnerstag jettete er dann doch lieber nach San Franzisko. 895 Mark fürs Ticket nach Kalifornien und zurück -- dieses Angebot hatte den Münchner nicht lange Zögern lassen. Muß man vielleicht mit einer Stewardess verlobt sein, um so billig um die halbe Welt zu jetten? Des Rätsels Lösung ist viel einfacher: Als den Münchner das Fernweh überkam, hatte er sich bei den Last-Minute-Büros umgehört. Bei der Tonband-Ansage von L'Tours wurde er fündig.

Noch rascher ging's beim Münchner Studenten Manfred Kanzler: er packte einfach Zahnbürste uns Scheckbuch ein und fuhr zum Flughafen. Da hatte er noch keine Ahnung, wohin die reise gehen sollte. Drei Stunden später saß er schon im Jet nach Eliat am Roten Meer -- für 498 Mark. Für Verkäuferin Beate Baskos vom ABR-Last-Minute-Service am Flughafen ist das nichts Ungewöhnliches: Sie vermittelt jedes Wochenende Ferienglück gleich dutzendweise in letzter Minute. Der Schluß-Verkauf von Urlaubsreisen, vor drei Jahren noch fast unbekannt, erlebt jetzt den großen Boom.

About the Author

Dr. Peter J. Heinz is the Dean of Languages at Pikes Peak Community College in Colorado Springs, Colorado. He holds a Ph.D. from The Ohio State University in Second Language Education and spent 16 years teaching German at the United States Air Force Academy, as an Associate Professor, Language Learning Center Director, Deputy Department Head, and as the Director of International Programs.