AMCIS 2004 Proceedings

Americas Conference on Information Systems (AMCIS)

12-31-2004

Computer Aids for Designing Effective Multiple Choice Questions

Morgan Benton
New Jersey Institute of Technology

Marilyn Tremaine
New Jersey Institute of Technology

Julian Scher
New Jersey Institute of Technology

Computer Aids for Designing Effective Multiple Choice Questions

Morgan C. Benton
Information Systems Department
New Jersey Institute of Technology
mcbenton@njit.edu

Marilyn M. Tremaine
Information Systems Department
New Jersey Institute of Technology
tremaine@njit.edu

Julian M. Scher
Information Systems Department
New Jersey Institute of Technology
scher@njit.edu

ABSTRACT

With minimal training in assessment design, professors must generate high quality examinations that fairly assess student accomplishment of course objectives. Multiple choice questions are frequently used for this assessment, but often do not address learning objectives, are misinterpreted, and if reused, may challenge the integrity of some students. It is proposed that computer aids for generating multiple choice tests might be of considerable use to an instructor. This study presents a first step in building such aids, that of demonstrating via a case study, the inadequacy of current multiple choice test item generation methods. The case study also uncovers the functionality to build into a computer test generation aid for instructors. Bloom's taxonomy of learning objectives, discriminant analysis and student evaluations of questions is used to analyze the question data used in this study.

Keywords

Learning objectives, Bloom's taxonomy, multiple-choice questions, instructor support software.

INTRODUCTION

Teachers in K-12 and higher education often lack the knowledge and skills necessary to create effective assessments. They also lack the understanding to correctly interpret assessment results and use them to adapt future instruction (Stiggins, 2001). In the context of a growing movement toward school accountability, exemplified by such legislation as the No Child Left Behind Act (NCLB), this shortcoming of our frontline instructors is the elephant in the classroom—while most would acknowledge the problem with classroom assessment, the enormity of the problem discourages action toward a solution. There are several forms such action might take, such as the improvement of psychometric education in teacher training programs or appropriate legislation to require teachers to have knowledge of assessment. Such solutions will not, however, produce quick results.

Software could be built to convert psychometricians' knowledge and expertise into a form accessible to classroom instructors. Already a great deal of heuristic understanding of how to generate high-quality assessments exists, much supported by empirical evidence (Haladyna, Downing, & Rodriguez, 2002). Furthermore, promising developments are being made towards developing empirically defensible cognitive taxonomies of educational objectives (Anderson & Krathwohl, 2001).

There are many forms of instructor assessment that could be supported. The current project will build an interface to aid instructors in the generation of better-quality selected response items, or multiple choice questions (MCQs). The choice of this particular type of assessment item is based on several factors:

1. Multiple-choice questions are frequently used, although difficult to construct (McKeachie, 2002).
2. Students are highly familiar with this assessment mechanism.
3. Multiple-choice items lend themselves well to online delivery and computer grading.

4. There is a great deal of collected wisdom as to how to create high-quality multiple-choice questions (Haladyna et al., 2002).
5. Much of this wisdom could easily be implemented in software.

Specifically, what would such software do? Ideally, the system could guide an instructor through the process of formulating appropriate and specific educational objectives, and then suggest multiple-choice question formats that will measure student mastery of those objectives. As the instructor is creating questions, the system could provide guidance and feedback that would help prevent the instructor from falling into common question-generation pitfalls. Once the test has been administered and scored, the system could generate statistical feedback, help the instructor interpret that feedback, and provide suggestions to modify either the instruction and/or the questions, to improve student learning and/or its assessment, respectively.

Task analysis is one of the first steps in a user-centered software design process. "Contextual Task Analysis," writes Mayhew, "is most appropriate when you already have a set of functions and features identified and scoped out for automation; you primarily need to understand the identified current work in order to optimally support it with a new product such as a software application" (Mayhew, 1999, p.67). This is the case in this research. Our goal is to support instructors' generation of multiple-choice questions. To do so we need to thoroughly understand the task.

This study presents a detailed task analysis of the multiple choice test item generation carried out by a university professor. The study compares the test items to educational objectives, to students' perceptions of the questions, and to a discriminant analysis that indicated which questions were better at sorting out the poor students from the good students. The results of the analysis both demonstrate that the test items generated without aids were poor and that software could readily be created to support this instructor in creating better multiple choice items.

This paper is organized as follows. The next section summarizes relevant literature and evaluates existing software with respect to the generation of high quality multiple-choice questions. The third section describes the actual task analysis study that was performed. The fourth section presents the principal results. The discussion that follows explores the implications of what was found in the context of designing a system to support instructors. Finally, future work to be done is briefly stated.

WHAT OTHERS HAVE DONE

Supporting instructors' creation of valid and appropriate MCQs involves three processes. First, for an assessment to be effective, it is necessary to specify the educational objective being assessed. Bloom's taxonomy of educational objectives theorizes that depending on the level of the cognitive processes being tested, different types of MCQ may be appropriate (Anderson & Krathwohl, 2001). Second, creating valid questions requires adhering to sound design guidelines. Over the years researchers, using empirical validation (Haladyna et al., 2002), have developed a collection of recommendations for the construction of valid MCQs. Third, even if a question validly measures a given objective, it may still be too difficult or too easy for the target student population. It is necessary to make some determination as to the level of question difficulty. Several methods exist for doing this, including expert evaluation (e.g. Newman & Taube, 1996), test-taker evaluation (e.g. Prestwood & Weiss, 1977), feature-based evaluation (e.g. Green, 1983), and statistical evaluation using item-response theory (IRT)(Hambleton, Swaminathan, & Rogers, 1991). The premise of this software design project is that a significant portion of these three processes could be automated.

Educational Objectives and Bloom's Taxonomy

Instructors should identify and plan instruction according to explicit objectives. Sufficient definition of educational objectives is needed if test items are to be connected to classroom instruction. Assessment not aligned with objectives can often be demoralizing and damaging to the self-esteem of students who cannot adequately prepare for examinations (McKeachie, 2002). At the K-12 level, states mandate extremely fine-grained objectives at every grade level because of the movement toward standards-based education. In higher education, however, determining appropriate learning objectives remains a challenge. Specific objectives for courses may or may not be set by accrediting bodies, and even then may only be examined or enforced in years when an academic department is to be accredited.

Location in the Taxonomy	**Example Objectives—Students will:**
1A—Remember Factual Knowledge	1. Know the meanings of acronyms related to the HCI field
2B—Understand Conceptual Knowledge	2. Be able to classify examples of key HCI concepts
2C—Understand Procedural Knowledge	3. Understand when to apply different interface evaluation techniques

Figure 1--Examples of objectives in different locations in Bloom's Taxonomy

The goal of Bloom's Taxonomy of Educational Objectives is to assist instructors in generating well-defined learning objectives for their students, and also in generating assessments that are aligned with the objectives. The original taxonomy divided cognitive processes into six levels, which were thought to be cumulative, i.e. achievement at higher cognitive levels required mastery of cognitive skills at all lower levels. Anderson and Krathwohl (2001) recently expanded the taxonomy to further specify a knowledge dimension. Using the taxonomy, an instructor may develop objectives, such as those given in Figure 1. The taxonomy also suggests the format for assessment items aligned with such objectives (Figure 2).

MCQs often assess only what is easy to measure, failing to measure mastery at higher levels of cognitive complexity. Explicit incorporation of Bloom's taxonomy into the instruction design process may help solve this problem. However, Anderson and Krathwohl (2001) point out that there has been only limited validation of Bloom's Taxonomy (c.f. Chapter 16). The theory that the cognitive levels are cumulative has received mixed support, and it is sometimes difficult to map MCQs onto distinct categories in the taxonomy—there seems to be overlap in some cases.

1. CSCW stands for:
   a. Computer Supported Cooperative Work
   b. Collaborative Systems Coordinating Work
   c. Computer Supported Creative Work
   d. Computerized Sequence Counter Work

2. The table of contents at the beginning of a book is an example of:
   a. feedback
   b. affordance
   c. a forcing function
   d. mapping

3. Protocol analyses are best conducted:
   a. While the subject is trying to figure out how to do a task with the interface.
   b. Immediately following a user walkthrough when details of the user interface are fresh in the user's mind
   c. By having two protocol experts examine the wording used in the interface
   d. After a user has worked with the interface extensively for at least six months

Figure 2--Examples of Multiple-Choice Questions that Assess Objectives at Bloom's Taxonomy Levels 1A, 2B, and 2C, respectively

Guidelines for Valid Multiple-choice Question Construction

Research has been performed to determine how best to construct MCQs so that they will have the best chance of providing valid measures of target knowledge. Haladyna et al. (2002) did a review of the literature, compiling a list of guidelines for writing MCQs (Figure 3).

Content Concerns	**Writing the Choices**
1. Specific content, single specific mental behavior	18. As many effective choices as possible; 3 is adequate
2. Important content; avoids trivial content	19. Only ONE right answer
3. Novel material for testing higher level learning	20. Location of right answer is varied
4. Content independent of other items	21. Choices in logical or numeric order
5. Not too specific, nor too general	22. Choices are independent; non-overlapping
6. Not based on opinion	23. Homogenous content and grammatical structure
7. Not a trick question	24. Choices are about the same length
8. Vocabulary simple for group being tested	25. *None-of-the-above* used very carefully
Formatting Concerns	26. Avoids *all-of-the-above*
9. MC, TF, MTF, matching, context-dependent; NOT complex	27. Choices phrased positively; avoids negatives such as NOT
10. Vertical formatting	28. Avoids giving away the right answer by using:
Style Concerns	28a. Specific determiners, e.g. always, never, completely, etc.
11. Edited and proofed	28b. Clang associations: identical or resembling words in stem
12. Correct grammar, punctuation, caps, and spelling	28c. Grammatical inconsistencies that clue test-taker to answer
13. Minimize reading for each item	28d. Conspicuous correct choices
Writing the Stem	28e. Paris or triplets of options that clue test-taker to answer
14. Directions in the stem are clear	28f. Blatantly absurd or ridiculous options
15. Central idea is in the stem, not in the choices	29. Make all distractors plausible
16. Avoids excessive verbiage	30. Uses typical errors of students to write distractors
17. Avoids or cautiously uses NOT or EXCEPT	31. Only uses humor if compatible with teacher and environs

Figure 3—Guidelines for MCQs (Haladyna et al., 2002)

Assessment of Question Difficulty

Instructors must ensure that test item difficulty matches the students' ability levels—tests that are either too easy or too hard for a given set of students yield less useful feedback and may have a serious impact on student motivation. Therefore, the software should help the instructor predict a given question's difficulty. Difficulty is commonly measured using four methods: (1) expert evaluation by instructors, (2) evaluation by test-takers, (3) evaluation based on item-features, and (4) statistical evaluation based on item response theory. Even with these methods, the factors that contribute to test item difficulty are not yet clearly understood.

One set of studies attempts to gauge the accuracy of various predictors of item difficulty by correlating them with statistical measures(Adams, 1993; Newman & Taube, 1996; Prestwood & Weiss, 1977). In this work, expert judges or students rated the difficulty of test items. The ratings were compared to statistical difficulty levels determined from the results of the test taking. Generally these studies found that human judges are accurate predictors of item difficulty. These studies could not, however, determine which elements make one question more difficult than another.

Another set of studies hypothesizes that one or more item properties are related to item difficulty(Enright, 1993; Green, 1983; Scheuneman, 1991). In these studies the typical properties that were selected for study included item structure, readability, semantic content, cognitive demand/complexity, knowledge demand, and language difficulty. The results were mixed. Semantic content of distractors, and the level of knowledge demanded, appear to be significant predictors of difficulty. Surprisingly, language difficulty does not seem to be consistently related to item difficulty. Cognitive demand/complexity as a factor gives mixed results. The cognitive properties of test items are particularly important for us to understand in light of their possible relationship to the cognitive taxonomies discussed in the previous section.

Evaluation of Existing Software for Question Generation

Though not comprehensive, Rocklin (1999) provides a list of some major software packages that support test generation and delivery. Rocklin cites a University of Iowa committee whose task was to generate a list of ideal features for software to support testing at a university. This list includes features that would support the generation of objectives aligned with a cognitive taxonomy, e.g. Bloom's taxonomy. The list also requested software capable of IRT-based statistical analysis, as well as providing guidance in the interpretation of those analyses. A survey of existing software shows that, at best, these packages provide standard templates for formatting MCQs with no content guidance. Several of the high end systems provide extensive statistical analysis. One notable system—CAPA, developed at the University of Michigan—actually contains over one hundred content templates for physics problems. The actual questions are generated on-the-fly by the system (Kashy, Albertelli, Thoennessen, Tsai, & Kashy, 2000). No systems were found that supported the generation of learning objectives aligned with any taxonomy. The IMS Global Learning Consortium Learning Design module specifically discusses learning-objectives, but there is no reference to taxonomies, and the precise role of objectives is not defined (IMS Global Learning Consortium, 2004).

DESCRIPTION OF THE STUDY

Twenty-four students participated in this study. All were enrolled in an upper-level undergraduate, distance-learning course in human-computer interaction (HCI) at a mid-sized, urban, northeastern, technical university. The students in the course completed all course requirements online, except for taking the midterm and final examinations on campus. As an incentive to participate in the study, students could opt out of two of six course assignments, in exchange for which they would take four, ten-item, multiple-choice quizzes and also fill out a questionnaire following each quiz. The questions on the quizzes corresponded to course content for the week of the quiz. To motivate students to take the quizzes seriously, the quiz scores counted towards the final semester grade. Twenty-one students returned completed questionnaires on at least one of the four quizzes, for a total of seventy-six questionnaires. Two students' questionnaires were eliminated because of integrity concerns, and one student completed the questionnaires without completing the quizzes. Six subjects were female. The class represented a very diverse ethnicity.

The quizzes were delivered online via WebCT™, a commercial course-management software package. Using the software, students were allowed access to the quizzes at specified times, and were restricted to a contiguous thirty-minute time limit for completing the quiz. Upon quiz completion, the system provided students with a link to the online questionnaire. The questionnaire asked seven questions about each of the ten quiz items, followed by four questions about the quiz as a whole. The questions asked the students to rate the difficulty of each question, to identify the reasons why they missed questions, and to rate the clarity and fairness of each question. Summary questions asked how long subjects studied for the quiz, how carefully they read the instructions, how clear the instructions were, and what could have been done by either the student or instructor to improve the quiz.

The instructor for this course was a veteran professor with several decades of teaching experience. At the time the quiz questions were generated, explicit educational objectives had not been specified for this course. The instructor generated questions by surveying the readings and lecture notes, selecting representative knowledge deemed to be important, and then writing questions thought to assess mastery of this knowledge. The collective experience of the researchers suggests that this is how instructors generally create exams.

RESULTS

This report focuses on two methods that were used to analyze the question generation task: categorization of questions into Bloom's taxonomy, and analysis of quiz results informed by the student questionnaire feedback.

Table 2 shows the breakdown of the questions into the Bloom categories. The test items were initially categorized separately by two researchers. Disputes over the categorization of the items were resolved by the instructor. The following lessons were learned during the categorization process:

1. It was impossible for an outsider to categorize some of the questions. The categorization required a distinction between Remember and Understand questions. Only the instructor would know if the students had been explicitly exposed to the material in the question beforehand, placing the item in the "Remember" category.
2. Classification was difficult when answer choices assessed different cognitive processes. In one case this ambiguity appeared to confuse students resulting in a low number of correct responses.
3. It is likely that some cells in the table will never have entries. For example, only procedural knowledge works for the "Apply" situation, if Anderson and Krathwohl's definition is followed.
4. In general, the robustness of the taxonomy as a categorization scheme seems an open question. It was a challenge at times to work backward from a question and fit it into a category. In fact, the authors of the taxonomy don't advocate this practice and intend that the taxonomy be used to develop objectives prior to item generation. If done in this direction it seems likely that questions would always fit cleanly into a single category.

Table 2 indicates that all of the questions fell into the first two cognitive categories: Remember and Understand. No questions tested cognitive skills at higher levels. Although it was expected that no questions would fall into the Create category (by definition one cannot assess creativity with a closed assessment item), it was somewhat surprising that no questions were categorized as Apply, Analyze, or Evaluate since much of the course material would require students to Apply procedures, or to Analyze or Evaluate interfaces. Several explanations for this lack of variety seem plausible:

1. Multiple-choice questions may not lend themselves well to assessing higher-level cognitive mastery. Although Anderson and Krathwohl's book indicates item formats for the higher-level processes, there doesn't seem to be satisfactory empirical evidence indicating their validity (see Chapter 17, p.298).
2. Without explicit practice or guidance in generating items to assess a variety of cognitive skills, it may be that instructors naturally fall into a habit of writing items that fit a limited number of situations. It is not that items to test higher levels aren't possible, it is just that instructors aren't aware of the differences and hence don't vary their items.
3. The instructor subconsciously wrote items on which it was felt that the students would do well, and hence didn't write questions to assess higher-level skills. This explanation is rather difficult to test, and not supported by the levels of success of the students on the quizzes (the overall average was just over 60%).
4. Questions at higher levels of cognition were not appropriate for the given course content. This does not seem likely, but given that the instructor had never defined explicit learning objectives, it may turn out to be the case.

Regardless of the reasons for the lack of variety in question types, and the difficulties in categorization, it is believed that a software system that guides the instructor in the development of objectives will result in questions that assess a greater variety of cognitive skills, and which are more accurately categorized.

The Knowledge Dimension	The Cognitive Process Dimension					
	1. Remember	2. Understand	3. Apply	4. Analyze	5. Evaluate	6. Create
A. Factual Knowledge	**2**	0	0	0	0	0
B. Conceptual Knowledge	**7**	**21**	0	0	0	0
C. Procedural Knowledge	**6**	**4**	0	0	0	0
D. Meta-Cognitive Knowledge	0	0	0	0	0	0

Table 1—Frequencies of Quiz Questions Appearing in Bloom's Taxonomy Categories (N=40)

We arbitrarily selected items that less than 45% or greter than 85% of the students answered correctly for more detailed examination. We deemed that these questions were either too hard or too easy. We look at three of the low scoring questions.

Question Title	**N**	**% Correct Of:**			**Discrimination**	**Score**	
		Whole Group	Upper 25%	Lower 25%		Mean	SD
Heuristic vs. Cognitive Evaluations	20	60	85	50	0.31	60.00%	50
User-centered design lifecycle	20	60	71	37	0.37	60.00%	50
User Walkthrough	20	65	100	37	0.61	65.00%	49
Conceptual Model (subway system)	20	65	85	50	0.27	65.00%	49
Paper-clip Icon Example	**20**	**30**	**42**	**25**	**0.38**	**30.00%**	**47**
Table of Contents	**20**	**90**	**100**	**75**	**0.54**	**90.00%**	**31**
Affordance definition	20	70	85	50	0.22	70.00%	47
Questionnaire Use	20	60	57	50	0.15	60.00%	50
Cell phone design goals	**20**	**25**	**57**	**0**	**0.59**	**25.00%**	**44**
Cognitive Walkthroughs	20	70	100	50	0.46	70.00%	47
					Overall Mean:	**59.50%**	

Table 2—Results from Quiz #1 as Reported by WebCT™

Figure 4 gives an example of a poor question that 25% of the students answered correctly. The student feedback on the question indicated that students interpreted the question to be asking for the most representative example of a usability goal whereas the instructor wanted the students to indicate that they understood that that usability goals had to be quantifiable. Changing "valid" to "measurable" fixed the question's ambiguity. If a software alert had noted that the answers were not parallel structures, the problem would have been avoided.

Which of the following is a valid usability goal for the design of cell phones?

a. On average, the time it will take a user of the new interface to obtain a voicemail message will be 30 seconds
b. Users will find it easy to learn its basic functions
c. Users will be completely satisfied with the interface
d. 80% of users will find it easy to learn

Figure 4—Bad Question Resulting from Assessment of Multiple Concepts

Figure 5 shows two questions that suffered from the problem of having poor examples to illustrate a concept. In the first question the student must choose the correct concept that applies to the given example; in the second question the student must choose the correct example that corresponds to the given concept. In the first question, the example is not really a good example of any of the concepts, and in the second, the examples are all too similar to one another, causing confusion.

Overall, the student feedback and the discriminant analysis indicated that the difficulty of most of the questions was appropriate, but that the questions did not assess the higher cognitive skills being taught to the students. When questions were problematic, it was typically their phrasing or mixing of concepts that caused students difficulty.

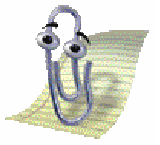
The following paper clip icon that is used to summon help in Microsoft Word is an example of:  a. affordance b. mapping c. conceptual model d. feedback e. is expressed by both items A and B above	A story board a. is a rough method for describing and laying out the design of a user interface b. is a prototyping method that describes the linkages between the various functions that are being built in the user interface c. is a way of providing help in a user interface that is both painless and fun d. is a method for designing user interfaces by building a story that explains to the user each and every function of the application e. is a rough method for evaluating the early design of a user interface

Figure 5—Bad Questions Resulting from Poor Exemplification of Concepts

DISCUSSION

This study reinforces the claims made by Stiggins (2001) that instructors do not know how to design effective assessments. Three key design implications were gleaned from the case study.

1. Guided Implementation of Educational Objectives

Instructors need guidance in the creation of explicit educational objectives and the associated questions that assess a student's mastery of the objectives. Few university level educators are aware of Bloom's taxonomy. An interface design that guides an instructor in stating specific course objectives and then supports the generation of test items via online templates that guide question writing for multiple levels of Bloom's taxonomy is a suggested design option. Statistics on the distribution of the items across six levels of Bloom's taxonomy would also be a useful aid for the instructor.

2. Real Time Feedback on Item Difficulty

As test items are written, feedback on the grammatical form and word usage of a question would also serve as a useful guide for an instructor. At a rudimentary level, this feedback could take the form of spelling and grammar checking. At a higher level, the system could check for grade level of word usage, parallelism of multiple choice answers, the use of confusing negative forms, cohesiveness of sentence forms and long complicated sentence structures. The system could also provide a final checklist of common pitfalls to be checked against each question. These aids would not remove the poor example shown in Figure 5 but would fix the other two questions discussed.

3. Post Hoc Feedback on Item Difficulty

Question difficulty is not based on language design alone but also on the educational concepts being tested. Instructors can learn from statistical analyses of their test results what test items are particularly difficult for the students. Two mechanisms exist for this: one is the performance of an IRT-based statistical analysis, and the second is the collection of feedback from test-takers via questions about each test item. A software tool, if connected with the test giving software can support both of these efforts. The identification of particularly difficult questions might lead the instructor to focus more on a topic in lectures, to provide alternative methods for learning the topic or to generate test items at lower levels of Bloom's taxonomy, recognizing that the concept is difficult for all but the elite of the class.

CONCLUSIONS AND FUTURE WORK

This task analysis was only conducted on one instructor. It was exploratory in nature and conclusions cannot be drawn from the study about the quality and depth of multiple choice questions used in university level courses. The student feedback and the discriminant analysis provided by WebCT™ suggest that a large majority of the questions were of reasonable quality. The distribution of the questions across Bloom's taxonomy indicated that the instructor was not testing the students at higher cognitive levels of understanding. The problem questions suggested that the instructor needed guidance with wording questions. The next steps in this work are twofold: (1) A prototype multiple choice test tool is to be designed and tested on users and (2) A comprehensive study needs to be conducted on the types of test items generated with the prototype vs. the types of questions generated without the aid of the prototype.

REFERENCES

1. Adams, R. (1993). *Item Difficulty Adjustment Study: GRE Verbal Discretes* (No. GRE Board Professional Report No. 89-04P). Princeton, NJ: Educational Testing Service.
2. Anderson, L. W., & Krathwohl, D. R. (2001). *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Educational Objectives* (Complete Edition ed.). New York: Longman.
3. Enright, M. K. (1993). *A Complexity Analysis of Items from a Survey of Academic Achievement in the Life Sciences*. Princeton, NJ: Educational Testing Service.
4. Green, K. E. (1983, April 11-15, 1983). *Multiple-Choice Item Difficulty: The Effects of Language and Distracter Set Similarity.* Paper presented at the 67th Annual Meeting of the American Educational Research Association, Montreal, Quebec, CAN.
5. Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A Review of Multiple-Choice Item-Writing Guidelines for Classroom Assessment. *Applied Measurement in Education, 15*(3), 309-334.
6. Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of Item Response Theory* (Vol. 2). Newbury Park: Sage Publications.
7. IMS Global Learning Consortium, I. (2004, 2/13/2003). *IMS Learning Design Specification*. Retrieved 2/22, 2004, from http://www.imsglobal.org/learningdesign/index.cfm
8. Kashy, E., Albertelli, G., Thoennessen, M., Tsai, Y., & Kashy, D. A. (2000, October 18-20). *ALN Technology on Campus: Successes and Problems.* Paper presented at the 30th ASEE/IEEE Frontiers in Education, Kansas City, MO, USA.
9. Mayhew, D. J. (1999). *The Usability Engineering Lifecycle: A Practitioner's Handbook for User Interface Design*. San Francisco: Morgan Kaufmann.
10. McKeachie, W. J. (2002). *McKeachie's Teaching Tips: Strategies, Research, and Theory for College and University Teachers* (11th ed. ed.). Boston: Houghton Mifflin.
11. Newman, L. S., & Taube, K. T. (1996, April 8-12, 1996). *The Accuracy and Use of Item Difficulty Calibrations Estimated from Judges' Ratings of Item Difficulty.* Paper presented at the Annual Meeting of the American Educational Research Association, New York, NY, USA.
12. Prestwood, J. S., & Weiss, D. J. (1977). *Accuracy of Perceived Test-Item Difficulties* (No. 77-3): Office of Naval Research, Arlington, VA. Personnel and Training Research Programs Office.
13. Rocklin, T. (1999). *Virtual Companion to article Computers and Testing in The National Teaching and Learning Forum, Volume 8 Number 5.* Retrieved 2/22, 2004, from http://www.ntlf.com/html/sf/vc85.htm
14. Scheuneman, J. (1991). *Effects of Prose Complexity on Achievement Test Item Difficulty*. Princeton, NJ: Educational Testing Service.
15. Stiggins, R. J. (2001). The Unfulfilled Promise of Classroom Assessment. *Educational Measurement: Issues and Practice, 20*(3), 5-15.