

To the Graduate Council:

I am submitting herewith a dissertation written by Monica Anne Wallace entitled "Multiple Choice Exams: Student Explanations for Answer Choices." I have examined the final paper copy of this dissertation for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, with a major in Education.



Robert L. Williams, Major Professor

We have read this dissertation
and recommend its acceptance:



Gary Klucken

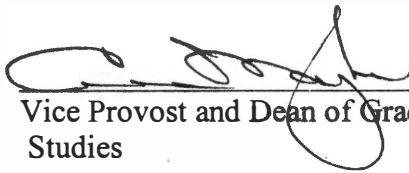


R. Steve McCallum



Sandra Twardosz

Accepted for the Council:



Vice Provost and Dean of Graduate
Studies

Thesis
2003b
.W26

**MULTIPLE-CHOICE EXAMS: STUDENT EXPLANATIONS FOR ANSWER
CHOICES**

**A Dissertation Presented for the Doctor of Philosophy Degree
The University of Tennessee, Knoxville**

Monica Anne Wallace

August 2003

ACKNOWLEDGEMENTS

Thank you to my Committee Chair, Dr. Robert Williams, for demonstrating and expecting quality and persistence. His contributions to my graduate education extend well beyond my dissertation research.

Thank you to my Committee, Dr. Gary Klukken, Dr. Steve McCallum, and Dr. Sandra Twardosz for completing their duties with interest and care.

Thank you to the following EP 210 graduate teaching assistants who contributed to my research: Carrie Booher, Lloyd Elliot Clark, Susan Stockdale, Chris Neddenreip, Eden Abramson, Renee Oliver, Beth Winn, and Jessica Allin.

ABSTRACT

This study examined relationships between three variable domains: (a) students' explanations for their answers to difficult multiple-choice items, (b) generic critical thinking, and (c) performance on different types of exam items. Students took a critical thinking test at the beginning of a large human development course and 5 multiple-choice exams during the course. Prior to turning in each exam, students prepared written explanations for their answers to the 4 items judged most difficult by the instructors. Written explanations were reviewed for accuracy of responses, number of options considered, and the presence of faulty thinking strategies.

Specific questions addressed by the study included the following: (a) What is the correlation between explanatory dimensions (i.e., accuracy, options, error) and different types of exam items (i.e., total, recall, comprehension, mixed, and target)? (b) What is the correlation between base-level critical thinking and the explanatory dimensions? (c) To what degree do the explanatory dimensions predict exam performance and critical thinking scores? (d) How well does critical thinking as a singular predictor account for scores on item types? and (e) What differences in the explanatory dimension, critical thinking, and item-type scores exist between high-, average-, and low-performing students on the exams?

Accuracy and options significantly and moderately correlated with all item types. The error dimension significantly, but weakly, correlated with all item types except the recall items. Critical thinking significantly and weakly correlated with accuracy and options, but did not correlate significantly with the error dimension. All explanatory dimensions significantly predicted exam performance measures, but the number of

options explained proved to be the strongest predictor of exam measures. Options also was the only explanatory dimension that significantly predicted critical thinking. Critical thinking as a singular predictor accounted for a small amount of the variance in all item types, but accounted for the highest percentage of variance in the comprehension and total items. Significant differences existed among all groups with respect to all dependent measures except for the error dimension between the low- and average-performing groups.

TABLE OF CONTENTS

<u>CHAPTER</u>	<u>PAGE</u>
CHAPTER ONE	
INTRODUCTION	1
What Multiple-Choice Exams Measure	2
Student Reasoning and Performance on Multiple-Choice Exams	3
Critical Thinking and Exam Performance	4
Purpose of the Study	6
CHAPTER TWO	
METHOD	7
Participants	7
Description of Psychoeducational Studies (PES) 210	7
Instructional Format	9
Exam Development	9
Description of Unit Exams	10
Assessment of Entry-Level Critical Thinking	11
Assessment of Student Explanations for Item Responses	11
Item Selection for Written Analysis	11
Analysis of Written Explanations	12
CHAPTER THREE	
RESULTS	15
Correlational Analysis	15
Regression Analyses	15
Group Comparisons	16

CHAPTER FOUR

DISCUSSION	21
The Linkage of Critical Thinking to Explanatory Dimensions and Exam Performance	21
Explanatory Dimensions as Predictors of Exam Performance	23
Explanatory and Item-Type Group Differences	24
Application of Study Findings	25
Limitations of the Study	26
Unanswered Questions	27
LIST OF REFERENCES	28
APPENDICES	33
Appendix A	34
Appendix B	35
Appendix C	38
VITA	45

LIST OF TABLES

	<u>PAGE</u>
Table 3.1 Correlations Between Explanatory Dimensions and Exam/ Critical Thinking Performance Measures	16
Table 3.2 Significant Explanatory Predictors for Exam Item Types	17
Table 3.3 High-, Average-, and Low-Performance Group Comparisons for Raw Score Means and Percentage of Maximum Scores for Explanatory Categories, Item Types, and Critical Thinking	19

CHAPTER ONE

INTRODUCTION

Multiple-choice tests constitute efficient methods of assessing performance in large undergraduate educational psychology courses. Although these exams typically yield a wide range of student performance, some students argue that this variation is more a function of the method of assessment than their mastery of content. Some students claim that they can do well on multiple-choice exams even when their course knowledge is marginal. Other students claim to know the information quite well and yet do poorly on the exams. Their explanations regarding poor performance often include such comments as “I didn’t know what the question was asking,” “the items were tricky,” and “the items were poorly worded.” An alternative explanation is that performance on multiple-choice exams primarily relates to students’ mastery of information and reasoning skills.

An understanding of high-, average-, and low-performing students’ knowledge and reasoning on multiple-choice exams could impact instructional management: particularly, the delivery of information and assessment of student mastery. For example, providing practice questions that require higher order thinking and course knowledge may help students utilize both knowledge and effective reasoning on the exams. Learning how reasoning influences performance on multiple-choice exams may be a step towards learning how to enhance the critical thinking skills of course participants, a goal of various commissions and college educators (Bowen, 1977; Jones, 1995; National Center for Education Statistics, 1991).

Because student knowledge and reasoning were examined in the context of responses to multiple-choice exams, a review of literature on multiple-choice exams is

fundamental to this study. More specifically, information regarding complex multiple-choice item development, the nature of research related to student performance on multiple-choice exams, and a recent study linking multiple-choice exam performance to base-level critical thinking were reviewed.

What Multiple-Choice Exams Measure

A trend towards performance-based assessment has resulted in criticism of more traditional methods such as multiple-choice tests (Killoran, 1992). The creation of complex items can be time consuming, tedious, and often a matter of trial and error. Item-development guidelines generally focus on issues of style and grammar (Kehoe, 1995), offering little information on how to construct complex multiple-choice items. Bloom's Taxonomy, or some form of it, often guides test construction (Webb, 1993); yet failure to clearly operationalize the levels and delineate the relationship between the various levels are among difficulties cited in the use of Bloom's Taxonomy in item development (Ennis, 1993; Paul, 1984). Further, there is negligible support for the few models of item development mentioned in the literature (e.g., Aiken, 1982; Karras, 1991; Killoran, 1992).

Although multiple-choice tests often fail in the area of testing students' higher order thinking skills (Karras, 1985; Killoran, 1992), some reports (Bowering & Wetmore, 1997; Fuhrman, 1996; Karras, 1978; Martinez, 1999; Morrison & Free, 2001) acknowledge that it is both plausible and desirable to construct items that require cognitive manipulation of information. In regard to student learning, well constructed multiple-choice items would seem to require a precise understanding of material, which may not be demanded in such open-ended assessments as essay exams. To answer

complex items “correctly,” one must possess accurate factual information and the ability to cognitively manipulate that information. The requirement to think about facts in a novel context or combine facts to reach a conclusion can be reflected in multiple-choice items, as well as the demand to demonstrate knowledge of factual information. Karras (1978) refers to these items as ones that “get it all together” (p. 212).

Student Reasoning and Performance on Multiple-Choice Exams

Reasoning, as do a number of educationally relevant constructs (e.g., intelligence, critical thinking), has multiple definitions. Abstract reasoning, a hallmark of Piaget’s formal operational stage of cognitive development (Rothstein, 1997), general reasoning or intelligence (“g”) as defined by Charles Spearman (Sattler, 1992), and such sub-categories of reasoning as deductive and inductive reasoning are examples of cognitive dimensions subsumed under the construct of reasoning. The definition of reasoning utilized in this study (the explanation a student gives for answer choices on multiple-choice exams) parallels the dictionary definition of reason as “a statement offered in explanation or justification” (Merriam Webster Collegiate Dictionary, 1996, p. 974).

An extensive review of educational literature has revealed no studies linking reasoning, as defined in this study, and student performance on multiple-choice exams. Much of the literature on multiple-choice exams (e.g., Bowering, 1997; McClain, 1983; Scruggs & Mastropieri, 1995) has addressed test-taking strategies (e.g., skipping difficult items, answering an item prior to reading options). However, the definition of reasoning used in this study does not encompass test-taking strategies. Another related line of research involved having students report aloud their thinking/reasoning process as they respond to multiple-choice items (e.g., Bloom & Broder, 1950; Connolly & Wantman,

1964; Kropp, 1956; McGuire, 1963, Norris, 1990). For example, Norris (1990) had participants report aloud what they were thinking as they responded to multiple-choice items. These verbal reports (analogous to the written explanations in the current study) were then analyzed for evidence that critical thinking, the construct being measured, was required to answer the questions. However, the primary purpose of these verbal report studies was to investigate item and construct validity rather than differences in reasoning.

Critical Thinking and Exam Performance

The examinations in the course (PES 210) targeted in this study have a reputation among students for being challenging at best and unfair and tricky at worst. Graduate teaching assistants (GTAs) and the instructor who have worked with the course can relate tales of face-to-face encounters with students who claim a 4.0 GPA being “ruined” by PES 210. Additionally, some students have complained that course procedures exist to foster research goals rather than their academic interests. Some students have also objected to the missed exam items appeals process that requires individual documentation of course information to support their selection of the non-keyed response (as compared to an in-class group rebuttal).

Although anecdotal in nature, the student responses described above reflect resistance to critical thinking. Keely, Shemberg, Cowell, and Zinnabuer (1995) suggested that resistance to critical thinking among college students is analogous to resistance seen among persons in psychotherapy. Psychotherapy clients want answers to their problems, and students want teachers to tell them the one right answer. The motivation and effort required to make personal change is parallel to the effort required to participate as critical thinkers in class discussions and in taking exams. To lessen resistance to classroom

activities that require critical thinking, Keeley et al. suggested ideas commonly offered to improve a therapeutic relationship: (a) communicate expertness and credibility, (b) demonstrate warmth, genuineness, and empathy, (c) create a sense of safety and trust, (d) assess and mold student expectations, (e) establish limits--avoid enabling, and (f) note the resistance and invite students to explore reasons for it.

A recent study (Worth, 2000) has linked base-level critical thinking to performance on multiple-choice exams. This study compared the predictive potential of two well-established predictors of performance (attendance and note taking) to that of base-level critical thinking on course performance. At the second class meeting of a highly structured human development course, students took the California Critical Thinking Skills Test (CCTST), a standardized measure that targets one's "ability to properly construct and evaluate arguments" (Facione, 1986, p. 222). Scores on the CCTST were correlated with a number of course outcome variables including scores on multiple-choice exams. Differences in critical thinking as measured by the CCTST accounted for 29% of the variance in exam performance, while note taking and attendance on non-credit days, both established predictors of student performance, accounted for an additional 20% and 2% of the variance, respectively.

It appears that critical thinking is consistent with "good" reasoning on multiple-choice exams. Worth (2000) observed that "critical-thinking skills probably helped students determine (a) what information was needed to respond to a particular test item and (b) what conclusion would follow from that information" (p. 22). Yet unknown are the specific components of exam reasoning that differentiate high-, average-, and low-performing students, as well as the relationship between student's reasoning on exam

items and their base-level critical thinking.

Purpose of the Study

To summarize, a basic premise of this study is that multiple-choice items can be written that require higher thought processes. Further, it is hypothesized that a relationship exists between how students explain their answer choices (i.e., reasoning) and their exam performance. To investigate this proposed relationship, students were asked to explain in writing their answer choices for selected items. These written explanations were subsequently analyzed for accuracy, completeness, and reasoning characteristics.

Specific questions addressed by this study included the following: (a) What is the correlation between explanatory dimensions (i.e., accuracy, options, and error) and different types of exam items (i.e., total, recall, comprehension, mixed, and target)? (b) What is the correlation between base-level critical thinking and the explanatory dimensions? (c) To what degree do the explanatory dimensions predict exam performance and critical thinking scores? (d) How well does critical thinking as a singular predictor account for scores on item types? and (e) What differences in the explanatory dimension, critical thinking, and item-type scores exist between high-, average-, and low- performing students on the exams?

CHAPTER TWO

METHOD

Participants

Students in two sections of an undergraduate human development course ($N = 157$) participated in the study. Participation was voluntary, with a small amount of course credit given to those participating in research activities. Alternative activities for credit were made available for those who chose not to participate. One hundred percent of students participated in at least one facet of the study. The sample consisted of 69 sophomores, 47 juniors, 27 seniors and 10 graduate students. Academic classification was not available for 4 students. More women ($n = 124$) than men ($n = 33$) participated.

Description of Psychoeducational Studies (PES) 210

PES 210 is a prerequisite course for entry into the teacher education program at The University of Tennessee, Knoxville. The course is organized around five units: physical, cognitive, psychological, social, and character development. The course is designed to facilitate use of credible information in reaching conclusions related to educational trends and issues. Examples of specific issues covered include benchmark achievements in physical development, teacher promotion of language skills, role of causal attributions in dealing with success and failure experiences, and the philosophy and strategies of character education.

Course reading materials came from the text Developmental Issues in Teaching by Rothstein (1997) and a set of journal articles in a reading packet developed for the course. A study guide contained questions that highlighted the most important concepts from the readings and instructor presentations. The readings study questions were not

specifically reviewed during the instructor presentations (unless students asked about a particular question), although information from the reading material often was reflected in the content discussed in class.

Graded activities included five 40-item multiple-choice exams, 5 essay quizzes over readings typically not covered in class, one 75-item multiple-choice comprehensive final, and one research paper. For the last two units in the course a special cooperative learning group contingency was available where bonus points could be earned if the group exam average met a certain criterion for improvement. Course grades were assigned on a criterion-referenced basis, with a certain number of points corresponding to a particular letter grade.

Each unit followed the same instructional sequence: videotape introduction, three days of instructor presentations, short-answer essay quiz over assigned readings, 40-item unit exam, and feedback regarding the exam. Students had the option of having their exams scored in class on the day of the exam and reviewing missed items prior to leaving class. Students also were provided the opportunity to appeal missed items in writing. Documentation from course readings and/or instructor presentation notes that supported the incorrect answer choice was a required component of appeals. If the appeal presented justification for the incorrect choice as viable as the justification offered by the instructors for the keyed response, credit was given for the incorrect response.

A course website provided hand-outs, announcements, and clarification of concepts discussed in class. Individual appointments with GTAs, videotapes of instructor presentations, and supplementary videotapes related to course content were additional resources available to support student learning.

Instructional Format

Six graduate teaching assistants (GTAs) and one instructor were responsible for course delivery. Three of the six GTAs working with the course participated in the course the previous year (Fall 1998, Spring 1999), with the primary responsibilities of grading student written work and learning course content. The opportunity to observe the supervising instructor teach during the 9:40 course section (Fall 1999) offered an additional opportunity to master course content. Occasional supervising instructor observation of GTAs, as well as discussions among GTAs and the supervising instructor, also provided ideas for improving instructional effectiveness. Supplementary videotapes, videotapes of the supervising instructor presentations, and readings that supported the course content also were available for teaching assistant review.

Each GTA was provided with notes, overhead transparencies and handouts that followed the student study guide. The student study guide consisted of 435 questions that outlined the most important concepts from lectures, readings, and videotapes for each of the five units. Each GTA took each unit exam, providing an additional opportunity to measure understanding and clarify concepts when missed items were discussed at weekly planning meetings.

For each of the five units in the 11:10 course section, one second-year GTA was the primary instructor. Two additional second-year GTAs and three first-year GTAs were available to support instruction. For example, “non-teaching” GTAs participated in discussions, assisted in answering student questions, or reminded the teaching GTA if important information was omitted.

Exam Development

The five unit exams were created by the supervising instructor; however, GTAs and the instructor participated in ongoing revision of items. The most-missed items from

the prior year's exams were reviewed and adapted when areas of ambiguity or inconsistency were noted by GTAs and the supervising instructor. Instructor and GTA copies of exams included page numbers from instructor presentation notes and readings corresponding to the concepts being tested by each item, thus constituting evidence of content validity.

Description of Unit Exams

Each item on the five 40-item unit exams was a multiple-choice question with four options. Items were classified as either direct recall or comprehension. Direct recall items assessed memory of facts, with the context of the question stem and options closely paralleling information presented in class or in the course readings. Comprehension items met some combination of the following guidelines: (a) the use of facts in a context different from the context presented in class or course readings; (b) the development and evaluation of an argument based on information explicitly stated or inferred from information given; and (c) a comparison, contrast, or transformation of several pieces of information in order to make a decision.

The six GTAs working with the course participated in the exam item classification process. Each rater was provided with an item-type definition sheet, a copy of each exam and an item-classification rating sheet. An item type was assigned if at least 4 of the 6 raters agreed on the classification category. Throughout the rating process, GTAs provided rationale for their classification choices during weekly planning meetings. Ratings were not altered at this time; however, these discussions provided guidance and training for making future ratings. Across the five unit exams, raters judged 26% of the items as direct recall and 53% as comprehension, with an average inter-rater agreement of 73%. Items receiving less than 66% agreement across raters were designated as mixed items (21% of the items). An example of recall and comprehension item types can be found in Appendix A.

Assessment of Entry-Level Critical Thinking

On the second class meeting of the semester, students in each of the two course sections participating in the study took the California Critical Thinking Skills Test (CCTST), a composite measure of argument construction and evaluation (Facione & Facione, 1994). This instrument contains 34 multiple-choice items. Critical thinking metrics for the current sample (CS) indicated a more homogeneous distribution and slightly higher central tendency measures than those reported for the standardization sample (SS): SS mean = 15.89 and CS mean = 16.88, SS median = 16 and CS median = 17, SS standard deviation = 4.46 and CS standard deviation = 3.94, and SS range of scores = 2 to 29 and CS range of scores = 8 to 27. Kuder-Richardson 20 reliability coefficients were between .68 and .69 in the standardization sample and .68 in a prior PES 210 course (Worth, 2000). Significant correlations between scores on the CCTST and SAT-verbal ($r = .55$), SAT-math ($r = .44$), and college GPA ($r = .20$) were found during test-validation research (Facione & Facione, 1994), thus providing evidence of predictive validity.

Assessment of Student Explanations for Item Responses

All students in the two course sections were asked to explain in writing their answer choices on four selected exam items prior to turning in their unit exam for grading. The selected items were listed on an attachment handed out with the exam and no further instructions were given regarding the level of detail expected in the explanation. Students' written explanations were subsequently assessed for knowledge and reasoning.

Item selection for written analysis. For each of the five 40-item multiple-choice unit exams, the six GTAs working with the course initially identified the six items they thought most difficult. A frequency count of these individual ratings yielded a composite rating of the six most difficult items. These six items subsequently were reviewed at a

weekly staff meeting and reduced to four.

Specific criteria utilized in reaching consensus regarding the four most difficult items are represented by the following questions: (a) Does the item require more than one step to answer? (b) Does the item require use of factual knowledge in a context different from that presented in class or course readings? and (c) Does the item require development and evaluation of an argument based on more than one piece of information? Four items made the final cut for each exam, yielding a total of 20 items across five exams. Sixteen of the selected items were classified as comprehension items, and four of the selected items were classified as mixed items.

Analysis of written explanations. One graduate student rated all written responses and two other graduate students rated 13% of the written responses. Raters were initially trained to rate written explanations during a two-hour workshop. Study purpose and methodology were reviewed, as well as rating procedures. During the initial training session raters were shown examples of items that had been rated and then provided with written explanations to rate as practice. Individual ratings were then discussed as a group. Raters were provided with additional written explanations to rate at home for practice. Once scored by the raters, these practice explanations were reviewed on an individual basis with the primary researcher. When the raters and the primary researcher achieved approximately an 80% level of agreement, a different set of written explanations were randomly assigned for the purpose of establishing inter-rater reliability. When applicable, points of disagreement between the primary researcher and raters were reviewed with individual raters, providing on-going training and clarification of procedures if necessary. However, a discussion of disagreement did not change the score assigned. Raters were first year GTAs working with Educational Psychology 210 (formerly PES 210).

Each written explanation was rated for accuracy and number of options considered. Where applicable, raters also computed a faulty-thinking or error score for

student explanations. A “Written Analysis Rating Scale” (see Appendix B) provided scoring guidelines. For example, criteria were included for the assignment of a 0, .5, or 1-point accuracy rating. Definitions for each of the three rating dimensions and levels of inter-rater agreement follow.

The accuracy rating represented the level of accurate and relevant information included in the written explanation. A composite rating from 0 to 4 (up to 1 point for each of the 4 options) was assigned for each item. For each of the 20 items rated, an information summary sheet included a copy of the item and for each option (A-D) factual information that either supported or refuted the option as the correct choice. Any underlined information on the fact sheet was considered crucial and, therefore, required to obtain an accuracy rating of 1. Inter-rater agreement for accuracy across the five course units was 81%, with a range of 73 to 94%.

The options rating signified how many options student’s ruled in or out for each item, with the ratings ranging from 0 to 4. It was not necessary for the student to identify by letter (i.e., A, B, C, D) the option being ruled in or out to receive credit toward the options score. Raters were instructed to infer that an option was considered based on a match between information included in the explanation and the option. Inter-rater agreement for options across the five course units was 91%, with a range of 72 to 98%.

Raters assigned an error score when an explanation for an incorrect answer reflected one or more of five faulty thinking strategies: misinterpretation, mismatch, bias, inconsistent standards, and naiveté. The research team developed the following definitions for the error categories: misinterpretation--altering the meaning of the stem or an option and/or ignoring words or phrases crucial to understanding the item; mismatch--invoking accurate but irrelevant information; bias--relying on personal experiences, values, or opinions as the basis for an answer; inconsistent standards--varying criteria for supporting or rejecting options within the same item; and naiveté--not knowing a

common word or common information crucial to answering a question.

If the primary rater had identified an error(s) in the explanation, she highlighted the error category column on the inter-rater rating form to cue the two other raters of the presence of an error in the explanation. It was possible to have more than one error type in a singular explanation, and each incidence of a faulty thinking strategy was given a score of one point. Inter-rater agreement for error across the five course units was 81%, with a range of 69 to 100%. Examples of student written explanations that contain an error, the corresponding information summary sheet, and accuracy and options ratings can be found in Appendix C.

CHAPTER THREE

RESULTS

Correlational analyses, regression analyses and group comparisons were computed for explanatory dimension scores. Each explanatory dimension was correlated with each item type and critical thinking test score. Each explanatory dimension also was used as a predictor variable in regression analyses for item type and total exam scores. Critical thinking was used as a singular predictor for item type and total exam scores. High-, average- and low-performing groups were then compared on explanatory dimensions, item types, and critical thinking. No significant male/female differences were obtained for any of the variables in the study, consistent with the overall findings of a relatively recent meta-analysis by Linn and Hyde (1991) that explored gender differences in psychology.

Correlational Analysis

Table 3.1 shows that both accuracy and options ratings for the four target items (20 across exams) correlated substantially, positively, and significantly ($p < .01$) with all item types on the exams. Accuracy and options also correlated significantly with critical thinking. The combined error rating correlated negatively and significantly with all item types, but not significantly with critical thinking. Of the three explanatory dimensions, options and accuracy correlated the highest with item types and critical thinking.

Regression Analyses

Including accuracy, options, and error scores in a step-wise regression analysis for all item types and total exam grade showed that options was the primary predictor for all item types and total exam performance (see Table 3.2). Accuracy also contributed to prediction of all item types and overall exam performance. Error contributed to prediction of comprehension, target, and mixed item types. With respect to scores on the California Critical Thinking Skills Test, options proved to be the only significant predictor,

Table 3.1

Correlations Between Explanatory Dimensions and Exam/Critical Thinking Performance Measures

Independent variables	Performance Measures				
	Total items	Recall items	Comp ^a items	Mixed items	Critical thinking
Accuracy	.54**	.50**	.57**	.50**	.21*
Options	.63**	.60*	.62**	.58**	.33**
Error	-.22*	.20*	-.33**	-.29*	NS

Note. The N for all variables except critical thinking (n = 150) was 157.

* $p < .05$. ** $p < .01$.

^aComp = comprehension

accounting for 12% of the variance. Critical thinking as a singular predictor accounted for a small but significant amount of the variance in all item types. Critical thinking accounted for 15% of the variance in recall items, 14% of the variance in mixed items, and 6% of the variance in target items. Critical thinking accounted for a higher percentage of the variance (i.e., 17%) in both the comprehension and total items.

Group Comparisons

A criterion-referenced grading system indicated that 14% of the students earned As or B+s (high performing) on the exams, 30% made Cs (average performing), and 15% made Ds or Fs (low performing). The following standards differentiated these grade levels: 90% and above correct = A, 88-89% correct = B+, 70-79% correct = C, 60-69% =

Table 3.2

Significant Explanatory Predictors for Exam Item Types

<u>Item type</u>	<u>Significant predictors</u>	<u>Explained variance</u>
Recall	Options	36%
	Options, accuracy	39%
Comprehension	Options	39%
	Options, error	46%
	Options, error, accuracy	51%
Mixed	Options	34%
	Options, error	39%
	Options, error, accuracy	42%
Target	Options	39%
	Options, error	51%
	Options, error, accuracy	56%
Total	Options	40%
	Options, accuracy	44%
	Options, accuracy, error	46%

D, and below 60% correct = F.

The General Linear Model multivariate analysis function in SPSS 10.0 was utilized to compare the different performance groups across explanatory dimensions, item types, and critical thinking (see Table 3.3). After multivariate tests (i.e., Pillai's Trace, Wilks' Lambda, Hotelling's Trace, Roy's Largest Root) showed significant differences among the groups, univariate analyses were conducted and yielded significant F statistics for each of the dependent variables. Post hoc pairwise comparisons utilizing the Tukey Honestly Significant Difference test (Tukey HSD) showed differences among specific means. All comparisons were significant, except error for the average- and low-performing groups. Specific results for each pairwise comparison follow.

High-performing students obtained significantly different ratings from low-performing students on accuracy ($p < .001$), error ($p < .01$), choices ($p < .001$), all item types ($p < .001$), and critical thinking ($p < .001$). The accuracy rating mean was 2.5 times higher and the options rating mean was 1.5 times higher for high performers than for the low performers. The error rating mean was almost 2 times higher for low performers than for the high performers. On average, high performers scored 1.5 times higher on recall and comprehension and 1.7 times higher on target items than low performers.

A comparison of high performers to average performers demonstrated that high-performing students also obtained significantly different ($p < .001$) ratings from average-performing students on all explanatory dimensions. The mean accuracy rating was 1.4 times higher, and the mean options rating was 1.3 times higher for high performers than for average performers. The error rating mean was almost 2 times higher for average performers than for high performers. On average, high performers scored 1.2 times higher on recall and comprehension items and 1.3 times higher on target items than did the average performers. High-performers also obtained significantly higher ($p < .05$)

Table 3.3

High-, Average-, and Low-Performance Group Comparisons for Raw Score Means and Percentage of Maximum Scores for Explanatory Categories, Item Types, and Critical Thinking

Comparison Variable	Group Comparisons		
	High mean	Average mean	Low mean
Explanation			
Accuracy	37.06 (46%) ^a	26.24 (33%)	14.83 (19%)
Option	54.77 (68%)	42.14 (53%)	35.20 (44%)
Error	2.36 ^b	4.44 ^b	4.45 ^b
Item Type			
Target ^c	16.13 (81%)	12.17 (61%)	9.50 (48%)
Recall	47.68 (92%)	39.85 (77%)	34.00 (66%)
Comp	95.91 (90%)	78.87 (74%)	65.37 (62%)
Mixed	38.59 (92%)	32.10 (76%)	27.04 (64%)
Critical Thinking	19.38 (57%)	16.63 (49%)	13.33 (39%)

Note. All group differences were significant except error for the average- and low-performing groups.

^aNumbers in parentheses = percentage of maximum score represented by raw score means. ^bThese numbers represent the average number of reasoning errors across the 20 target items. ^cTarget = score for the 20 difficult items targeted for written analysis.

critical thinking scores than did average performers.

A comparison of average-performing students to low-performing students demonstrated that average-performing students obtained significantly different mean ratings from low-performing students on accuracy ($p < .001$) and options ratings ($p < .01$). The accuracy rating mean was almost 2 times higher, and the options rating mean was 1.2 times higher for average performers than for low performers. There were no significant differences for error ratings between the two groups. Average performers scored 1.2 times higher on recall and comprehension items and 1.3 times higher on target items than low performers. The average performers also obtained significantly ($p < .01$) higher critical thinking scores than did the low performers.

The most frequent reasoning error for all groups was misinterpretation of items (57% for high performers, 62% for average performers, and 59% for low performers). For high performers, mismatch accounted for 33% of errors as compared to 22% for average performers and 25% for low performers. Ten percent of the errors made by low performers and 7% of the errors made by average performers were bias, whereas only 1% of high performers made this type of error. The percentage of inconsistent standards errors made was similar for all groups. The percentage of naiveté errors made by the average performers (3%) was slightly higher than those made by the other two groups (< 1%).

CHAPTER FOUR

DISCUSSION

Overall, the findings indicated that accuracy and options significantly and moderately correlated with all item types. The error dimension significantly, but weakly, correlated with all item types, except the recall items. Critical thinking significantly and weakly correlated with accuracy and options, but did not correlate significantly with the error dimension. All explanatory dimensions significantly predicted exam performance measures, but the number of options explained proved to be the strongest predictor of exam measures. Options also was the only explanatory dimension that significantly predicted critical thinking. Critical thinking as a singular predictor accounted for a small but significant amount of the variance in all item types, but accounted for the highest percentage variance in the comprehension and total items. Significant differences existed among all groups with respect to all dependent measures, except for the error dimension between the low- and average-performing groups.

This chapter focuses on three aspects of the study's findings: (a) how critical thinking as defined by the CCTST correlated with the explanatory dimensions and exam performance, (b) the importance of both reasoning skills and accurate information as illustrated by the explanatory dimensions of options and accuracy in accounting for exam performance, and (c) the differences in performance groups with respect to item-type scores and number and type of reasoning errors. Applications of study findings and questions for future research also are addressed.

The Linkage of Critical Thinking to Explanatory Dimensions and Exam Performance

Although critical thinking seems a necessary component of successful exam performance, the correlations between the explanatory dimensions and critical thinking test scores were mixed. The weak relationship observed between accuracy and critical thinking ($r = .21$) is not surprising, given that the CCTST does not require retrieval of

factual information as do the course examinations. The CCTST provides all the information needed to answer the questions. A somewhat stronger, although moderate correlation, existed between critical thinking and options ($r = .33$). This finding is congruent with the purpose of the CCTST, in that a careful consideration of all pieces of information provided in the question is fundamental to answering the question.

Somewhat surprising, was the non-significant relationship between faulty thinking as represented by the five error categories in this study and critical thinking. It would seem that misinterpretation, bias, inconsistent standards, mismatch, and naiveté would be inconsistent with effective evaluation and construction of arguments. The definition of critical thinking used by the CCTST may have precluded a relationship between critical thinking and the error categories. The CCTST test manual reported no factor analysis of sub-components of critical thinking, thus precluding a more direct comparison of what is being measured by the CCTST and the error dimension in this study. Thus, critical thinking as measured by the CCTST seems to reflect a more broadly defined, and qualitatively different global construct than the specific thinking strategies denoted by the error dimension used in this study.

The extent to which the CCTST predicted exam scores was weaker for the current sample (explaining 17% of the variance in exam scores) than for Worth's (2000) sample. When used as the singular predictor of exam performance, the CCTST explained 21% of the variance in exam scores in Worth's study. The regression comparisons are consistent with a marginally higher correlation between critical thinking and exam scores in Worth's study ($r = .46$ in Worth's study as compared to $r = .41$ in the current study). Congruent with this higher correlation is a larger range of scores and larger standard deviation for both the critical thinking test and the exam scores in Worth's study than in the current study. Range of critical thinking scores in the Worth study was 5 to 33, whereas range of critical thinking scores in the current study was 8 to 27. Range of exam

scores in the Worth study was 0-189, whereas range of exam scores in the current study was 98-191. Standard deviation for critical thinking scores in the Worth study was 4.31, whereas standard deviation for the critical thinking scores in the current study was 3.94. Standard deviation for exam scores in the Worth study was 23.76, whereas standard deviation for exam scores in the current study was 18.88.

Explanatory Dimensions as Predictors of Exam Performance

The explanatory dimensions proved to be viable predictors for scores on all question types as well as exam total, with the strongest performance predictor being the number of options considered. The options variable also correlated most highly with critical thinking. Accounting for the correctness or incorrectness of each option may reflect good thinking as well as good effort. Additionally, results support the test-taking strategy of carefully considering each option when taking a multiple-choice exam, which is consistent with earlier findings by McClain (1983). As expected, error offered no significant explanation in the model generated for recall items, whereas error was part of the model for comprehension and target items. Unexpected was the failure of accuracy to be a more viable predictor for recall items than for the other item types, based on the item-classification definitions. One explanation for this finding may be that students knew enough to rule in/out an option on the exam (receiving credit for considering the option) without having the specific information required to receive a maximum accuracy rating.

The same prediction model was generated for target items (rated as the most difficult items on the exam) as for comprehension and mixed items; however, the amount of total variance explained was about 5% higher (56%) for target items. The biggest difference was that error explained an additional 12% of the variance for the target items as compared to around 5% for comprehension and mixed items. Thus, for target items the model placed a premium on reasoning characterized by careful analysis of each option, as

well as clear and thoughtful application of accurate information.

Explanatory and Item-Type Group Differences

The comparison of high-performing (those earning As or B+s) to low-performing (those earning Ds or Fs) and average-performing groups (those earning Cs) yielded significant differences on all explanatory dimensions. The comparison of the average-performing and low-performing (those earning Ds or Fs) groups produced significant differences only for the options and accuracy dimensions.

Although the primary reasoning error made by all groups was misinterpretation, some differences in the type of errors made did emerge. Low performers, as indicated by a higher percentage of bias errors, allowed personal views and feelings to impact answer selection a greater proportion of the time than did the high performers. High performers, as indicated by a higher rate of mismatch errors and accuracy scores, knew more facts than lower performers even though the information was sometimes misapplied. Also suggested by this finding is that accuracy is a necessary, but not a sufficient contributor to successful exam performance.

Low-performing students missed about the same percentage of questions irrespective of the item type (i.e., recall = 34%, comprehension = 38%, mixed = 36%). This finding challenges a claim often heard from low-performing students: “I knew the information really well but the exam was tricky.” In reality, these students did as poorly on direct-recall items as on the comprehension items. These findings show that both insufficient information and reasoning errors plagued the performance of low-performing students. Whenever the highest performing group was compared to a lower-performing group, the biggest difference in item type scores was always found within the target-item category. Thus, the most difficult items apparently required not only accurate information, but also higher order thinking skills to answer correctly. These comparisons of high-, average-, and low-performing students point to multiple cognitive differences

between the groups. Lower-performing students use less accurate information, analyze fewer options, and make more reasoning errors than high-performing students.

Application of Study Findings

First of all, the findings of this study empirically confirm the efficacy of commonly recommended guidelines for improving exam performance. For example, two common guidelines are to read items slowly and carefully (our findings indicated that item misinterpretation was the number one reasoning error) and make sure the information used matches the question or option (our findings identified mismatch of information as the second most common reasoning error).

A primary strategy for helping students better understand and improve their performance is to provide samples of student explanations for incorrect choices. When shown such explanations in a class workshop, a number of students identified the specific faulty thinking strategies used and pinpointed examples that illustrated their own thinking errors. One student reported that this analysis of faulty thinking helped her achieve her best score ever on a multiple-choice test the following day in another course.

Review of the written explanations in this study has also provided valuable input to instructors in refining both exam items and the instruction related to those items. Thus, having students write explanations for selected exam items could produce multiple benefits: provide a basis for instructor and students to examine information and reasoning used in responding to items; pinpoint items that need to be clarified (many students may have misinterpreted the same item); and highlight the need for better instruction related to an item (the item may be well designed but course content may have omitted some information crucial to the item). Another result of requiring written explanations for difficult items is that students appear more likely to acknowledge the legitimacy of such items and exert greater effort to improve their exam performance.

To this point, attempts to improve exam performance have been more effective at

the upper end of the grade scale than at the lower end. For example, 21% of students this past semester earned As and B+s (compared to 14% in this study), but the percentage earning Ds and Fs remained unchanged. The critical thinking deficiencies of the low performers—their mean fell in the bottom quartile of the standardization distribution—may impede their application of our explanatory dimensions to improve exam performance. Students with low critical thinking may need individualized assistance to understand and apply the explanatory dimensions. Having low performers submit written explanations for their answers to difficult exam items and then examining their rationales with them on a one-to-one basis would likely be more effective than simply providing sample explanations in a group setting.

Limitations of the Study

Although exam-item classification was not crucial to the central purpose of the study, levels of agreement achieved in classifying items were marginally acceptable. In fact, some items received such low levels of agreement they were classified as mixed items. However, it should be noted that raters were successful at picking the most difficult items used for written analysis. As indicated by a frequency count, these items were the most-missed items on each exam.

Another limitation of the inter-rater reliability was directly related to the purpose of this study: the inter-rater agreements for the explanatory dimensions were quite variable across raters. Although the average level of agreement for each of the explanatory dimensions was greater than 80%, some of the individual percentages used in computing the average were, in fact, quite low as compared to those typically reported in the behavioral literature (Kazdin, 2001).

The task of reaching agreement regarding qualitative cognitive dimensions appears inherently more difficult than reaching agreement regarding discrete behavioral events. Fuhrman (1996) and Martinez (1999) have noted that the knowledge and

reasoning skills of the respondent influence the perceived cognitive demands of an exam item. For example, as illustrated in the mixed-item classification category in the current study, an item that primarily requires recall for one respondent may involve higher order reasoning for another.

Unanswered Questions

Questions prompted or left unanswered by the current study include the following questions: What is the best to teach students to apply these findings? Is it enough to include suggestions in the syllabus for doing well in the course? Would supplemental workshops explaining how to apply the findings of this study contribute to effective reasoning on multiple-choice exams? Can teaching assistants be trained to help students examine their responses on exams and pinpoint whether the errors made relate to accuracy, options, or faulty thinking? and How can the instructional format be altered to provide additional opportunities in class to utilize the type of thinking required on the exams? An additional point of debate relates to both the efficacy and ethics of teaching students how to think/take exams. Does the learning that occurs while taking the test become lessened if students are taught ahead of time how to approach certain question types?

With respect to the system of analysis used to rate the student explanations in the current study, several questions need further attention. Can a critical thinking test be developed that measures a construct that corresponds with the faulty thinking strategies documented in this study? Are there dimensions related to student reasoning on exams other than those represented by the accuracy, options and error scores that should be explored? How universally employed (e.g., other courses) are the faulty thinking strategies identified in the current study? Can information learned regarding student reasoning be translated into steps for creating multiple-choice items that require higher order thinking skills?

LIST OF REFERENCES

Aiken, L. (1982). Writing multiple-choice items to measure higher order educational objectives. Educational and Psychological Measurement, 42, 803-806.

Bloom, B. S., & Broder, J. L. (1950). Problem solving processes of college students. Chicago, IL: University of Chicago Press.

Bowen, H. R. (1977). Investment in learning: The individual and social value of American higher education. San Francisco, CA: Jossey-Bass.

Bowering, E., & Wetmore, A. (1997). Success on multiple choice examinations: A model and workshop intervention. Canadian Journal of Counseling, 31, 294-304.

Connolly, J.A., & Wantman, M. J. (1964). An exploration of oral reasoning process in responding to objective test items. Journal of Educational Measurement, 1, 59-64.

Dempster, F. (1992). Using tests to promote learning: A neglected classroom resource. Journal of Research Development in Education, 25 (4), 121-125.

Ennis, R. H. (1993). Critical thinking assessment. Theory into Practice, 32, 79-186.

Facione, P. A. (1986). Testing college-level critical thinking. Liberal Education, 72, 221-231.

Facione, P. A., & Facione, N. C. (1994). The California Critical Thinking Skills Test: Test manual. Pacific Grove, CA: Midwest Publications.

Fuhrman, M. (1996). Developing good multiple-choice tests and test questions. Journal of Geoscience Education, 44, 379-384.

Haney, W., & Scott, L. (1987). Talking with children about tests: An exploratory study of test item ambiguity. In R. Freedle, R. Duran (Eds.), Cognitive and linguistic analyses of test performance. (pp. 299-321). Norwood, NJ: Ablex Publishing Corporation.

Jones, E. (Ed.). (1995). The national assessment of college student learning:

Identifying college graduates' essential skills in writing, speech and listening, and critical thinking. Washington, DC: National Center for Education Statistics, U. S. Department of Education, Office of Educational Research and Improvement.

Karras, R. (1978). Writing multiple choice questions: The problem and a proposed solution. History Teacher, 11, 211-218.

Karras, R. (1985). A realistic approach to thinking skills: Reform multiple-choice questions. Social Science Record, 22, 38-43.

Karras, R. (1991). Let's improve multiple-choice tests. OA Magazine of History, 6 (1), 8-9, 43.

Kazdin, A. (2001). Behavior modification in applied settings. Belmont, CA: Wadsworth/Thomson Learning.

Keeley, S. M., Shemberg, K. M., Cowell, B. S., & Zinnbauer, B. J. (1995). Coping with student resistance to critical thinking. College Teaching, 43, 140-145.

Kehoe, J. (1995). Writing multiple-choice test items. Washington, DC: US Department of Education. (ERIC Document Reproduction Service No. ED 3982360).

Killoran, J. (1992). In defense of multiple choice questions. Social Education, 56(2), 106-108.

Kropp, R.P. (1956). The relationship between process and correct item responses. Journal of Educational Research, 49, 385-388.

Linn, M. C., & Hyde, J. S. (1991). Trends in cognitive and psychosocial gender differences. In M. Gauvain & M. Cole (Eds.), Readings on the development of children (pp. 240-248). New York, NY: W. H. Freeman and Company.

Martinez, M. E. (1999). Cognition and the question of test item format. Educational Psychologist, 34, 201-218.

McClain, L. (1983). Behavior during examinations: A comparison of "A," "C," and "F" students. Teaching of Psychology, 10, 69-71.

McGuire, C. (1963). Research in the process approach to the construction and analysis of medical examinations. National Council on Measurement in Education Yearbook, 20, 7-16.

Merriam-Webster Collegiate Dictionary (10th ed.). (1996). Springfield, MA: MerriamWebster.

Morrison, S., & Free, L. W. (2001). Writing multiple-choice test items that promote and measure critical thinking. Journal of Nursing Education, 40, 17-24.

National Center for Educational Statistics. (1991). Commissioned papers for November 1991 National Conference. (ERIC Document Numbers: 340753 through 340768). Washington, DC: Office of Educational Research and Improvement, U. S. Department of Education.

Norris, S. P. (1990). Effect of eliciting verbal reports of thinking in critical thinking test performance. Journal of Educational Measurement, 27(1), 41-58.

Norris, S.P. (1992). A demonstration of the use of verbal reports of thinking in multiple-choice critical thinking test design. The Alberta Journal of Educational Research, 38(3), 155-176.

Paul, R. W. (1984). Critical thinking: Fundamental to education for a free society. Educational Leadership, 42, 4-14.

Rothstein, P. (1997). Developmental issues in teaching. New York: McGraw Hill.

Sattler, J. (1992). Assessment of children. (3rd ed.). San Diego, CA: Jerome M. Sattler, Publisher, Inc.

Scruggs, T., & Mastropieri, M. (1995). Teaching test-taking skills: Helping students show what they know. Cambridge, MA: Brookline Books.

Webb, L. (1993). The use of cognitive taxonomies in licensure and certification test development: reasonable or customary? Paper presented at the Annual Meeting of

the American Educational Research Association, Atlanta, GA.

Worth, S. (2000). Critical thinking ability as a comparative predictor of course performance. Unpublished dissertation. The University of Tennessee, Knoxville.

APPENDICES

Appendix A

Example of Recall and Comprehension Exam Items

Sample of an Exam Item Classified as Recall

When your instructor discusses Piaget's concrete operations and formal operations stages, the term operations can best be understood as synonymous with the concept of

- A. object permanence.
- B. symbolic thought.
- C. make-believe play.
- D. logical thought.

Sample of an Exam Item Classified as Comprehension

A behaviorist and humanist would most likely disagree about the validity of which of the following claims?

- A. People are innately good.
- B. The environment has some influence on behavior.
- C. There is some connection between behavior and feelings.
- D. A counselor can influence clients' discussion of their feelings.

Appendix B

Instructions for Rating Written Explanations

Written Analysis Rating Scale

Rating Guidelines

1. Compare each item with the information summary sheet for that item. Each option (A-D) should be rated 0, .5, or 1. Compute a composite score for each item. In cases where students do not explicitly match the explanation with a specific option, the rater may infer that an option was considered based on a linkage between the explanation and the option. One fact or statement may be used to support more than one option. In these instances, the same rating should be given. On items with an “all of the above” or “none of the above” option, the following procedures apply:
 - a. Follow the guidelines for rating any other option if the student provides an explanation for the “all” or “none” option.
 - b. In cases where the “all” or “none” option is selected and is the correct option, and the student does not give an explanation, a rating equal to the average rating for the other options should be given for the “all” or “none” option.
 - c. If the “all” or “none” option is the correct option and the only explanation given is for this option, then a rating of .5 should be assigned to the other options.
 - d. If the “all” or “none” option is not the correct option, the correct option is chosen, and no explanation is offered for the all or none option, the average ratings procedures described in (b) apply.
2. Indicate the number of options that were ruled in or ruled out for each item. The number may range from 0 to 4. It is not necessary for the student to explicitly identify the

option being ruled in/out. The rater may infer that an option was considered based on a linkage between information (whether accurate or inaccurate) included in the explanation and the option.

3. Identify whether the keyed response was chosen by indicating a “Y” for yes and an “N” for no in the appropriate column on the rating sheet.
4. For items where the keyed response was not chosen and the error category column is highlighted, determine the corresponding error category (as defined on the attached sheet) or categories for that item.

Scale Description for Accuracy Rating

0- inaccurate or no information. Examples include guesses, personal views that are inconsistent with course information, correct information that doesn’t provide support for the question/answer choice (i.e., mismatch), generalizations that are not supported by information presented in class or the course readings.

.5- partly correct information and/or correspondence with the question/option. Examples include explanations that lack a portion of the crucial information found on the information summary sheet, and/or personal views/background knowledge that are somewhat consistent with course information, repetition of the option (in cases where the option represents accurate information**) and/or accompanied by a statement such as “I know that answer choice is true,” or “That’s what the reading or the instructor said.”

1- information is accurate, corresponds with the question/option and includes the crucial information as outlined on the information summary sheet. Any generalizations made are consistent with the content presented in class or in the course readings.

** An exception: If the “all” or “none” response is the correct option and the explanation

contains a statement such as “all of the above are true,” a rating of 1 should be assigned.

Error Categories

Bias- personal experiences and/or values and/or opinions overrule factual information presented in explanation and/or personal experiences/values/opinions are presented in the absence of factual information.

Naiveté- explanation indicates that the student does not understand the meaning of a singular word (exclusive of vocabulary specific to course content) that is crucial in the stem or answer choices (e.g., majority, mortality) and/or student does not appear to possess information that is considered common knowledge (e.g., warning labels are currently on packages of cigarettes).

Misinterpretation- stem/answer choices were interpreted in such a way that the meaning of the stem and/or answer choice is either missed or altered and/or words/phrases that are crucial to comprehending the stem/answer choices are ignored and/or information inferred from stem or answer choices distracts one’s focus from the relevant information presented in the stem or answer choices.

Mismatch- information is accurate, yet is not the crucial information to be considered when answering the question.

Inconsistent standards- criteria by which one answer choice is ruled in or out is not applied to other answer choices.

Appendix C

Sample Target Items with Corresponding Student Explanations and Explanatory

Dimension Ratings

Sample Target Item # 1

A tobacco lobbyist is attempting to get you to vote against legislation that would require tobacco companies to fund independent medical research about the health risks of smoking. Which of the lobbyist's claims would you find credible?

- A. Cigarette use does not substantially increase the mortality rate among youngsters.
- B. Some individuals who are life-long smokers never develop heart disease, lung disease, lung cancer, or emphysema.
- C. Cigarette smoking ultimately is a personal choice.
- D. **All of the above claims have some basis in reality.**

Information Summary for Sample Target Item #1

- A. Cigarette smoking does not cause death in youngsters.
- B. Major long-term consequences of prolonged smoking can include respiratory and cardiovascular disease. No evidence that states that all or even a majority of long-term smokers develop these diseases.
- C. People can be influenced to smoke, but cigarette smoking is ultimately a personal choice.
- D. Supported by evidence in answer choices a-c.

Sample Student Explanations and Ratings for Target Item #1

Student 1

I chose answer C for this question because it is a personal choice. Answer A is wrong because smoking is the number one most preventable cause of death. Letter B, may happen but it's not very likely for a smoker to remain completely healthy.

Accuracy- 1.5 (1.0 for C and .5 for B)

Options- 3 (A, B, C)

Error- A mismatch error was assigned based on the option A response. Information regarding smoking being a preventable cause of death was presented in course information, yet not relevant to answering this particular question. A misinterpretation error was assigned based on the choice B response. It appears that the student misunderstood what option B was saying in that her paraphrase seemed to be restating option B, yet she didn't recognize the parallel between her response and option B.

Student 2

I chose answer C because people (most) know what they are doing to their bodies and don't seem to care until they have a harsh disease and then realize the effects of cigarette smoking.

Accuracy- 0

Options- 1 (C)

Error- A bias error was assigned in that the answer selection appears to be based on personal experience/views independent of factual information presented in the course.

Student 3

I chose B because cigarette use does cause health risks and greater threats of death on in life. Cigarette smoking is a personal choice but that doesn't have any basis for research on health, but it is true some people smoke and live to be 100 so B seemed reasonable.

(Student 3 continued)

Accuracy- 3 (1 for A, B, C)

Options- 3 (A, B, C)

Error- A misinterpretation error was assigned based on the portion of the explanation that states that cigarette smoking as a personal choice has no relevance for health research. Student also seemed to accept choice A; however, did not include any rationale in the explanation that justifies why B was a more reasonable response than A. This type of thinking is consistent with an inconsistent standards error.

Sample Target Item #2

You are listening to a local civic leader's speech on teenage pregnancy. Which of the following claims offered in the speech will you be most likely to challenge in the question and answer session following the speech?

- A. **A majority of teen mothers will become pregnant again within 12 months of giving birth.**
- B. Fewer than half of pregnant teenagers elect to have an abortion.
- C. Older teenagers are more likely to become pregnant than are younger teenagers.
- D. The teen pregnancy rate is highest in the U. S. of any developed country.

Information Summary for Sample Target Item #2

- A. One study showed that there is a 20% probability of a repeat pregnancy within 12 months of giving birth.
- B. 40% of pregnancies end in elective abortions.
- C. Percentage of 18-19 year olds who become pregnant is more than double the percentage of 15-17 year olds.
- D. Pregnancy rate in America is higher than any developed country.

Sample Student Written Explanations for Sample Target Item #2

Student 1

I would question B because from the facts obtained both in class and in our reading it shows that 40% are elective abortions and 10% are spontaneous abortions. These statistics are not fewer than half.

Accuracy- 1 (1 for B)

Options- 1 (B)

Error- A misinterpretation error was assigned because it appears the student overlooked the word elect in option B.

Student 2

I chose C. Older teenagers have a higher pregnancy rate but it is because more are having sex, not because they are older or any other reason.

Accuracy- .5 (.5 for C)

Options- 1 (C)

Error- A misinterpretation error was assigned because it appears the student changed the meaning of the chosen option.

Student 3

I chose C. I feel that in this day and age that is a false statement. Nowadays there are just as many 13 and 14 year olds getting pregnant as 18 and 19 year olds. I think this is partly because of too little, too late education on sex and teen pregnancies.

Accuracy- 0

Options- 1 (C)

Error- A bias error was assigned in that the explanation contained no factual information. Personal views/experiences were the basis for answer selection.

Sample Target Item #3

As a member of Congress, you are being asked to support legislation intended to reduce youth smoking in the U.S. Which of the following legislative efforts has the best chance of substantially reducing teenage smoking?

- A. Fund smoking cessation programs for all teenagers who regularly smoke.
- B. Require tobacco companies to put multiple warnings on cigarette packages.
- C. **Substantially increase taxation on cigarettes.**
- D. Remove restrictions on the sale of cigarettes to teenagers, thus reducing the attractiveness of smoking to youngsters.

Information Summary for Sample Target Item #3

- A. These programs tend to have low success rates. Recruiting and retaining adolescents in these programs are difficult. Even for the adolescents who stay with the programs, addiction and social influences make it difficult to stop smoking.
- B. Surgeon General warnings are currently on individual packs of cigarettes. One may reasonably conclude that the multiple warning labels would have minimal impact.
- C. Some studies indicate that adolescents are much more sensitive to a price hike than are young adults. AND/OR because adolescents prefer the higher-priced premium brands, a substantial price hike is likely to be acutely felt.
- D. No evidence reported in class suggests lack of legal access increases attractiveness of smoking for teens. AND/OR a logical conclusion: greater access could yield a greater number of addicted teens.

Sample Student Explanation and Ratings for Sample Target Item #3

Student 1

I chose B because warnings would warn teenagers each time they bought a pack. Taxation wouldn't help because teenagers go for pricey brands anyway. If restrictions were lifted, accessibility would increase and usage would increase. A cessation program would not work because you would have trouble getting teenagers to admit that they are regular smokers. They would be afraid of getting in trouble with parents, etc.

Accuracy- 1.5 (.5 for A, 1 for D)

Options- 3 (A, B, D)

Error- A naiveté error was assigned because it appears from the first line in the student explanation that the student did not have the common knowledge that warning labels already exist on cigarette packages.

Vita

Monica Anne Wallace was born in Winston-Salem North Carolina on November 7, 1960. She graduated from North Davidson Senior High School in Lexington, North Carolina. From there, she went to the University of North Carolina at Greensboro and received a B. S. in psychology in 1983. Monica received a Masters in Education from the University of South Florida in 1998. She graduated from The University of Tennessee, Knoxville with a Doctor of Philosophy degree, major in Education, in 2003.

7797 0196 20

11/05/03



MAB

