

A Comparison of the Quality of Multiple-Choice Questions from CPA Exams and Textbook Test Banks

Charles D. Bailey
University of Central Florida

Julia N. Karcher
University of Louisville

Barbara Clevenger
Louisiana College

Abstract

Multiple-choice test questions have substantial support as a reliable and valid evaluation method. In practice, however, instructors may prepare or select multiple-choice questions based on content and without adequate attention to item quality. This paper has two specific objectives: (1) to familiarize accounting educators with item-writing “rules” (commonly accepted guidelines intended to achieve validity); and (2) to provide evidence on the relative rule compliance of test bank questions and CPA exam questions (“quality”). We selected random samples of 100 multiple-choice questions from each source and employed judges to evaluate them for rule compliance. Because not all rules apply to all questions, we categorize the questions into eight types based on their format (question versus completion), numerical calculation requirements, and the presence or absence of words in the responses. The test bank questions showed consistently lower quality than CPA exam questions across the six main question types. Violations were about 4.6 times more frequent for test bank questions (6.42% versus 1.41%). The analysis indicates that some types of questions are especially prone to rule violation.

Multiple-choice questions (hereafter MCQs) are widely used in both professional certification examinations and classrooms. They are especially convenient today for accounting educators, since most textbook packages include a test bank. The advantages of using MCQs are ease of grading, standardization, and the opportunity to cover a large amount of material quickly. Collier and Mehrens (1985) strongly

defend the value of MCQs and encourage readers to develop and use such items in accounting courses. They “believe that multiple-choice testing is the best available testing methodology to achieve the desired qualities of reliability and validity in the assignment of grades based on tests to measure student learning” (p. 41). As Aiken (1987) notes, MCQs can measure not just simple recognition but higher order learning skills—by requiring

one to examine multiple premises, to perform classifications, to determine relations and correlations, or to consider if-then situations.

The larger class sizes dictated by funding restrictions also ensure the continued prominence of MCQs. In Florida, for example, population growth and budgetary constraints are expected to increase college class sizes dramatically during the coming decade (Date, 1995). A MCQ exam can quickly test whether students have achieved some level of technical knowledge before they apply that knowledge through case studies, cooperative learning projects and other less traditional classroom methods in the development of critical thinking and communication skills stressed by the Accounting Education Change Commission.

Research examining the validity and reliability of test banks is limited. Clute, McGrail, and Robinson (1988), assuming test bank questions were reasonably well written, found a significant level of bias in the distribution of correct answers for MCQs in principles of accounting textbooks. Clute and McGrail (1989) reported similar findings for cost accounting textbooks, and Geiger and Higgins (1998) discover such biases in professional certification exams. The concern is that “test wise” test-takers (Gibb 1964) could unfairly use this bias to improve their scores and decrease the validity of a test.

Gibb (1964) examined students’ awareness of seven secondary clues, such as grammatical errors and longer correct alternatives, to determine the correct answer to a MCQ. Gibb developed an Experimental Test of Testwiseness instrument which subsequent researchers

have used to determine that certain secondary clues may be more important than others (Miller, Fagley & Lane, 1988; Miller, Fuqua & Fagley, 1990). Writers of MCQs need to be aware of these secondary clues to construct valid, reliable test questions.

A number of education researchers have developed rules for MCQ construction, and although broad support exists for many of those rules, we find no studies that have investigated whether examination items used in accounting follow the rules. Further, the Framework for the Development of Accounting Education Research (Williams et al., 1988) notes that “[a]n unfortunate truth is that many classroom tests are poorly prepared” (p.105) and that “[t]he subjects of testing and grading have been barely scratched in accounting literature” (p.108).

This study samples MCQs from Certified Public Accountant (CPA) examinations and from test banks accompanying accounting textbooks, to assess their adherence to the MCQ-writing rules. It has two objectives: (1) to familiarize accounting educators with item-writing “rules,” or commonly accepted guidelines intended to achieve validity; and (2) to provide evidence on the relative rule compliance of test bank questions and CPA exam questions (“quality”). Despite the nondisclosure of future CPA exam questions, the existing body of questions has ongoing utility, and the questions provide an interesting benchmark because of the effort that went into their construction.

The remainder of this paper is divided into four sections. The first discusses the relevant research and summarizes the various item construction rules. The second describes the methods and data

sources used in this study. The third details the results of our analysis, and the last discusses the limitations, conclusions and implications of the study.

The Development of MCQ-Writing Rules

Haladyna and Downing (1989) reviewed 96 theoretical and analytical studies and developed a taxonomy of 43 rules designed to improve MCQ writing. They found that 33 of the rules enjoyed strong consensus among the sources. Aiken (1987) provides 14 guidelines for writing MCQs and also discusses MCQ arrangement, test taking, scoring and sociological effects of multiple-choice testing. In the accounting literature, Baldwin (1984) details eight guidelines for writing MCQs.

Since the current study concerns individual MCQ construction, not overall test validity,¹ we consider only the guidelines which deal with the individual MCQ, its stem (the material preceding the responses) and the responses. Based on the guidelines cited in the literature, we developed a set of 15 rules dealing with the overall test question, the question stem and the responses.

The final set of rules appears in Table 1. We selected these items based on their widespread acceptance, relevance to the accounting setting, and our ability to judge whether MCQs adhere to them. Because we chose to sample from a wide variety of

topic areas, we did not attempt to assess the technical accuracy of the answers.

In addition to the general criteria calling for clarity of communication, researchers have tried to develop measures of readability. Readability formulas, which assign a measure of difficulty or required grade level to text, now are readily available with the major word processing packages. Although readability formulas appear frequently in accounting research (Jones & Shoemaker, 1994), they are quite limited, being based on statistics that include sentence length, word length, and other syntactical variables. The Flesch Reading Ease score decreases with sentence length and syllables per word, and ranges from 100 (very easy) to 0. Thirty and below is "very difficult." Gunning's Fog Index is intended to indicate the school-grade needed for comprehension, with college graduates at grade 17, and the Flesch-Kinkaid Grade Level is a similar measure. Flory, Phillips & Tassin (1992) argued for the usefulness of such formulas in evaluating textbook readability, prompting a sharp response from Stevens, Stevens, and Stevens (1993). The latter favor the use of the "cloze" procedure, which utilizes a sample of readers from the intended audience. For the items in our study, however, the intended audience consists of students who have studied specific technical materials, and the broad inclusiveness of our sample makes this procedure impractical. Thus, although we agree that the readability formulas are severely limited in their ability to predict how well students will understand a particular writer's style, we believe that the statistics may shed some light when applied to the samples drawn from our two sources.

¹. Formal validation typically is done for standardized tests, but not for classroom tests. The rules that we are considering are at an entirely different level, and have the purpose of improving validity. For a recent, thorough summary of the research on developing and validating MCQs, see Haladyna (1994).

Table 1: List of Rules

The stem

- 1 deals with one central problem. ^{a, b}
- 2 has the to-be-completed phrase placed at the end. ^a
- 3 uses simple numbers to test a concept. ^b

The responses

- 4 are grammatically consistent with the stem. ^{a, c, e, f}
- 5 do not unnecessarily repeat language from the stem. ^{a, f, g, h}
- 6 are all of approximately the same length. ^{a, c, e, f, h}
- 7 consist only of numbers that are plausible based on the data in the problem. ^{a, b, c, e, f, h}
- 8 avoid the use of “all of the above.” ^{a, f, g}
- 9 avoid the use of “none of the above.” ^{a, f, g}
- 10 contain one allowable answer rather than the use of “A and B,” “A, B, and C,” etc. ^{f, g}

The question overall

- 11 avoids excessive verbiage. ^{a, e, f, g, h}
- 12 is grammatically correct throughout. ^{a, e, f, g, h}
- 13 is clearly written. ^{a, e, f, g, h}
- 14 contains no negative or other counter-intuitive wording without underlining or other special emphasis. ^{c, d, e, g, h}
- 15 contains no other offensive or undesirable characteristics not listed above. ⁱ

Main sources in addition to the review by Haladyna & Downing (1989):

^a Aiken (1987)

^b Baldwin (1984)

^c Denova (1979)

^d Green (1963)

^e Grolund (1968)

^f Lutterodt & Grafinger (1985)

^g Miller & Erickson (1985)

^h Wood (1960)

ⁱ Added to encompass sexist language, offensive or distracting scenarios, etc.

Note: Appendix A shows instructions to judges and clarifying examples to aid interpretation.

Methods

We used the following steps to obtain the data set: (1) assembling the sampling frame, (2) sample selection, (3) retyping and proofing, (4) matching of relevant rules with each question, (5) preparation of instructions for judges, and (6) judgments and reconciliation of differences. In addition, we prepared the text for readability-checking. Details of these methods follow.

Sampling

We began by assembling the sampling frames (the two sets of MCQs from which we would select the samples). For the CPA exams, we chose all sections of the exams for years 1986 (the first year of the current exam format) through 1990. With 60 MCQs in each section of each exam, the sampling frame contained 3,000 MCQs.

For the textbook test banks, we began by asking our colleagues at two institutions for all test banks that they had available. Then, for publishers that appeared to be underrepresented, we asked sales representatives for copies of all accounting-related test banks. We eliminated test banks that claimed to use some questions from professional exams, but which failed to identify such questions. The result was a sampling frame of sixteen test banks. We identified the MCQs and assigned sequential numbers to the 13,770 questions in the sampling frame.

From each sampling frame, we selected 100 different MCQs through simple random sampling. If a selected test bank question carried the notation that it was taken from a professional

examination, we selected the next question in the test bank.

Instructions for Judges

A pilot test, using 20 randomly selected MCQs, served to refine the rules and instructions. The three researchers independently evaluated the twenty questions and discussed the clarity of the rating instructions. The Instructions to Judges, explaining each rule, appear in Appendix A. After reading these instructions, the judges discussed them with one of the researchers before proceeding.

Question Retyping, Proofing, and Preparation

CPA exam questions have a characteristic appearance because of their font and layout. To eliminate any bias related to such appearance factors, we disguised the sources from our judges. We had the questions retyped into a uniform format, proofread, and spell-checked. Spelling errors were verified in the original to decide whether to correct them or identify them as “chargeable” violations.

Question types. After proofing, we appended to each MCQ the list of rules *relevant* to that question. There are eight question “types,” defined by the presence or absence of three characteristics. See Table 2 for these types. The first, for example, is designated type CNW, because it is in Completion format, involves Numerical calculations, and includes Words in the responses. All 15 rules are relevant to this type of question. At the other extreme is type X, which has none of these characteristics, so that rules 2 through 7 do not apply. To avoid confusion, all fifteen rules always appeared in the list, but with the rules not relevant to the particular item struck through. This uniform format reduced

cognitive strain for the judges, and also allowed them to double-check our decisions about the relevance of the rules for each question.

Judgments and Reconciliation of Differences

The two judges for the study were Ph.D. students in accounting, who were paid for the task and were otherwise not associated with the project. One of them had previously authored supplemental

materials for a textbook. They independently evaluated each of the 200 MCQs to determine whether it conformed to each relevant rule, and agreed on 93.5 percent of the judgments. The κ (kappa) coefficient (Wickens, 1989) of 0.89 is very high, indicating considerable rater agreement after adjusting for chance. Because of the unexpected death of one of the judges, one of the authors reconciled the differences in conference with the other judge.

Table 2: Question Types

Question Type	Defining Characteristics				Number Occurring in:	
	Completion Format (C)	Numerical Calculations (N)	Includes Words In Responses (W)	Rules Not Relevant	CPA Exams	Test Banks
CNW	x	x	x		1	9
CN	x	x		4, 6	18	13
CW	x		x	3, 7	41	36
C	x			3, 4, 6, 7	0	1
NW		x	x	2, 5	1	9
N		x		2, 4, 5, 6	13	9
W			x	2, 3, 5, 7	24	22
X				2—7	2	1
				Total	100	100

Preparation for Readability Analysis

To allow some analysis of the material by computerized readability checking, we generated files containing revised text in sentence form. Our main guidelines were as follows: (1) For continuation-format questions with verbal responses, we generated complete sentences by pasting each of the responses in turn onto the stem. For numerical responses, however, we used only the first response to generate a sentence, because the others would be grammatically identical. (2) When the format was not completion, we removed

answers that were not complete sentences. (3) We deleted journal entry answers. (4) We made no attempt to provide reasonable paragraphs.

Data Analysis

Question-Writing Rules

Table 3 contains an analysis of violations overall, by question type, and by rule. The “violation %” at the bottom of each rule column shows the percentage of questions that violated each rule. For two of the rules, we found no violations in our sample; they were rule 5 (the responses do not unnecessarily repeat language from the stem) and rule 10 (the responses contain one

allowable answer rather than the use of “A and B,” “A, B, and C,” etc.). The most-violated rules were 9 (the responses avoid the use of “none of the above”), 13 (the question overall is clearly written), 7 (the responses consist only of numbers that are plausible based on the data in the problem) and 3 (the stem uses simple numbers to test a concept). Notably, the CPA exam questions never violated rules 3 or 9.

The violation proportions are higher for the test bank MCQs than for the CPA exam questions, and this holds true for every type of question. For all of the MCQs combined, the difference in violation proportions is highly significant ($\chi^2=40.22$, $p<0.0001$), as shown in Table 3 and based on the following contingency table:

	Violations	Total Potential Violations	Percentage
Test Banks	79	1230	6.42
CPA Exams	17	1202	1.41
Total	96	2432	3.95

To conserve space, Table 3 omits details on question types C and X, as these categories contained only four questions with no violations. Of the six remaining types of questions, the individual χ^2 statistics are significant for the two largest groups, CW ($p<0.0001$), and W ($p=0.0033$). An interesting observation, however, involves the two types that have numerical problem content but also contain words in the answers (CNW and NW).

These have the highest violation rates, but the CPA exam has largely (and perhaps wisely) avoided them, with only one of each represented in the CPA sample. Thus the sample size is too small for meaningful individual comparison of this group, but the violation rates of 8.2% and 7.1% by these test bank MCQs contribute to the test banks' overall weakness.

Readability Statistics

As a supplemental analysis, we used the Grammatik V facility of WordPerfect to generate certain key statistics on the sample MCQs. Subject to the caveats discussed above, we present comparative statistics in Table 4.

The statistics indicate that the CPA exam MCQs contain slightly more difficult reading. This is not surprising, as the exam tests more advanced knowledge and must include more technical language. The Flesch-Kincaid Grade Level of 14 for the CPA exam seems appropriate, however, as *Fortune Magazine* registers 14.0 and *Barron's* 11.6 (Heath & Phelps, 1984). The Flesch Reading Ease score is on the margin of being “very difficult” at 30 for the CPA exam. Gunning’s Fog

Index scores of 16 for the test banks and 18 for the CPA exam also seem appropriate, as grade 17 represents college graduates.

The differences in sentence length are interesting. The CPA exam includes fewer “short” sentences and more “long” sentences. Fully 90% of the test bank sentences are twelve words or less in length, compared to 57% for the CPA exam. This difference may reflect a common, and questionable, view that shorter sentences are easier to read. As Stevens, Stevens, and Stevens (1993) note, the sentence “I think, therefore I am” requires more effort to comprehend (is at a higher reading level) than “I went to the baseball game and bought hamburgers, sodas, popcorn and peanuts” (p.288).

Table 4: Readability Statistics

<u>Statistic</u>	<u>CPA Exams</u>	<u>Test Banks</u>
Flesch-Kincaid Grade Level	14	12
Flesch Reading Ease	30	34
Gunning’s Fog Index	18	16
Proportion of “Short” Sentences (<12 words)	240/420 = 57%	348/390 = 90%
Proportion of “Long” Sentences (>30 words)	39/240 = 16%	6/390 = 1.5%
Average Words/Sentence	17.5	14.7
Average Letters/Word	5.11	4.97
Average Syllables/word	1.88	1.87

Summary and Conclusions

Although pedagogical changes are likely to incorporate less objective approaches to grading—e.g., portfolios, case studies, cooperative learning outputs and other nonstandardized assignments — MCQs will continue to have a place in student assessment. Dow and Feldmann (1997) found that about two-thirds of their respondents make heavy use of MCQs. Collier and Mehrens (1985) argue strongly

for MCQs as a format for reliable and valid assessment of knowledge. Students need to be exposed to all types of questions that are included in the professional licensing examinations, including MCQs. And finally, MCQs allow economical testing of a large volume of material.

In an effort to provide assistance for those who write their own MCQs and for those who wish to avoid validity-threatening mistakes when choosing

MCQs from textbook test banks, we have summarized generally accepted rules for MCQ writing (i.e., Table 1). We have then analyzed 200 randomly-selected MCQs from accounting textbook test banks and CPA examinations to determine whether their construction met those rules.

In general, past CPA exam MCQs, and many of the test bank MCQs, meet the question-writing rules described in this paper. The majority of questions from all sources (96%) showed no violations. Not surprisingly, the CPA examination questions violated significantly fewer rules than the test banks. The incidence of errors should not deter educators from using test banks, but they should use their knowledge of the rules to edit questions for rule violations.

The most common violations of the accepted rules were the inclusion of “none of the above” as a response, unclear writing, implausible numbers given the problem data, and the use of complex numbers to test a concept. The questions that resulted in the largest number of test bank violations were those that required numerical calculations but included words in the responses (i.e., the type CNW and NW questions). These were rare in the CPA examinations (2 in our sample of 100), and anyone selecting or writing MCQs may wish to avoid them.

In the past, most CPA exam questions have been developed in-house. Currently, practitioners and academics are encouraged to submit questions, and the American Institute of Certified Public Accountants (AICPA) conducts two-day workshops to train potential question-writers. Based on our findings, it appears that the guidelines used here have probably formed a part of the AICPA’s editorial policy in the past, and question

writers will be expected to consider them in the future.

Reviewers of textbooks might wish to examine MCQs in the books and related test banks, using the guidelines we have described. Such an analysis would provide an uncommon and insightful perspective on one aspect of the writing quality.

All of the available indices to test readability suggest that both test bank and CPA examination MCQs are at the appropriate collegiate reading level. We have noted, however, that the readability formulas do not necessarily translate to understandability in the mind of a reader.

This research is subject to several limitations. One set of limitations concerns the authors’ choice of criteria (rules) for judging “quality” and the judges’ ability to assess violations. Another concerns the sampling of questions. The CPA exam questions came from a known and completely defined universe, but the test bank sampling frame represents a convenience sample. Differences in quality may exist among publishers. Also, this study does not address overall test validity, nor do we assess the relative importance of the fifteen rules or the seriousness of their violation. The rules are consensus judgments from the education literature, and empirical evidence of their actual impact on test reliability and validity is sketchy.

Future research might compare questions from other professional exams such as the Certified Management Accountant, Certified Internal Auditor,

etc.² Knowledge of differences related to various subject areas (tax, financial, management, etc.) would provide more specific guidance to instructors. An interesting related question, which may not be answerable empirically, is whether it is better to “protect” and reuse exams (which allows them to be improved and validated), or to return them for students to keep and study. In sum, the area of overall test validity deserves further attention.

MCQs have their limitations in measuring achievement, but they continue to be useful tools. This study has provided a common list of MCQ-writing rules and discussed the violations of those rules in accounting textbook test banks and CPA examinations. Instructors who follow these guidelines when writing or selecting MCQs should produce more readable examinations that will in turn allow students to more easily and accurately demonstrate their knowledge.

2. We originally hoped to include a sample of questions generated by instructors for their own classroom use. A pilot request for such items, FAXed to the chairs of ten randomly-selected accounting departments, brought no response. We suspect that one reason is that individuals are unsure which questions are their “own” work, rather than being adapted from other sources.

Appendix A: Instructions To Judges

A number of guidelines exist for composing “good” multiple-choice exam questions. We have reviewed the literature on this topic, and have summarized certain of these rules. You will be judging numerous multiple-choice questions, deciding whether each question complies with the rules.

A multiple-choice question consists of two main parts: a stem, which sets up the question, followed by a set of possible responses. The question may or may not use a completion format, in which the response completes a sentence left incomplete in the stem.

You will be rating each multiple-choice question in three categories. Category one concerns the stem. Please indicate whether or not the stem satisfies each criterion by marking “true” or “false.” Category two concerns the response alternatives. Please mark “true” to indicate that all of the responses satisfy the criterion or “false” to indicate that one or more responses violate the criterion. For any response of “false,” write an explanation on the question unless the item is self explanatory.

The following pages contain descriptions and explanations of the criteria. Please read these carefully, and we will discuss them. Some of the criteria relate only to certain types of exam questions. Only an appropriate sub-set will appear for each exam question that you evaluate.

Criteria for the Stem

The stem:

deals with one central problem.

The reader should be able to state the theme of the question without having to read the responses first. (Note, however, that some questions may be so short that the reader can easily grasp the stem and the responses simultaneously. In that case, no potential confusion exists.)

has the to-be-completed phrase placed at the end.

uses simple numbers to test a concept.

Example: “What is the balance in Accounts Receivable, given the following information? “

Complex numbers:

Beginning balance	\$ 950,345
Charge sales	2,250,987
Payments received	2,345,324

Simple numbers:

Beginning balance	\$ 950
Charge sales	2,250
Payments received	2,000

Criteria Concerning the Response Alternatives

The responses . . .

are grammatically consistent with the stem.

do not unnecessarily repeat language from the stem.

are all of approximately the same length.

Use your discretion. *The key factor is whether one of the answers is noticeably longer or shorter than the others.* For example, if four responses range from 1 to 3 or 4 lines in length, you might judge that no one of them stands out from the others. If three are one-line responses and the fourth is 4 lines long, then you might judge that it stands out prominently.

consist only of numbers that are plausible based on the data in the problem.

Example: Determine the balance in the cash account, given the following data:

Beginning balance, \$2000; Deposits, \$350; Disbursements, \$1050.

“Plausible” responses would include \$ 3400, \$1300, or \$600, but would *not* include \$6.77, which could not result from any plausible addition and subtraction operations.

avoid the use of “all of the above.”

avoid the use of “none of the above.”

contain one allowable answer rather than the use of “A and B,” ”A, B, and C,” etc.

For example, a setup like this meets the criterion:

	<u>Plan One</u>	<u>Plan Two</u>
A.	Yes	No
B.	Yes	Yes
C.	No	Yes
D.	No	No

The following violates the criterion because the forbidden format appears in D and E:

- A. Common Stock
- B. Preferred Stock
- C. Treasury Stock
- D. A & C
- E. A, B, & C

Criteria Related to the Test Question Overall**The question overall:**

Note: Give positive responses here only for criticisms not already covered by your earlier responses. Do not duplicate, for example, a criticism that language in the stem is repeated in the responses. This would have been covered under the stem-related items.

avoids excessive verbiage.

is grammatically correct throughout.

is clearly written.

contains no negative or other counter-intuitive wording without underlining or other special emphasis.

Example: "Which of the following is not an asset?" Not should be underlined.

contains no other offensive or undesirable characteristics not listed above.

Examples that come to mind are sexist language, grotesque or offensive scenarios, unrealistic examples, and outdated technical language or inappropriate terminology. Be very specific in explaining any such criticism.

Appendix B: Selected Examples of Rule Violations

Below are examples of questions that the judges decided violated rules of construction. We have chosen examples for rules that may not be self-evident from the Instructions in Appendix A, and that we believe to be sufficiently interesting and nontrivial.

The following item was found to violate rule 1 ("The stem deals with one central problem."):

G has adjusted gross income of \$40,000. G incurs the following expenses in G's hardware business:

- \$300 sales tax.
- \$300 vehicle registration.
- \$20 vehicle inspection fee.
- 200 water and sewer fee.
- \$100 parking fees incurred while visiting clients.

How much will G be able to capitalize or deduct?

- a. \$820.
- b. \$800.
- c. \$0.
- d. \$300.
- e. \$920.

The following item was found to violate rule 2 ("The stem has the to-be-completed phrase placed at the end.")

_____ is a notice, issued by a bank, that a checking account has been reduced.

- a. Credit memoranda
- b. Debit memoranda
- c. Bank reconciliation

d. Bank statement

The following item was found to violate rule 4 (“The responses are grammatically consistent with the stem.”):

An example of a pricing objective is

- a. to maintain a stated rate of return on investment.
- b. to maintain a price that should support the trend of total sales increases.
- c. to identify and adhere to short-run and long-run pricing strategies.
- d. all of the above **(sic)**

The following item was found to violate rule 7 (“The responses consist only of numbers that are plausible based on the data in the problem.”):

Calvin Soles has a long-term debt of \$500,000 that is due in 10 years. He wants to make 10 annual deposits into a fund that will earn annual interest (compounded) of 8 percent to retire this debt. If the table factor for the future amount of an ordinary annuity of 10 rents of 1 at 8 percent is 14.486562, then the amount to be deposited at the end of each of these years is

- a. \$50,000.00.
- b. \$60,000.00.
- c. \$34,514.75.
- d. some other figure.

The following item was found to violate rule 11 (“The question overall avoids excessive verbiage.”):

An auditor would be most likely to consider modifying an otherwise unqualified opinion if the client's financial statements include a footnote on related party transactions

- a. Representing that certain related party transactions were consummated on terms equivalent to those obtainable in transactions with unrelated parties.
- b. Presenting the dollar volume of related party transactions and the effects of any change in the method of establishing terms from that used in the prior period,.
- c. Explaining the business purpose of the sale of real property to a related party.
- d. Disclosing compensating balance arrangements maintained for the benefit of related parties.

The following item was found to violate rule 13 (“The question overall is clearly written.”):

In 19X3, M transfers a \$200,000 life insurance policy on M's life to G. The policy is worth \$50,000 at the time of transfer and G pays \$50,000. When M dies in 19X5, the \$50,000 cash is still in a savings account. The consideration offset when computing M's gross estate is

- a. \$50,000.
- b. \$200,000.
- c. \$150,000.
- d. \$0.

- e. \$100,000.

The following item was found to violate rule 14 (“The question overall contains no negative or other counter-intuitive wording without underlining or other special emphasis.”):

Which of the following is not a variable in the determination of whether investigation of business expenses can be taken?

- The amount of reasonable expenditures.
- The nature of the business being investigated.
- Whether or not the acquisition actually takes place.
- The current business of the taxpayer.
- None of the above.

The following item was found to violate rule 15 (“The question overall contains no other offensive or undesirable characteristics not listed above.”) It was deemed trivial, as it requires only common knowledge; only “b” concerns communication.

The communication stage of accounting is accomplished by

- sorting data.
- reporting to decision makers.
- recording data.
- processing data.

References

- Aiken, L. R. (1987). Testing with multiple-choice items. *Journal of Research and Development*, Vol. 20, No. 4, pp. 44–58, reprinted in *Research in Testing*. Harrisonburg, Va.: Center for Research in Accounting Education, James Madison University, 1990, pp. 19–45.
- Baldwin, B. A. (1984). The role of difficulty and discrimination in constructing multiple-choice examinations with guidelines for practical application. *Journal of Accounting Education*, v.2, n.1: 19–28.
- Clute, R. C., and McGrail, G. (1989). Bias in examination test banks that accompany cost accounting texts. *Journal of Education for Business* 64 (March):245–247.
- Clute, R. C., McGrail, G., and Robinson, A. V. Jr. (1988) Choice bias in multiple-choice questions from principles of accounting test banks. *Accounting Educators' Journal* 1(2): 92-96.
- Collier, H. W., & Mehrens, W. A. (1985). Using multiple-choice test items to improve classroom testing of professional accounting students. *Journal of Accounting Education*, v.3, n.2,: 41–51.
- Date, S. (1995). Universities look ahead with dread. *Orlando Sentinel*. December 4: A-1, A-20.
- Denova, C. C. (1979). *Test construction for training evaluation*. Van Nostrand Reinhold Company.
- Dow, K. J., and Feldmann, D. A. (1997) Current approaches to teaching intermediate accounting. *Issues in Accounting Education*, v.12 n.1: 61–75.
- Flory, S. M., Phillips, T. J. Jr., & Tassin, M. F. (1992) Measuring readability: A comparison of accounting textbooks. *Journal of Accounting Education* 10(1):151-161.
- Geiger, M. A., and Higgins, M. M. (Forthcoming). On the answer-arrangement bias of professional certification examinations in accounting. *Accounting Educators' Journal*

- Gibb, B. (1964). *Test-wiseness as a secondary cue response*. (Doctoral dissertation, Stanford University) Ann Arbor, MI: University Microfilms, No. 64-7643
- Green, J. A. (1963). *Teacher-Made Tests*. Harper & Row.
- Grolund, N. E. (1968). *Constructing achievement tests*. Prentice Hall.
- Haladyna, T. M. (1994). *Developing and validating multiple-choice test items*. Hillsdale, N.J.: Lawrence Erlbaum Associates.
- Haladyna, T. M., & Downing, S. M. (1989). A taxonomy of multiple-choice item writing rules. *Applied Measurement in Education*, v.2, n.1: 37-50.
- Heath, R. L., & Phelps, G. (1984, Summer). Annual reports II: readability of reports vs. business press. *Public Relations Review* 10: 56-62.
- Jones, M. J., & Shoemaker, P. A. (1994). Accounting narratives: a review of empirical studies of content and readability. *Journal of Accounting Literature* 13: 142-184.
- Lutterodt, S. A. & Grafinger, D. J. (1985). *Developing objective test items I*. General Physics Corporation.
- Miller, P. M., Fuqua, D. R., and Fagley, N. S. (1990) Factor structure of the Gibb experimental test of testwiseness. *Educational and Psychological Measurement* 50 (Spring): 203-208.
- Miller, P. M., Fagley, N. S., and Lane, D. S. (1988). Stability of the Gibb (1964) experimental test of testwiseness. *Educational and Psychological Measurement* 48 (Winter): 1123-1127.
- Miller, P. W. & Erickson, H. E. (1985). *Teacher-written student tests: a guide for planning, creating, administering, and assessing*. National Education Association.
- Stevens, K. C., Stevens, K. T. and Stevens, W. P. (1993). Measuring readability: a comparison of accounting textbooks: a response. *Journal of Accounting Education* 11(2):287-292.
- Wickens, T. D., *Multiway contingency tables for the social sciences*. Hillsdale, NJ: Lawrence Erlbaum Associates (1989) p. 238.
- Williams, J. R., Tiller, M. G., Herring, H. C. III, & Scheiner, J. H. (1988). *A framework for the development of accounting education research*. Sarasota, Florida: American Accounting Association.
- Wood, D. A. (1960). *Test construction development and interpretation of achievement tests*. Charles E. Merrill Books, Inc.

Table 3: Rule Violations

Ques. Type*	Source (number): T=Test bank C=CPA Exam	Violations by Rule Number															Total Violations	Possible Violations†	Violation Rate (%)	χ^2	p			
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15								
CN	T(13)	1		1													6	169	3.6	0.738	0.3901			
	C(18)	3									1		3				5	234	2.1					
CNW	T(9)			1				2		6							11	135	8.2	1.319	0.2508			
	C(1)												2				0	15	0.0					
CW	T(36)		2		2				5	7	1		2	3	3		25	468	5.3	18.674	0.0000			
	C(41)										2		2				4	533	0.8					
N	T(9)	2		2				1			1		1				7	99	7.1	2.462	0.1166			
	C(13)							1			1		2				4	143	2.8					
NW	T(9)	1		1				2		5	1	2	1			1	14	117	12.0	1.743	0.1867			
	C(1)																0	13	0.0					
W	T(22)	1					1			3	3		4	4			16	242	6.6	8.638	0.0033			
	C(24)										1		2		1		4	264	1.9					
Violation %		4.0	1.7	6.8	1.4	0.0	0.7	8.2	2.5	10.5	0.0	6.0	1.0	10.0	3.5	2.5								
Overall Violations, CPA Exam vs. Test Banks																					<u>40.219</u>		<u>0.0000</u>	

*The two omitted types (C and X) included a total of four MCQs, containing no violations.

Computations of "violation %" by question include these in the denominator where appropriate.

†(Number of questions) x (relevant rules per question)