A Consumer's Guide to Multiple-Choice Item Formats That Measure Complex Cognitive Outcomes

(or What to Use When Bad Things Happen to Good Certification Testing Programs)

Ronald A. Berk

WARNING:

NES has determined that this material is **HIGHLY FLAMMABLE.**

DO NOT use near forest fires or blow torch.

CAUTION:

FOR EXTERNAL USE ONLY

AVOID ingestion, inhalation, and prolonged contact with skin.

Over the past couple of years a tidal wave of public opinion, political resistance, and/or professional criticism washed away several state-initiated performance assessment programs. In a few cases, states may have proceeded too quickly with the use of constructed-response or portfolio measures; in others, "performance assessment phobia" thwarted the states' efforts to capitalize on the latest testing technology. The symptoms of PA-phobia are typically manifested in the fear of unstandardized, subjectively-scored tests thrust upon local districts by

Ronald A. Berk is a Professor in the School of Nursing at The Johns Hopkins University.

federal law and/or state officials. A central concern in all states is the cost of development, administration, and/or scoring of performance assessment programs. Whatever the objection to performance assessment, politicians in a few states have stepped on the brakes firmly and brought the use of constructed-response formats to a screeching halt.

The decisions of governors and legislatures to return to the good ol' days of standardized multiple-choice tests seem irrevocable. When these decisions take effect, the circumstances surrounding those decisions become moot. We need to shift gears and refocus our methodology on multiple-choice item formats that can measure as closely as possible the competencies or outcomes for teacher and administrator certifications.

Certainly the transition from the new performance-assessment item formats back to multiple-choice item formats is not easy, especially considering the complex cognitive outcomes of most teacher and administrator assessment programs. Like most professionals, I try to keep up with the scholarly research and literature in my field. Representing an academic institution, I thought it would be appropriate to search for a meaningful quotation from the literature that may capture the feelings that some of you may be having or possibly will be having during the transition phase. I finally found one. According to Dr. Seuss (Geisel & Geisel, 1990):

You'll be on your way up!
You'll be seeing great sights!
You'll join the high flyers who soar to great heights.
You won't lag behind, because you'll have the speed.
You'll pass the whole gang and you'll soon take the lead.
Wherever you fly, you'll be the best of the best.
Wherever you go, you will top all the rest.
Except when you don't
Because, sometimes, you won't.
I'm sorry to say so
but, sadly, it's true
that bang-ups
and hang-ups
can happen to you.
You can get all hung up
in a prickle-ly perch.
And your gang will fly on.
You'll be left in a Lurch.
You'll come down from the Lurch
with an unpleasant bump.
And the chances are, then,
that you'll be in a Slump.
And when you're in a Slump,
you're not in for much fun.
Unslumping yourself is not easily done.

How prophetic Dr. Seuss was in expressing our ups and downs. The issue is not whether we will experience a "Slump" in our work; the issue is how we will respond to it. Is the response irritation, anger, or frustration and murmuring and grumbling about what we cannot do? Or, do we glide smoothly through the transition period concentrating on what we CAN do? Our attention has to be focused on the latter. We must do our very best under what we might perceive are the worst conditions.

The purpose of this chapter is to evaluate the utility of multiple-choice items for measuring the complex cognitive outcomes of most teacher and administrator certification programs. Multiple-choice items? Yes. I know you think I am trying to toy with your gag reflex. Or, you are probably thinking, "What? Ron, Heeellooo! Wake up and smell the test items! You can't use multiple-choice items for those outcomes! You have to use constructed-response or some other performance-assessment format." This issue of whether multiple-choice items can measure complex outcomes brings to mind the advice of an old farmer who said, "Never try to teach a pig to sing. It wastes your time and annoys the pig!" Just as the pig in the movie *Babe* questions the validity of this advice, I question the validity of the notion that multiple-choice items can measure only low-level cognitive processes. The polemics surrounding this issue continue to rage between the critics and defenders of multiple-choice and constructed-response items. Snow (1993) has summarized the key arguments:

> Critics of conventional standardized testing often argue that multiple-choice format requires only a superficial, list-of-isolated-facts kind of knowledge structure, and promotes learning and instruction of this same sort, inasmuch as teachers inevitably teach to the test. Such tests are also open to invalidation due to various forms of test-wiseness, social bias, and the ease of cheating. Constructed-response format, on the other hand, involves deeper understanding and higher order, critical thinking, and thus promotes learning and instruction aimed at these higher goals. It is also fairer, because instruction and assessment are in this way more closely connected, and even potentially integrated.
>
> Defenders of conventional standardized testing argue that multiple-choice format can be used to assess many, if not all, deeper or higher order goals of learning and instruction; the two types of tests are usually highly correlated anyway, so multiple-choice tests are entirely adequate as well as much more economical for many educational purposes. Such tests are also fairer: Major forms of bias can be detected objectively and eliminated. Constructed-response format, on the other hand, does not necessarily assess deeper or higher order understanding—it may even promote rote memorization. Furthermore, it is unreliable as well as uneconomical, and

> open to major sources of scoring bias, as well as unwanted correlations with non-instruction-related student characteristics such as intelligence, anxiety, and socioeconomic status. (p. 50)

You are not going to believe this (but I am going to tell you anyway), but there is no definitive evidence that constructed response is superior to multiple choice in assessing higher-order thinking skills. In fact, there is a conspicuous research need for comparative studies of constructed-response versus multiple-choice formats. Snow (1993) has suggested some plausible rival hypotheses to direct this line of inquiry, and Mislevy (1993) has proposed a framework for investigating the differences between the formats.

Although it is assumed that performance assessment formats will get the job done, it may be possible to get better, more cost-effective mileage out of multiple-choice formats used alone or in conjunction with constructed-response items. My task was to investigate the state of the art of the multiple-choice format of yesteryear and to find ways to tweak out its full potential in the context of the certification programs. This translated into reviewing a relatively modest collection of studies, many of which were conducted in the healthcare professions, and the latest variations on the multiple-choice theme. Fortunately, several excellent reviews by Albanese (1993), Frisbie (1992), and Haladyna (1992a, 1992b, 1994) facilitated this process. The result is the following "consumer's guide" to multiple-choice item types. I hope it will be useful for those of you who are shopping for new formats. Maybe it will provide a few fresh insights and tell you something you did not already know about this superstar, the M–C item.

This guide is organized into three sections. The justification for this has been the significant life forces throughout history that have occurred in threes, such as Steve Martin's movie *The Three Amigos,* Dickens's classic work *A Tale of Three Cities,* and, of course, those unforgettable three brothers on *Newhart*—Larry, Darryl, and Darryl. These phenomena can hardly be ignored. So why should this chapter be any different? The sections are as follows: (1) M-C items where the choices are manipulated, (2) M-C items where the stem is manipulated, and (3) M-C item sets based on context-dependent material.

M–C Items Where the Choices are Manipulated

There are several multiple-choice formats that were designed expressly to measure complex cognitive outcomes: (1) complex multiple choice (CMC), (2) Type K, (3) multiple true-false (MTF) or Type X, and (4) multiple response (MR). The components of these item formats, examples, and major advantages and disadvantages are presented in Table 1. What distinguishes these formats are the ways in which the choices are structured and scored. They are all more difficult to answer than comparable conventional M-C items (Type A) due to the choice formats. Despite the intent to assess higher-order thinking skills with these formats, however, they all fall short and exhibit limitations that render them less desirable and more confusing than other formats for measuring complex knowledge.

TABLE 1

M-C Formats that Manipulate the Choices to Measure Complex Cognitive Outcomes (Listed in order of decreasing overall complexity)

Item Format	Components	Example	Advantages	Disadvantages
Complex Multiple Choice (CMC) (Albanese, 1993)	1. Stem 2. List of potentially correct answers (primary responses) 3. List of combinations of primary responses (secondary choices)	**Which of the following behaviors suggest your gray cells have gone on sabbatical (or you're beginning to lose it!)?** (a) You use a match to see better when looking for a gas leak in your house. (b) You try to analyze (as in pick apart) your relationship with your "significant other." (c) You buy your kid a pet because he/she promises to take care of it. (d) You advise your teenager to use his/her own best judgment. (e) You anger your proctologist just before your exam. A. (a) and (b) B. (a) and (e) C. (b), (c), and (d) D. (b), (d), and (e)	1. Permits use of more than one correct answer in primary responses. 2. Facilitates machine scoring of one correct choice from among secondary choices, similar to conventional M-C items. 3. Purportedly measures complex cognitive outcomes.	1. Produces more difficult items with lower discrimination than comparable simple, single-answer M-C items (Haladyna & Downing, 1989; Parker & Somers, 1983). 2. Cluing of secondary choices increases lower-ability students' scores and decreases higher-ability students' scores (Albanese, 1993). 3. Yields lower reliability coefficients than MTF and conventional M-C items (Albanese, Kent, & Whitney, 1977, 1979; Harasym, Norris, & Lorscheider, 1980; Dawson-Saunders, Nungester, & Downing, 1989; Parker & Somers, 1983; Tripp & Tollefson, 1985). 4. Difficult to construct and edit and more likely to contain identifiable flaws that reflect on the content validity. 5. Validity coefficients suggest CMC format does not measure complex cognitive knowledge not measured by other item formats (Dawson-Saunders et al., 1989); limited amount of variance accounted for may be due to lower-ability students' taking advantage of cluing (Albanese, 1993). 6. Requires more time to answer than other formats, thereby reducing item sampling of content (Frisbie & Sweeney, 1982). 7. Takes more space per page, thereby reducing efficiency and increasing cost of product (Haladyna, 1994).

continued on next page

Item Format	Components	Example	Advantages	Disadvantages
Type K (Albanese, 1982, 1993)	1. Stem 2. List of four potentially correct answers (primary responses) 3. Fixed list of five combinations of primary responses (secondary choices)	**Which of the following behaviors suggest your gray cells have gone on sabbatical (or you're beginning to lose it!)?** (a) You try to analyze (as in pick apart) your relationship with your "significant other." (b) You buy your kid a pet because he/she promises to take care of it. (c) You advise your teenager to use his/her own best judgment. (d) You anger your proctologist just before your exam. A. (a), (b), and (c) B. (a) and (c) C. (b) and (d) D. (d) only E. All of the above	Same "Advantages" as CMC.	Same "Disadvantages" as CMC, plus one additional problem: Fixed combinations of responses weight knowledge of primary responses keyed as false higher than knowledge of options keyed as true (Albanese et al., 1977).

continued on next page

Item Format	Components	Example	Advantages	Disadvantages
Multiple True-False (MTF) or Type X (Frisbie, 1992)	1. Stem 2. List of primary correct and incorrect choices answerable as true and false responses	**If you're REEEALLY serious about following your diet, which of these "diet tips" are True (A) and which are False (B)?** 1. ___ If no one sees you eat it, it has no calories. 2. ___ If you drink a diet soda with a candy bar, they cancel each other. 3. ___ The calories from candy or donuts consumed while driving are burned up at the rate of one item per mile. 4. ___ Any food used to relieve stress NEVER counts, e.g., hot chocolate with marshmallows, banana split, strawberry cheesecake. 5. ___ Eating with your fingers while standing up is half the calories as eating with a fork or spoon while sitting down. **According to the "Laws of Psychology," which of these statements are True (A) and which are False (B)?** 1. ___ Never ring a bell when one of Pavlov's dogs is sitting in your lap. 2. ___ The Laws of Behavior Modification only apply to kids in other families. 3. ___ The right hand does know what the left hand is doing; it just doesn't care. 4. ___ Adults get older faster than children, and adults with children get older the fastest.	1. Permits use of more than one correct answer in primary choices. 2. Facilitates machine scoring of each choice. 3. Purports to measure complex cognitive outcomes. 4. Yields higher reliability coefficients than Type K and conventional M-C items (Albanese et al., 1977; Frisbie, 1992; Frisbie & Druva, 1986; Frisbie & Sweeney, 1982; Hill & Woods, 1974; Kreiter & Frisbie, 1989; Mendelson, Hardin, & Canady, 1980). 5. Perceived by students as harder but more effective in measuring their achievement than conventional M-C items (Frisbie & Sweeney, 1982; Kreiter & Frisbie, 1989).	1. Local dependence of item responses can overestimate reliability (Haladyna, 1994). 2. Produces more difficult items than conventional M-C items (Albanese, 1993; Frisbie, 1992). 3. Concurrent validity coefficients with Type K items suggest neither measures complex cognitive knowledge (Albanese et al., 1977), and MTF is interchangeable with conventional M-C items (Albanese & Sabers, 1978; Frisbie & Sweeney, 1982; Hill & Woods, 1974; Kreiter & Frisbie, 1989). 4. Limited to measuring understanding of concepts by listing examples and nonexamples, characteristics and noncharacteristics (Haladyna, 1994). 5. Requires approximately the same time to answer four T-F items in one MTF item as to convert Type K answers to the appropriate secondary choice (Albanese, 1993).

continued on next page

Item Format	Components	Example	Advantages	Disadvantages
Multiple Response (MR)	1. Stem 2. List of primary correct and incorrect choices where more than one answer is selected and incorrect selections are penalized	**Which of the following indicate you're NOT YOUNG anymore?** A. You enter a room and forget what you went in for. B. You've become quite creative at extricating yourself from boring people at parties. C. Your idea of a vacation is to be a couch potato for a week. D. You've gone from a do-it-yourselfer to a hire-someone-elser. E. You're less particular about who you include in your sexual fantasies. **Which bumper sticker(s) would you expect to find on Roseanne's limousine?** A. My Karma Ran Over Your Dogma B. If Money Could Talk, It Would Say: "Bye Bye!" C. It's Lonely At the Top, But You Eat Better! D. I May Be Fat, But You're Ugly, and I Can Lose Weight E. All Men Are Idiots, and I Married the King!	1. Permits use of more than one correct answer in primary choices. 2. Facilitates machine scoring although more than one answer is selected. 3. Measures complex cognitive outcomes in terms of lines of reasoning used by test takers in selecting different answers. 4. Retains the major qualities of conventional M-C items.	1. Produces more difficult items than conventional M-C items (Ryan, 1993). 2. Correction for incorrect selection of answers may be problematic. 3. Sparse research evidence comparing it to other M-C formats.

CMC and Type K Formats

The most recent review of the research on CMC and Type K formats was conducted by Albanese (1993). The long list of disadvantages of using these formats is associated with the evidence of cluing from the overlapping secondary choices (Albanese, Kent, & Whitney, 1977, 1979; Harasym, Norris, & Lorscheider, 1980; Kolstad, Briggs, Bryant, & Kolstad, 1983; Kolstad, Wagner, Kolstad, & Miller, 1983) and its effects on the psychometric characteristics of the items and scores (Albanese, 1982; Case & Downing, 1989; Haladyna, 1992b; Shahabi & Yang, 1990; Subhiyah & Downing, 1993).

MR Format

The MR format is the least studied of any of the formats. It is complicated to score when a penalty is assessed for incorrect answer selection. Following the reasoning for choosing different choices may be best analyzed and tracked using computerized adaptive testing.

MTF Format

The MTF format is a transformation of the CMC format into a true-false mode without the baggage of overlapping secondary choices. The research on its characteristics has been summarized by Frisbie (1992). There are several advantages that prompted Haladyna (1992b) to recommend it as an effective substitute for the CMC format. However, such an endorsement for the use of MTF items in teacher and administrator certification tests seems premature. There are two key issues related to scoring (and reliability) and the utility of MTF for assessing higher-order thinking skills that need to be addressed.

The first issue is the technical concern of local dependence. If there are five choices to be answered for a single item, for example (see Table 1, MTF example 1), each choice may be scored dichotomously (0,1) and treated as an independent item, or the item may be treated as an item cluster with a maximum score of 5, or the item may be scored 1 when all five true-false choices are answered correctly and 0 when any part is incorrect.

In other words, a MTF item can be scored as five separate true-false items, as an item set of five true-false items, or as one M-C item. The intractable problem underlying this scoring is the dependency among the five answers with a common stem that gives the appearance of a M-C item set. This local dependence of the MTF item with the potential for response cueing violates the assumption of independence among responses of different M-C items. Although Frisbie and Druva (1986) and Albanese and Sabers (1988) did not find this violation in their test data, Haladyna (1994) cautions that dependency can result in an overestimation of reliability. Perhaps the best strategy to minimize the effect of local dependence is to score the MTF item either as a single M-C item (0,1) or as an item set (Rosenbaum, 1988).

The second issue is whether the MTF format can measure complex skills. Previous experience with this format indicates it may be limited to testing understanding of concepts. There have been no applications demonstrating how it can measure cognitive outcomes requiring analysis, prediction, or evaluation.

The preceding remarks on CMC, Type K, MTF, and MR item formats and the information in Table 1 suggest that there are serious flaws, weaknesses, and limitations in these formats as they have been used to measure complex cognitive outcomes. The research evidence accumulated to date questions the potential of these formats in teacher and administrator testing programs. The formats described in the next two sections may provide more effective alternatives.

Use of Dangerous and/or Fatal Distractors

Table 1 illustrated how choices can be arranged in a variety of formats to elicit different types of responses. The structure and scoring of the choices were manipulated to create difficult items. Indeed, content must also be considered in constructing any set of choices. Most recently, the use of "dangerous" choices represents an approach that is idiosyncratic to licensure and certification tests. Essentially the notion consists of offering a menu of choices in M-C items that designate "bad decisions" with harmful effects on the clientele being served.

This application of dangerous answers originated in healthcare testing programs (Skakun & Gartner, 1990; Slogoff & Hughes, 1987). For example, a medical or nursing student who chooses dangerous or even "fatal" distractors on a series of items measuring decision-making outcomes can assist the professions in identifying possibly incompetent practitioners. Since the actions in these items can cause harm to patients, and in some cases kill them, those choices seem valid on licensure and certification examinations intended to protect the public from incompetents.

This practice suggests intriguing possibilities for teacher and administrator certification tests. Certain choices on any M-C item calling for an action or decision can be written to reflect the serious consequences that could result from inaction or inappropriate/incorrect decisions. These consequences may be harmful to students, parents, teachers, administrators, the school, the school board, or the school district. For example, consider the consequences of selecting the five choices in the following item.

> Mr. Vinnie Barbarino is a fifth-grade teacher at Sweathog Elementary School in Brooklyn. He suspected his principal, Guido Sardino, was involved in drug dealing within the school. This was confirmed when Guido invited Vinnie to dine with him at a local restaurant and asked him to assist in the drug distribution. Vinnie said, "No, not on your life!" to which Guido responded, "That wasn't the answer I had in mind."
>
> What action should Vinnie now take?
>
> A. Prepare a memo to the school board describing the whole story.
>
> B. Contact NYPD Blue and have Guido arrested for possession and distribution.
>
> C. Wear a bulletproof vest to school from now on.
>
> D. Avoid taking long walks near the East River.
>
> E. Ask his wife to start his car for him every morning.

A distractor analysis of the responses of candidates to such items may be very informative not only about what they know but about what kinds of decisions they may make as practitioners. A study that correlates their selection of dangerous distractors to their pass-fail status on the entire test may also be of interest. One should expect passing candidates to choose fewer dangerous distractors than failing candidates.

M-C Items Where the Stem is Manipulated

The stimulus element of any item format can direct the examinee to a simple recall of facts or to higher levels of cognition. In a M-C item, there are key "operative" words in the stem that can elicit these different levels of thought processes to arrive at a particular answer, such as "most effective," "consequence of," "best explanation," "most appropriate action." Although there is no evidence that explains how an examinee decides on the correct choice, test developers can approximate what the examinee is thinking in response to those operative words. This approximation of the mental processes an examinee must exhibit is the "acid test" of whether an item measures recall or higher level behaviors.

Among the various categories of complex cognitive behaviors identified in "cognitive" taxonomies (Bloom, Engelhart, Furst, Hill, & Krathwohl, 1956; Gagne, 1968; Royer, Cisero, & Carlo, 1993) and typologies for writing test items (Miller & Williams, 1973; Sanders, 1966; Williams & Haladyna, 1982), the two that seem most relevant to the outcomes addressed by teacher and administrator certification tests are prediction and evaluation. These higher-order thinking skills require examinees to predict the consequences of a particular action or situation based on certain principles and to evaluate the most effective, appropriate, or best course of action according to one or more criteria. The latter emphasizes decision-making abilities.

The simplest approach to measure these behaviors in M-C format is to manipulate the stem of the conventional stem-and-four-or-five-choice item to pose prediction or evaluation problems. Several generic stems containing key operative words have been recommended by Haladyna (1994). A few of these stems are listed in Table 2 along with illustrative M-C examples.

TABLE 2

M-C Formats that Manipulate the Stem to Measure Complex Cognitive Outcomes

Type of Outcome	Generic Stem	Example
Prediction	If..., then what happens?	**If you choose to enjoy one of the frequently advertised "Antarctic Adventure Cruises," what would you expect to happen?** A. Pay lots of money to freeze off your keester on a ship. B. Experience the thrill of seasickness at 40° F. below zero as fellow passengers watch you heave your guts out while clutching an ice-covered railing. C. Steer around humongous ice cubes to get a glimpse of animals you can see in your local zoo. D. Watch scores of red-jacketed tourists slipping, sliding, and falling on penguin guano while chasing the little critters to get a picture.
	What will probably happen?	**Hitchhiking is a fairly simple process: (1) find a nice highway, (2) stick out your thumb, and (3) get into any vehicle that stops. Should you have the IQ of a cornflake and elect to hitchhike, what will probably happen?** A. You'll reach your destination and discover you arrived two days late. B. You'll get to know the driver and will eventually be adopted by his/her family. C. You'll soon find FBI agents Scully and Mulder chasing you down because your driver is an extraterrestrial. D. You'll wind up mangled and beaten in a backwoods ditch three states over.
	What would happen if...? (Posing a hypothetical situation)	**If Billy Crystal replaced Peter Jennings on the *ABC Evening News*, what newsworthy event would be reported?** A. At the Academy Awards, Shirley MacLaine received the "Previous Lifetime Achievement Award." B. *The New England Journal of Medicine* reports scientific evidence that a runaway genetic experiment is responsible for Americans' interest in line dancing. C. The captain of the Exxon *Valdez*, responsible for the Alaskan oil spill, was chosen by the U.S. Naval Academy to deliver this year's commencement address. D. A nutritionist's overemphasis on eating "dietary fiber" contributed to a 100-year-old patient's passing wicker furniture.
	What is the cause of...? (Cause-effect relationships)	**Which computer virus can cause a total institutional shut down?** A. AT&T Virus: Every three minutes it tells you what great service you're getting. B. MCI Virus: Every three minutes it reminds you that you're paying too much for the AT&T virus. C. PBS Virus: Your PC stops every few minutes to ask for money. D. Star Trek Virus: Invades your system in places where no virus has gone before. E. Health Care Virus: Tests your system for a day, finds nothing wrong, then sends you a bill for $4,500.

continued on next page

Type of Outcome	Generic Stem	Example
Evaluation and Decision Making	Which is the most or least important, significant...?	**Which proverb has the MOST SIGNIFICANT meaning to a test director?** A. Many are called, but few are at home! B. Where there's a will, there's a won't! C. In God we trust, all others need data! D. To err is human, but it feels divine!
	Which of...is the most or least effective...?	**Which one of these "freebies" is the LEAST COST EFFECTIVE for commercial airlines to continue?** A. Complimentary prewaxed headphones to listen to "Wayne Newton Sings Conway Twitty's Greatest Hits." B. Complimentary movie that is a critically acclaimed blockbuster you and everyone around you have never heard of. C. Complimentary snack consisting of an unrecognizable, disgusting mound of crusty guck. D. Complimentary magazine featuring "Forbes 500's Top 10 Snottiest CEOs" and "Motivational Wall Plaques for Your Pathetic Employees."
	Which action, decision, procedure...demonstrates, illustrates...?	**Which one of the following actions demonstrates one of the "Seven Habits of Highly INEFFECTIVE Teachers?"** A. If you forget your principal's name, just say, "Hey, Bucko!" or "Hey, Hon!" B. When asked to give a presentation at a faculty meeting, do it in mime. C. When a colleague makes an insulting remark, say, "Isn't that spray paint on your bald spot?" D. If you are constantly harassed by a parent, just say, "Your son needs to be smacked and so do you!"
	What is the most or least appropriate action, response...?	**Based on the extensive training in child psychology and child-rearing practices many parents now receive, what is the MOST APPROPRIATE response in the following situation?** **Parent says:** "Please pick up those newspapers." **Child says**: "Eat dirt!" **Parent's response**: A. Kick the kid to "kingdom come" (at least for the weekend). B. Whup the kid until he/she picks up the newspapers. C. Try reasoning with the kid (Yeah, right!). D. Eat the nearest pot of dirt (Why not just gag me with a shovel!).

continued on next page

Type of Outcome	Generic Stem	Example
Evaluation and Decision Making	What is the best or worst step, procedure...?	**What is the WORST "hint" for preparing to take any teacher certification test?** A. Block out all of your experiences as a teacher and knowledge from articles and books on educational issues you may have read; instead, concentrate on your vacation fantasies. B. Eat a hearty breakfast, such as a three-egg ranch omelet, steak, bacon, sausage, hash browns, and a liter of coffee. C. Take your mutilated admission ticket, borrowed driver's license, and one dull no. 3 pencil without an eraser to the test center. D. Set your alarm to get up five minutes before the exam begins and don't look at the directions to the test center until you're in your car. E. Open the "Preparation Manual" for the first time while driving to the test center and glance at the sample items during stop lights.
	What is the most useful or useless strategy, approach...?	**What is the MOST USEFUL test-taking tip to MINIMIZE your true level of performance?** A. Guess early and frequently; if you're sure your answer is correct, erase it completely and guess the worst of the remaining choices. B. Mark several answers to each item to confuse the scoring so, when in doubt, the testing officials will mark your answer as incorrect. C. Narrow your answer choices in each item to the fewest number (one or two) of incorrect choices; then randomly mark your answer accordingly. D. Make stray marks all over the answer sheet, don't fill in the answer boxes completely, and induce a nosebleed to create red blotches on your previously marked answers. E. Start at the end of the test and work backwards, answering the most difficult sections first, leaving only a few precious seconds to race through the dozens of remaining items.
	What is the best or worst action, decision, advice...?	**If you are one of two people on a malfunctioning airplane with only one parachute, what is your BEST professional advice to survive?** A. If you are an **IRS agent**, confiscate the parachute along with the other person's luggage, wallet, clothes, and gold fillings. B. If you are a **cigarette manufacturer**, it doesn't matter who uses the parachute because studies have shown no relationship between airplane crashes and death. C. If you are a **government bureaucrat**, order the other person to conduct a feasibility study on parachute use in multi-engined aircraft under code-red conditions. D. If you are a **state legislator**, seize the parachute because the other person shouldn't be traveling on these new-fangled, untested, costly flying machines anyway.

Another source for generating stems at the higher levels of cognition is the range of stimuli typically employed in constructed-response (limited and extended) item formats. Since they are expressly designed to tap problem-solving, application, analytical, and evaluation abilities, their operative words can be used to build M-C versions with answer choices. Examples of 20 generic stems adapted from constructed-response item stimuli are presented in Table 3.

TABLE 3
Generic Stems for Measuring
Complex Cognitive Outcomes

1. What would be the most likely effect of . . . ?
2. Which principle can explain . . . ?
3. Which of the following is an example of the principle . . . ?
4. What is the difference between . . . and . . . ?
5. What is the similarity between . . . and . . . ?
6. Which of the following procedures should be used to . . . ?
7. Which of the following is a valid generalization from the data . . . ?
8. What is the major strength (or weakness) of . . . ?
9. What is the major advantage (benefit) or disadvantage (loss) of the following . . . ?
10. Which of the following approaches will lead to . . . ?
11. This situation will probably produce (or result in) . . . ?
12. What is the consequence of . . . ?
13. What is the BEST way to . . . ?
14. How would (?) react to . . . ?
15. What is the most (cost) effective (efficient, complicated) solution to . . . ?
16. This scenario BEST illustrates the principle of . . . .
17. Based on the criteria of . . . , which strategy (plan) is most effective?
18. What is the first step that should be taken to . . . ?
19. What is the BEST explanation (?) could give under these circumstances?
20. What is the most appropriate action to resolve . . . ?

Beyond the stem and choice structure described previously for developing M-C items that measure complex cognitive outcomes, context-dependent material can be added to expand the options using the aforementioned generic stems. This type of material can be tailored to specific problem situations. Procedures for incorporating this material in teacher and administrator certification test items are discussed in the next section.

M-C Item Sets Based On Context-Dependent Material

Context-dependent item formats have been used extensively for more than 50 years to measure higher-order thinking skills such as understanding, critical thinking, reasoning, and problem solving (Wesman, 1971). The conventional format of this item consists of a series of multiple-choice or true-false questions about a single, common piece of information. This information may take the form of written material (such as a reading passage) or visual/pictorial material (such as a table, chart, graph, map, drawing, figure, diagram, picture, photograph, work of art, or cartoon). The former type of material represents the classic "interpretive exercise"; the latter permits questions that tap both interpretation and problem-solving abilities.

Within the past several years, another form of context-dependent material that has been increasing in popularity is the "scenario" or "vignette," which describes a problem-solving situation to which examinees respond (in M-C format) with certain actions or decisions. This form has established a significant track record in professional licensing and certification examinations in medicine, nursing, and other healthcare occupations. Ironically, Swanson, Norman, and Linn (1995) note that "one probable consequence of enthusiasm for performance-based assessment methods in the health professions [over the past 20 years] has been the improvement of multiple-choice tests" (p. 11). Many of the examinations consist almost exclusively of M-C items with vignettes that provide a detailed description of a clinical situation, including history, physical and laboratory findings, and a set of questions about the diagnosis, prognosis, or next step(s) in care. As "low-fidelity" simulations of decision-making

situations in patient care (Swanson & Case, 1993), these items are viewed conceptually as somewhere between the M-C items of the past and performance assessment methods of the present. Despite this experience with this type of M-C format, there has been a sparsity of research on its contribution to the measurement arsenal of item formats (Haladyna, 1992a, 1992b).

The advantages and disadvantages of context-dependent item sets are listed in Table 4. There are few points in those lists that are new (see Gronlund & Linn, 1990). Even the problem of local dependence and the scoring of item sets examined previously with the MTF format in Table 1 has its own history, although solutions to the problem are still being investigated.

TABLE 4
Advantages and Disadvantages of
Context-Dependent M-C Item Sets

Advantages	Disadvantages
1. Measures complex cognitive outcomes encountered in everyday real-life situations. 2. Furnishes common background information required to demonstrate understanding, thinking skills, and problem-solving abilities. 3. Measures outcomes in greater depth and breadth at different levels of cognition with a set or series of related items based on the same information. 4. Measures separate components of problem-solving and decision-making abilities. 5. Facilitates machine scoring by using the conventional M-C format or other variations (see Table 1).	1. Difficult to construct and edit. 2. Requires more time and greater skill to construct compared to simpler M-C formats. 3. Requires considerably more time to read and answer than other formats, thereby reducing item sampling of content. 4. Measures specific diagnostic elements of problem-solving abilities in contrast to holistic integration of those elements in essay item formats. 5. Measures problem-solving outcomes at the recognition level only, not demonstration of actual problem-solving skills. 6. Local dependence or context effects of the item set can reduce reliability (Thissen, Steinberg, & Mooney, 1989) unless scored as a testlet (Sireci, Thissen, & Wainer, 1991; Wainer & Lewis, 1990). 7. Takes more space for stimulus material and set of items per page.

The previous applications of context-dependent item sets (a.k.a. item bundles, superitems, testlets) suggest potential for teacher and administrator certification tests that has yet to be fully realized. Although some states have developed these item sets for licensing and certification exams, more varied stimulus materials remain unexplored. In particular, consider the types of job-related material now being used in constructed-response formats and assessment centers, such as memoranda, letters, notes, transcripts of telephone and face-to-face conversations, board policies, referenda, and laws. Any combination of players may be involved in the interactions depicted or problems posed, from students and parents to teachers, principals, and school boards.

If this variety of material were reformatted into the structure of a context-dependent item set with five to ten M-C items, numerous and diverse complex cognitive outcomes could be assessed. These might include teaching, administrative, interpersonal, and communication skills such as analyzing complex information, reaching logical conclusions, making appropriate decisions, evaluating written communication, resolving conflicts, dealing effectively with people, and communicating appropriately for different audiences (e.g., students, teachers, parents, etc.). "Simulations" of many of these decision-making skills could be created in M-C format that arguably would be "the next best thing to being there."

Conclusions

What can be concluded from this shopping guide to M-C formats for teacher and administrator certification tests? Well, I guess it's time, as Paul Harvey would say, to hear "the rest of the story." Here are my "Top 10 Tips for Using M-C Formats to Measure Complex Cognitive Outcomes":

10. Avoid CMC, Type K, MTF, and MR formats like the bubonic plague.
9. Use available lists of generic stems (Tables 2 and 3) to generate actual stems to measure the outcomes of your certification program.
8. Focus on the "operative" words in the stems and emphasize them as they are presented to examinees, especially when the BEST choice is requested.
7. Based on the match of stems to outcomes, estimate how many can be measured by M-C formats and how many require performance assessment formats, such as oral communication, in-baskets, fact-finding exercises, and portfolios.
6. Use the simplest, conventional M-C format with context-dependent material, where appropriate, to flesh out the item content.
5. Select realistic, job-related, context-dependent material that teachers and administrators would actually encounter.
4. Incorporate dangerous and even fatal distractors into the answer choices where important decision-making skills are being assessed.
3. Follow all of the conventional rules of M-C item construction and find an "anal retentive" colleague to edit the items.
2. Score item sets of five to ten items as testlets, where possible, to minimize the problem of local dependence.
1. Step out of your comfort zone and take a few risks so that "wherever you fly, you'll be the best of the best. Wherever you go, you will top all the rest."

References

Albanese, M. A. (1982). Multiple-choice items with combinations of correct responses: A further look at the Type K format. *Evaluation & the Health Professions, 5*(2), 218–228.

Albanese, M. A. (1993). Type K and other complex multiple-choice items: An analysis of research and item properties. *Educational Measurement: Issues and Practice, 12*(1), 28–33.

Albanese, M. A., Kent, T., & Whitney, D. (1977). A comparison of the difficulty, reliability, and validity of complex multiple-choice, multiple-response, and multiple true-false items. *Proceedings from the Sixteenth Annual Conference on Research in Medical Education* (pp. 105–110). Washington, DC: Association of American Medical Colleges.

Albanese, M. A., Kent, T., & Whitney, D. (1979). Cluing in multiple-choice test items with combinations of correct responses. *Journal of Medical Education, 54,* 948–950.

Albanese, M. A., & Sabers, D. L. (1978, March). *Multiple response vs. multiple true-false scoring: A comparison of reliability and validity.* Paper presented at the annual meeting of the National Council on Measurement in Education, Toronto.

Albanese, M. A., & Sabers, D. L. (1988). Multiple true-false items: A study of interitem correlations, scoring alternatives, and reliability estimation. *Journal of Educational Measurement, 25,* 111–123.

Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). *Taxonomy of Educational Objectives: The Cognitive Domain.* New York: Longmans, Green.

Case, S., & Downing, S. (1989). Performance of various multiple-choice item types on medical specialty examinations: Types A, B, C, K, and X. *Proceedings from the Twenty-Eighth Annual Conference on Research in Medical Education* (pp. 167–172). Washington, DC: Association of American Medical Colleges.

Dawson-Saunders, B., Nungester, R., & Downing, S. (1989). A comparison of single best answer multiple-choice items (A-type) and complex multiple-choice items (K-type). *Proceedings from the Twenty-Eighth Annual Conference on Research in Medical Education* (pp. 161–166). Washington, DC: Association of American Medical Colleges.

Frisbie, D. A. (1992). The multiple true-false item format: A status review. *Educational Measurement: Issues and Practice, 11*(4), 21–26.

Frisbie, D. A., & Druva, C. A. (1986). Estimating the reliability of multiple-choice true-false tests. *Journal of Educational Measurement, 23,* 99–106.

Frisbie, D. A., & Sweeney, D. C. (1982). The relative merits of multiple true-false achievement tests. *Journal of Educational Measurement, 19,* 29–35.

Gagne, R. M. (1968). Learning hierarchies. *Educational Psychologist, 6,* 1–9.

Geisel, T. S., & Geisel, A. S. (1990). *Oh, the Places You'll Go! By Dr. Seuss* (pp. 11–19). New York: Random House.

Gronlund, N. E., & Linn, R. L. (1990). *Measurement and evaluation in teaching* (6th ed.). New York: Macmillan.

Haladyna, T. M. (1992a). Context-dependent item sets. *Education Measurement: Issues and Practice, 11*(1), 21–25.

Haladyna, T. M. (1992b). The effectiveness of several multiple-choice formats. *Applied Measurement in Education, 5,* 73–88.

Haladyna, T. M. (1994). *Developing and validating multiple-choice test items.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Haladyna, T. M., & Downing, S. M. (1989). The validity of a taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education, 1,* 51–78.

Harasym, P., Norris, D., & Lorscheider, F. (1980). Evaluating student multiple-choice responses: Effects of coded and free formats. *Evaluation & the Health Professions, 3*(1), 63–84.

Hill, G. C., & Woods, G. T. (1974). Multiple true-false questions. *Education in Chemistry, 11,* 86–87.

Hubbard, J. P. (1978). *Measuring medical education: The tests and experience of the National Board of Medical Examiners* (2nd ed.). Philadelphia: Lea and Febiger.

Kolstad, R., Briggs, L., Bryant, B., & Kolstad, R. (1983). Complex multiple-choice items fail to measure achievement. *Journal of Research and Development in Education, 17*(1), 8–11.

Kolstad, R., Wagner, M., Kolstad, R., & Miller, E. (1983). The failure of distractors on complex multiple-choice items to prevent guessing. *Educational Research Quarterly, 8*(2), 44–50.

Kreiter, C. D., & Frisbie, D. A. (1989). Effectiveness of multiple true-false items. *Applied Measurement in Education, 2,* 207–216.

Mendelson, M. A., Hardin, J. H., & Canady, S. D. (1980, April). *The effect of format on the difficulty of multiple-completion test items.* Paper presented at the annual meeting of the National Council on Measurement in Education, Boston.

Miller, H. G., & Williams, R. G. (1973). Constructing higher level multiple choice questions covering factual content. *Educational Technology, 13*(5), 39–42.

Mislevy, R. J. (1993). A framework for studying differences between multiple-choice and free-response test items. In R. E. Bennett & W. C. Ward (Eds.) *Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment* (pp. 75–106). Hillsdale, NJ: Lawrence Erlbaum Associates.

Parker, C., & Somers, J. (1983, December). *A comparison of the difficulty and reliability of type K and best response test items.* Paper presented at the Iowa Evaluation and Research Association Conference, Des Moines, IA.

Rosenbaum, P. R. (1988). Item bundles. *Psychometrika, 53,* 349–359.

Royer, J. M., Cisero, C. A., & Carlo, M. S. (1993). Techniques and procedures for assessing cognitive skills. *Review of Educational Research, 63,* 201–243.

Ryan, K. E. (1993, April). *A comparison of the single right answer format and the multiple answer format with one correct answer: Is there a one right answer mentality?* Paper presented at the annual meeting of the National Council on Measurement in Education, Atlanta.

Sanders, N. M. (1966). *Classroom Questions: What Kinds?* New York: Harper & Row.

Shahabi, S., & Yang, L. (1990, April). *A comparison between two variations of multiple-choice items and their effects on difficulty and discrimination values.* Paper presented at the annual meeting of the National Council on Measurement in Education, Boston.

Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement, 28,* 237–247.

Skakun, E. N., & Gartner, D. (1990, April). *The use of deadly, dangerous, and ordinary items on an emergency medical technicians-ambulance registration examination.* Paper presented at the annual meeting of the American Educational Research Association, Boston.

Slogoff, S., & Hughes, F. P. (1987). Validity of scoring "dangerous answers" on a written certification examination. *Journal of Medical Education, 62,* 625–631.

Snow, R. E. (1993). Construct validity and constructed-response tests. In R. E. Bennett & W. C. Ward (Eds.). *Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment* (pp. 45–60). Hillsdale, NJ: Lawrence Erlbaum Associates.

Subhiyah, R. G., & Downing, S. M. (1993, April). *K-type and A-type items: IRT comparisons of psychometric characteristics in a certification examination.* Paper presented at the annual meeting of the National Council on Measurement in Education, Atlanta.

Swanson, D. B., & Case, S. M. (1993). Trends in written assessment: A strongly biased perspective. In R. Harden, I. Hart, & H. Mulholland (Eds.) *Approaches to the Assessment of Clinical Competence: Part I* (pp. 38–53). Norwich, England: Page Brothers.

Swanson, D. B., Norman, G. R., & Linn, R. L. (1995). Performance-based assessment: Lessons from the health professions. *Educational Researcher, 24*(5), 5–11, 35.

Thissen, D., Steinberg, L., & Mooney, J. A. (1989). Trace lines for testlets: A use of multiple-categorical-response models. *Journal of Educational Measurement, 26,* 247–260.

Tripp, A., & Tollefson, N. (1985). Are complex multiple-choice options more difficult and discriminating than conventional multiple-choice options? *Journal of Nursing Education, 24*(3), 92–98.

Wainer, H., & Lewis, C. (1990). Toward a psychometrics for testlets. *Journal of Educational Measurement, 27,* 1–14.

Wesman, A. G. (1971). Writing the test item. In R. L. Thorndike (Ed.) *Educational measurement* (2nd ed., pp. 81–129). Washington, DC: American Council on Education.

Williams, R. G., & Haladyna, T. M. (1982). Logical Operations for Generating Intended Questions (LOGIQ): A typology for higher level test items. In G. H. Roid & T. M. Haladyna, *A technology for test-item writing* (pp. 161–186). New York: Academic Press.

Acknowledgements

The twisted ideas in this presentation were partially funded by grants from Moisha & Izzy's Bagel Shop (No. SESAME 10-95), Chickens-That-Almost-Crossed-the-Road Café (No. WINGS 'N' THINGS 4U), and Leroy's Bait and Tackle (No. HOOK-LINE & SINKER 27). The rest of the dough was provided by the National Society for the Prevention of Cruelty to Psychometricians.

The author gratefully acknowledges the assistance of Ellen Spies in the preparation of this chapter and especially the tables, although she has requested to be disassociated from my pathetic attempts at humor.