



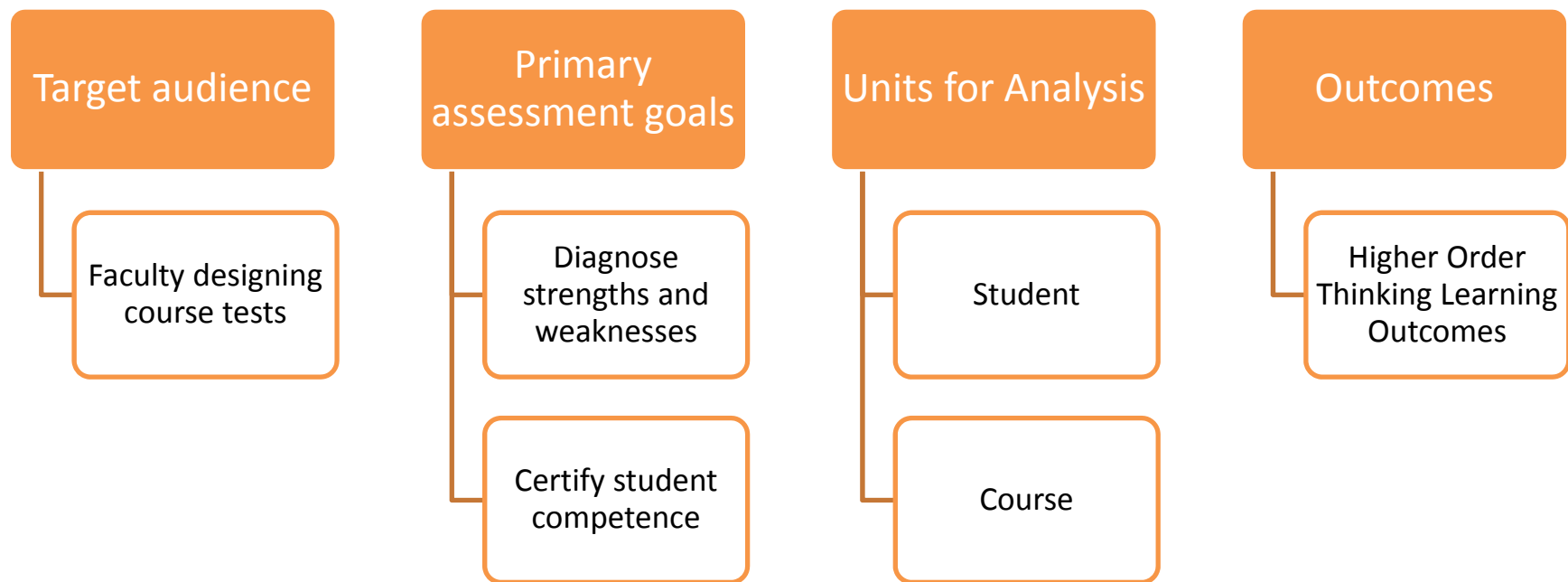
Designing Multiple Choice Tests to Measure Higher Order Thinking

Carol A. Kominski, Ph.D.
Assessment Specialist

Learning Outcomes

1. Apply Bloom's conceptual model to construct higher order learning outcomes.
2. Analyze structured vs. unstructured assessments.
3. Evaluate multiple choice test items for quality and skill level.
4. Construct multiple choice test items to assess higher order thinking.

Assumptions About Workshop Participants



Workshop Agenda

1. Bloom's conceptual Model of Higher Order Thinking

- Learning Outcomes that Assess Higher Order Thinking

2. Structured vs. Unstructured Tests (Pluses and Minuses)

- Types of Structured Tests

3. Analysis of Multiple choice Test Items

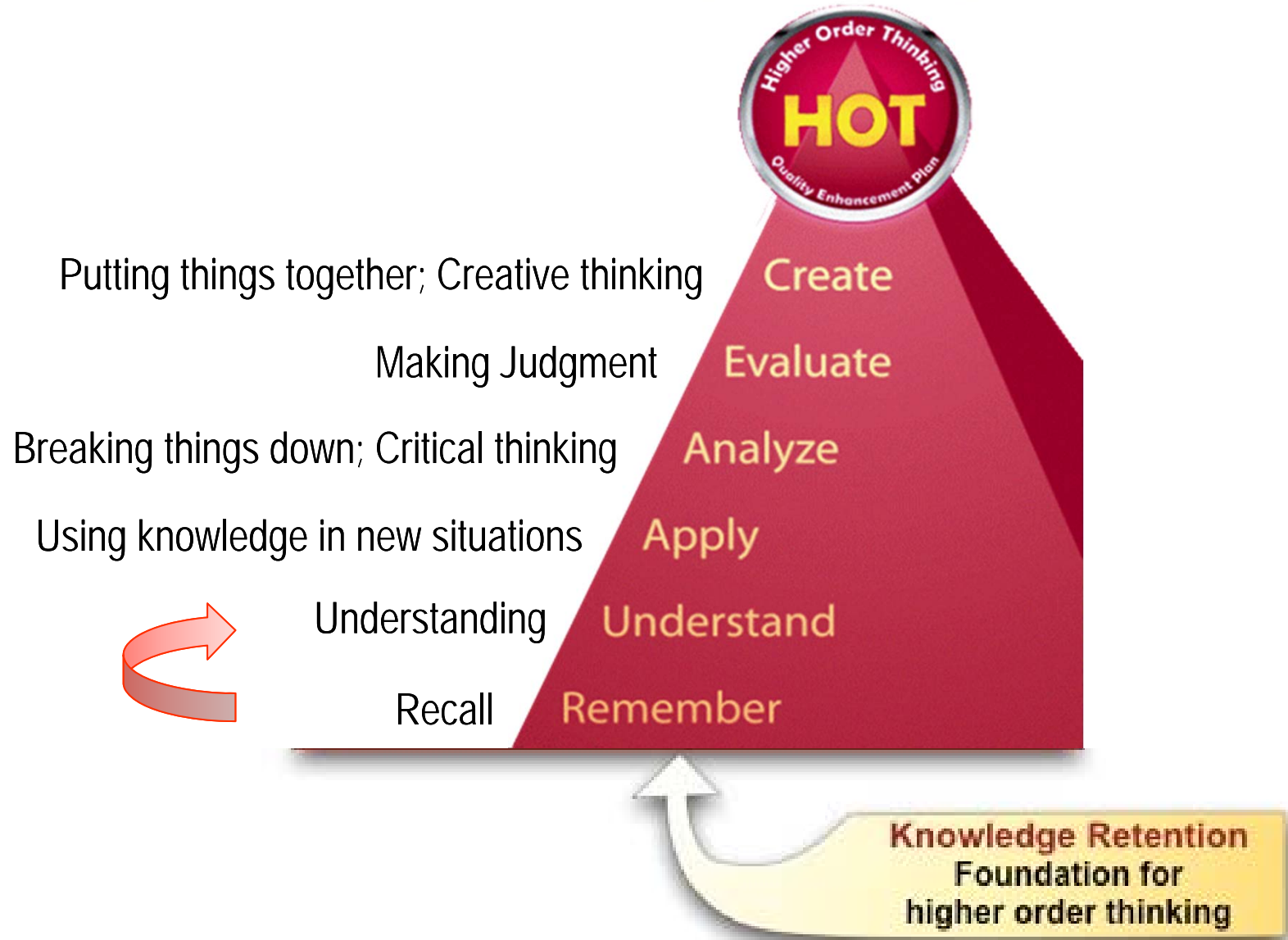
- General Quality Standards
- Level of Thinking

4. Construction of Multiple choice Test Items for Higher Order Thinking

Just what is higher order thinking?



Bloom's taxonomy remix



Learning Outcomes: Questions Should Have “Yes” Answers

Clarity

- Do they specify what students are expected to know and/or be able to do?

Communication

- Are they included in syllabus?
- Are they communicated in course activities?

Relationship to Test

- Can you report test performance on each outcome?

Learning Outcomes that Assess Higher Order Thinking: Key Words

- Useful websites
 - <http://www.celt.iastate.edu/teaching/RevisedBLOOMS1.html>
 - [http://cte.uwaterloo.ca/KSU/Bloom's Taxonomy Cognitive Domain.pdf](http://cte.uwaterloo.ca/KSU/Bloom's%20Taxonomy%20Cognitive%20Domain.pdf)

Definition	Remember	Understand	Apply	Analyze	Evaluate	Create
Bloom's Definition	Remember previously learned information.	Demonstrate an understanding of the facts.	Apply knowledge to actual situations.	Break down objects or ideas into simpler parts and find evidence to support generalizations.	Make and defend judgments based on internal evidence or external criteria.	Compile component ideas into a new whole or propose alternative solutions.
Verbs	<ul style="list-style-type: none"> • Arrange • Define • Describe • Duplicate • Identify • Label • List • Match • Memorize • Name • Order • Outline • Recognize • Relate • Recall • Repeat • Reproduce • Select • State 	<ul style="list-style-type: none"> • Classify • Convert • Defend • Describe • Discuss • Distinguish • Estimate • Explain • Express • Extend • Generalized • Give example • Identify • Indicate • Infer • Locate • Paraphrase • Predict • Recognize • Rewrite • Review • Select • Summarize • Translate 	<ul style="list-style-type: none"> • Apply • Change • Choose • Compute • Demonstrate • Discover • Dramatize • Employ • Illustrate • Interpret • Manipulate • Modify • Operate • Practice • Predict • Prepare • Produce • Relate • Schedule • Show • Sketch • Solve • Use • Write 	<ul style="list-style-type: none"> • Analyze • Appraise • Breakdown • Calculate • Categorize • Compare • Contrast • Criticize • Diagram • Differentiate • Discriminate • Distinguish • Examine • Experiment • Identify • Illustrate • Infer • Model • Outline • Point out • Question • Relate • Select • Separate • Subdivide • Test 	<ul style="list-style-type: none"> • Appraise • Argue • Assess • Attach • Choose • Compare • Conclude • Contrast • Defend • Describe • Discriminate • Estimate • Evaluate • Explain • Judge • Justify • Interpret • Relate • Predict • Rate • Select • Summarize • Support • Value 	<ul style="list-style-type: none"> • Arrange • Assemble • Categorical • Collect • Combine • Comply • Compose • Construct • Create • Design • Develop • Devise • Explain • Formulate • Generate • Plan • Prepare • Rearrange • Reconstruct • Relate • Reorganize • Revise • Rewrite • Set up • Summarize • Synthesize • Tell • Write

Structured vs. Unstructured Tests

Structured Tests

Have limited number of response options. Examples are

True-False

Multiple choice

Matching

Fill-in-the-blanks

Unstructured Tests

Have wider variety of response options controlled by test taker. Examples are

Technical Writing

Oral presentation

Procedural demonstration

Case study analysis

E-Portfolio

Pluses and Minuses for Structured Response Tests

Pluses

Comprehensive knowledge assessed efficiently

Scoring economical and speedy

Moderate to high reliability

Amenable to statistical analysis

Amenable to collection of comparative and trend data

Minuses

Test items laborious to construct

Higher order thinking skills items even more difficult to construct

Impact of cueing, guessing, test savvy, & motivation uncertain

Test security a requirement

Less related to tasks of professional life

Pluses and Minuses for Unstructured Response Tests

Pluses

Higher order thinking more easily assessed

Moderate to high authenticity for “real” life tasks

Requires greater student activity and engagement

Minimal influence of guessing and motivation on performance

Ease of construction

Minuses

Necessity for rubric/scoring key construction & calibration

Scoring requires significant time

Pre-calibration of evaluators needed to increase reliability

More difficult to assess broad range of knowledge quickly

Comparative and trend data harder to collect

Examples of Structured Test Items

Forced choice

- True-false
- Multiple Choice (usually 3-5 choices)
- Matching
 - Allow for use of same options for more than one question
 - Options can be extended (15-20 options)
 - One to one match or unevenly matched lists

Fill-in-the-blanks

- Complete a diagram
- Cloze test for comprehension where every nth word is omitted
- Complete a sentence

Structured Assessments: Focus on Multiple Choice Items

Commonly used

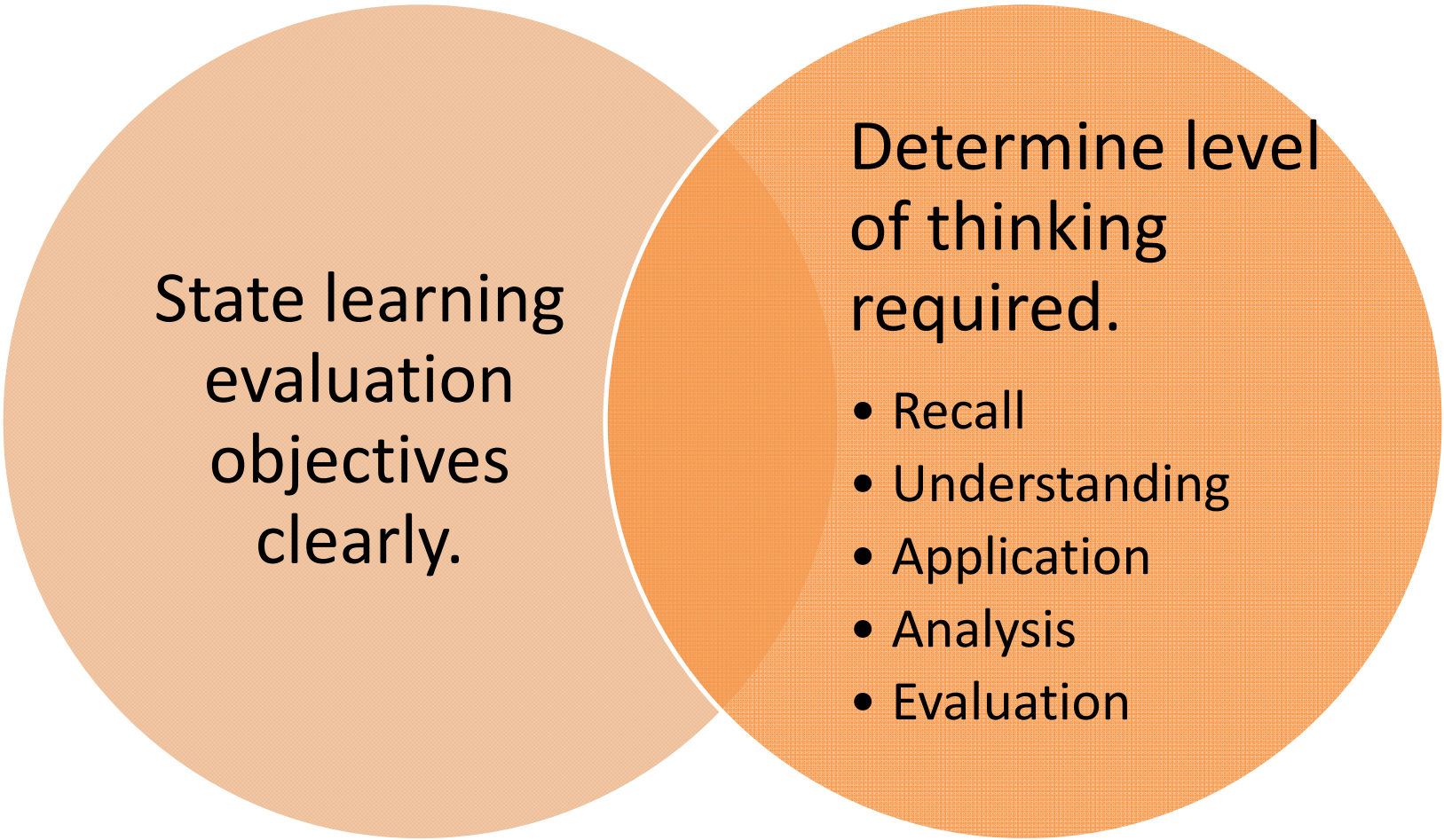
- Large classes
- High stakes testing
 - Admission to professional schools
 - Professional licensure



Item analysis

- Highly developed
- Facilitates systematic item improvement

Multiple Choice Items: Basic Guidelines

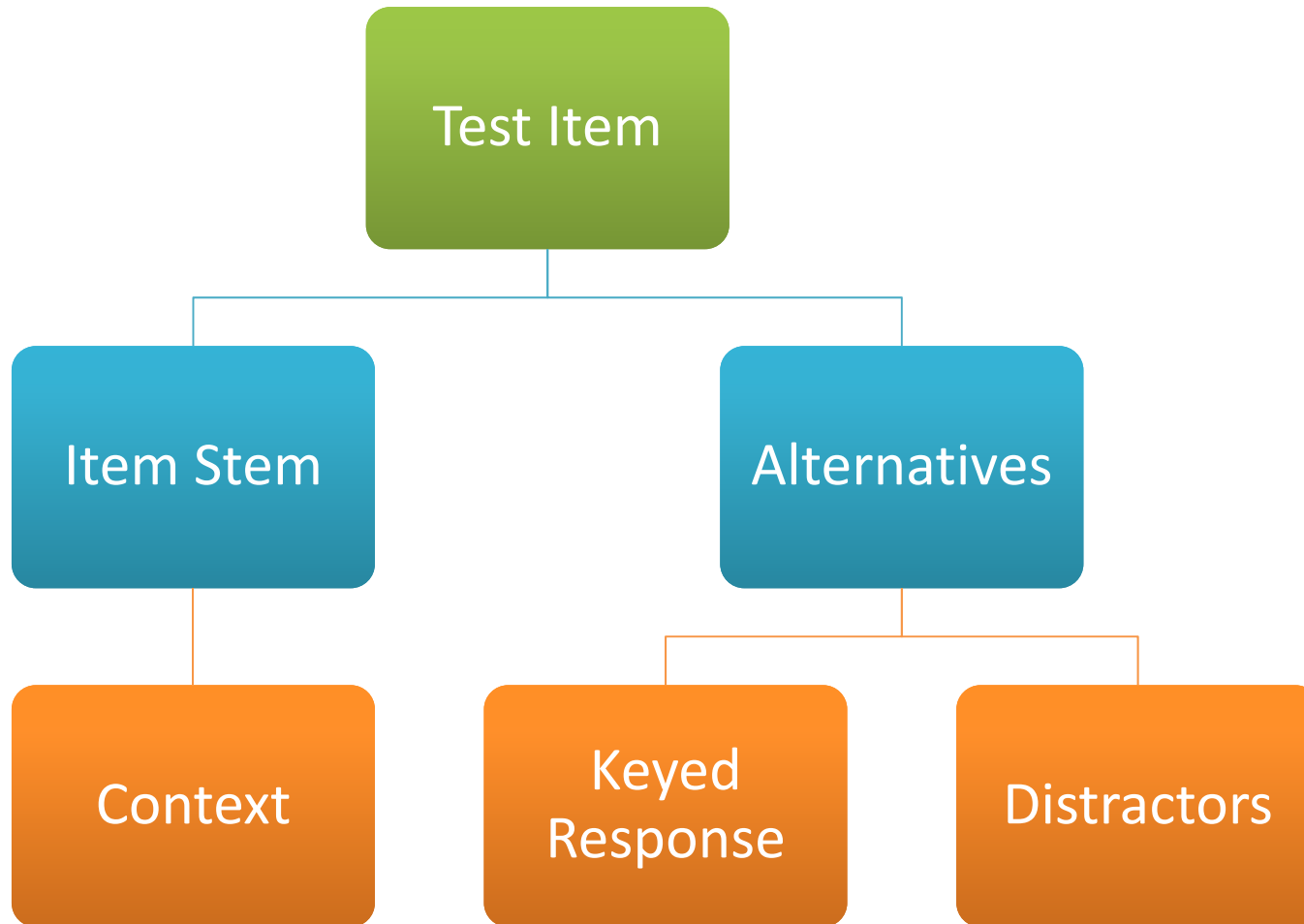


State learning
evaluation
objectives
clearly.

Determine level
of thinking
required.

- Recall
- Understanding
- Application
- Analysis
- Evaluation

Multiple Choice Question Terminology



Context Helpful for Higher Order Thinking Questions

No Context

- Usually not needed for testing of factual knowledge

Context Skeleton

- A small amount may be desirable for testing understanding.

Context

Rich

- Rich context is usually helpful for assessment of higher order thinking skills like application, analysis, and evaluation.

Three Desirable Qualities of Item Stem

1. Succinctness

2. Clear statement of question, problem, or task

3. Positive wording

Six Desirable Qualities of Alternatives

1. Similar lengths

2. Correct grammar

3. One correct answer

4. Absence of extremes like never, always, only

5. No “all of the above.”

6. Mutually exclusive alternatives

Let's Try Some Questions



Question 1: What is wrong with this question?

The way to a man's heart is through his

- a. aorta
- b. pulmonary arteries
- c. pulmonary veins
- d. stomach

Source: Constructing Written Test Questions for the Basic and Clinical Sciences.
Third Edition (Revised). National Board of Medical Examiners, 2002, p.15.

Question 2: What is wrong with this question?

Structured tests

- a. Usually assess higher order thinking.
- b. Are better for large classes.
- c. Do not require a high level of test security.
- d. Requires rubrics or scoring key.
- e. Are easy to construct.
- f. All of the above.

Question 3: What is wrong with this question?

Assume you are a biology professor interested in deciding whether or not team-based learning has a significant impact upon your students. You give half the students a lesson in which you employ team-based learning and the other half a lesson in which you teach using a traditional lecture. After both lessons, you give students a 100 point test to determine how well they have learned the material covered in each class. If you were to do a 2 tailed t test on the students' test results, what is the hypothesis that you are seeking to test?

- a. Students in the team-based learning class will score higher on the test.
- b. Students in both classes will score about the same on the test.
- c. Students in the traditional lecture class will score higher on the test.
- d. All of the above.

Question 3: New and Improved

An instructor teaching half of his students using team-based learning and the other half using traditional lecture gives each group the same test at the end of each class. He performs a 2 tailed t test to compare the two groups. What hypothesis is he testing?

- a. Students in the team-based learning class will score higher.
- b. Students in both classes will score about the same.
- c. Students in the traditional lecture class will score higher.

Question 4: What level of thinking is assessed?

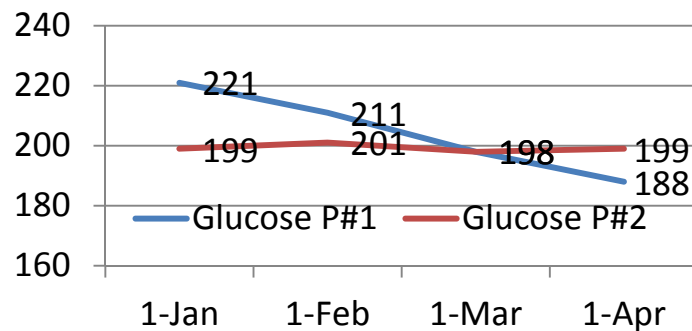
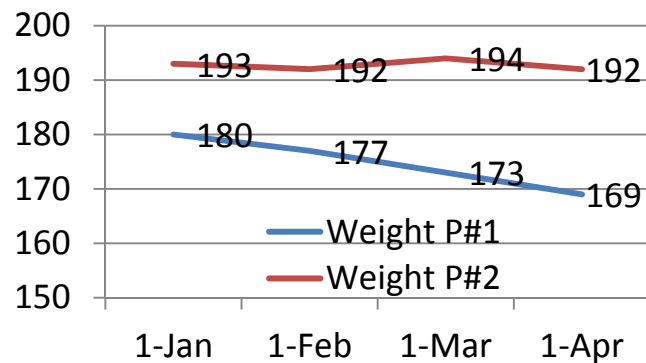
- Which of the following blood tests is used in diagnosis and treatment of diabetes?
 - a. Hemoglobin A1C
 - b. C reactive protein (CRP)
 - c. Antinuclear antibodies (ANA)
 - d. Aspartate aminotransferase (AST)

Question 5: What level of thinking is assessed?

- In a routine physical exam John Smith, age 47, had a blood glucose level of 140 and an A1C level of 4.1%. What is the most plausible explanation of these numbers?
 - a. He has Type I diabetes which is probably controlled by insulin.
 - b. He shows early signs of development of Type II diabetes.
 - c. He probably fasted before his blood glucose test.
 - d. He probably did not fast before his blood glucose test.

Question 6: What level of thinking is assessed?

- Two 60 year old male patients have Type 2 diabetes. Each have a BMI of 27. The primary treatment for each is a diet to reduce blood glucose levels. What is the most likely reason Patient #2 did not show a decline in glucose after three months?



- a. P#1 may have exercised more than P#2.
- b. P#2 probably leads a more sedentary life than P#1.
- c. P#1 lost more weight on the glucose reduction diet.
- d. P#2 may have a more resistant form of diabetes.

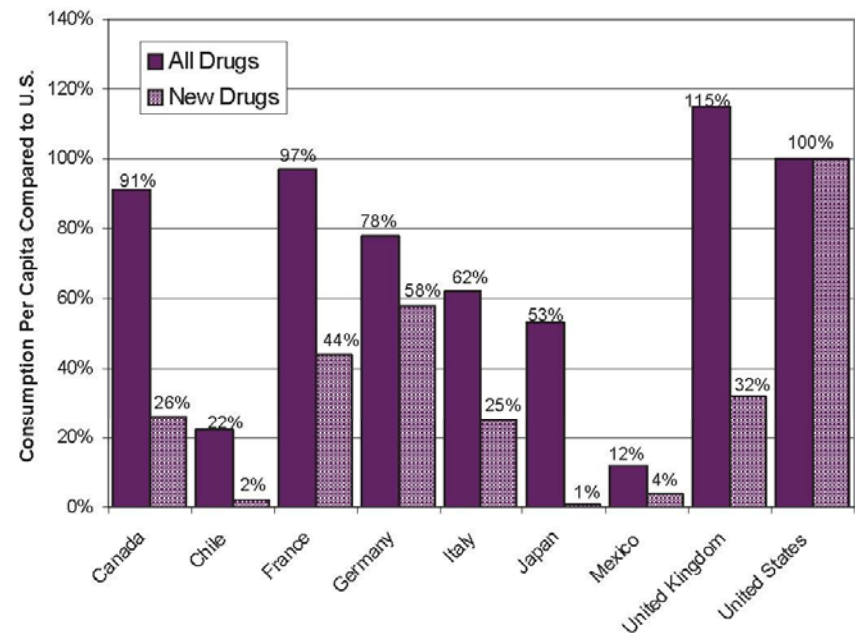
Question 7: What level of thinking is assessed?

Without any other data, which conclusion can you make from reviewing Figure 17?

- a. The average American uses more drugs than citizens of any country except the United Kingdom.
- b. The average Mexican or Chilean consumes fewer drugs than citizens from other countries.
- c. Americans are more likely than residents of other countries to use new drugs.
- d. Japanese have regulations that make it very difficult to obtain new drugs.

CRS-26

Figure 17. International Pharmaceutical Consumption as a Percentage of U.S. Consumption, for 249 Leading U.S. Molecules, 1999



Source: Patricia M. Danzon and Michael F. Furukawa, "Prices and Availability of Pharmaceuticals: Evidence from Nine Countries," *Health Affairs*, Web exclusive, Oct. 29, 2003, pp. W3-521-W3-536, available at [<http://content.healthaffairs.org/cgi/reprint/hlthaff.w3.521v1.pdf>], Exhibit 7.

Notes: "New drugs" are those two years old or newer. From the 350 leading molecules (active ingredients) based on 1999 U.S. sales volume, Danzon and Furukawa chose 249 that were approved in at least four of the study countries or had been approved in the United States since 1992. All products with that active ingredient, including brand-name, generic, and over-the-counter products (if available), and all presentations (capsules, tablets) and strengths in each country were included. Note that consumption of new drugs and all drugs are not the same in the United States, despite their equivalent bars in this chart.

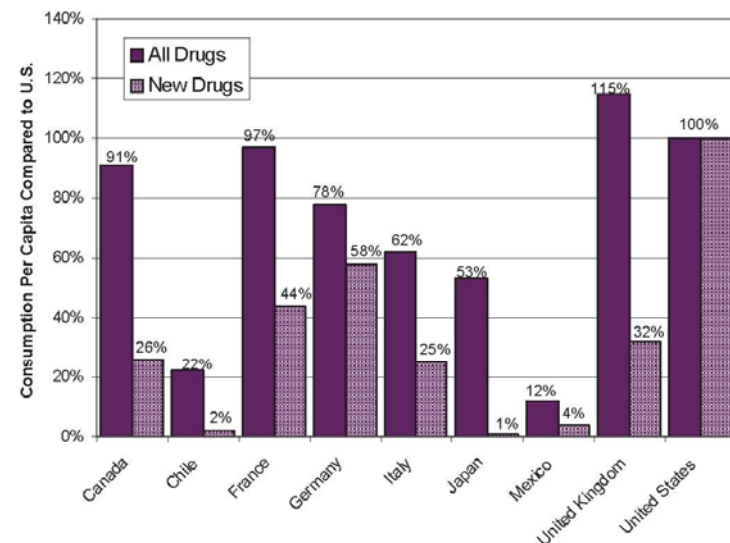
Question 8: What level of thinking is assessed?

What data would be most helpful in estimating average levels of personal drug consumption for the countries identified in Figure 17?

- a. Percent of population in each country buying the covered drugs.
- b. Average cost of new drugs for each country.
- c. Average cost of all drugs for each country.
- d. Population of each country.

CRS-26

Figure 17. International Pharmaceutical Consumption as a Percentage of U.S. Consumption, for 249 Leading U.S. Molecules, 1999



Source: Patricia M. Danzon and Michael F. Furukawa, "Prices and Availability of Pharmaceuticals: Evidence from Nine Countries," *Health Affairs*, Web exclusive, Oct. 29, 2003, pp. W3-521-W3-536, available at [<http://content.healthaffairs.org/cgi/reprint/hlthaff.w3.521v1.pdf>], Exhibit 7.

Notes: "New drugs" are those two years old or newer. From the 350 leading molecules (active ingredients) based on 1999 U.S. sales volume, Danzon and Furukawa chose 249 that were approved in at least four of the study countries or had been approved in the United States since 1992. All products with that active ingredient, including brand-name, generic, and over-the-counter products (if available), and all presentations (capsules, tablets) and strengths in each country were included. Note that consumption of new drugs and all drugs are not the same in the United States, despite their equivalent bars in this chart.

Question 9: What level of thinking is assessed?

- Susan and Clara each want to lose weight. Susan goes on a low carbohydrate diet and Clara goes on a Vegan diet. After six months Susan loses 30 and Clara loses 15 pounds. Relative to losing weight, which of the following conclusions is supported?
 - The low carbohydrate diet is more effective at producing weight loss than the Vegan diet.
 - The Vegan diet contains more calories than the low carbohydrate diet.
 - The low carbohydrate diet is easier to maintain than the Vegan diet.
 - Additional information is needed before making any conclusions.

Characteristics of Multiple Choice Items That Measure Higher Order Thinking

Difficult to construct

- Must develop context

Require lots of context

- Reading selections
- Scenarios, vignettes
- Tables, charts, graphs

Require more testing time

- Reading selections, studying tables and charts
- Thinking itself more complex

Require review by others

- Other faculty, colleagues, or small sample of students

Objectives and Test Items: Dimensions and Guidelines

Dimensions

- Number of learning evaluation outcomes
- Relative importance of each outcome
- Total testing time
 - Higher thinking items require more time

Guidelines

- Minimum two items per objective
- 5-10 items for important learning objectives
 - Additional items increase reliability and validity