Mahatash Esfandiari, UCLA Department of Statistics

Susan Phares, Office of Instructional Development,
UCLA

Testing for upper level thinking in lower division and upper division statistics classes and the role of technology

November 2004

Old Vs. New Trends in Statistics Education

	Old Trends in Statistics Education	New Trends in Statistics Education
1	Teach HOW	Teach HOW and WHY
2	Artificial data sets	Real data sets
3	Focus on numbers	Focus on context
4	Theorem proof example	Collect data, select analysis, fit models
5	Train technicians who produce statistics	Train people who become researchers and consumers of statistics and address real world problems
6	Mostly irrelevant to student discipline	Relevant to students' discipline
7	More emphasis on memorization	Must be able to think critically
8	Emphasis on numerical accuracy than conceptual insight	Emphasis on expressing results in non-technical terms and communicating what numbers mean
9	Emphasis on stepwise formulaic pedagogy	Linking day to day questions to statistical models
10	Dealing with abstract, mechanical and boring problems	Capitalizing on problems with a real context and related to students' discipline
11	Not much use of the computer	Widespread use of the computer

Purposes of the study

- Assessing the extent to which the students in a lower division class (statistics 10) are expected to engage in recall of statistical information, comprehension & interpretation of statistical information, and application, analysis, synthesis, and evaluation of statistical concepts and methods.

- Assessing the extent to which the questions asked on statistics 10 examinations are stated within context and with reference to real world problems.

- Analysis of the type of questions (multiple-choice, true-false, word problems, and calculation problems) asked in the statistics 10 exams by level of challenge, context, and content taught.

Level I: Low level of challenge

Characteristic of Level I questions

- It is a passive process and the student is not expected to change the original information in any form or shape.

- Recall of information including theorems, definitions, methods, procedures using specific facts, conventions, categories, and classifications.

Typical verbs used to make Level I questions

Define, describe, identify, label, list, match, recall, recognize, remember, select

An example of Level I question

A simple random sample of 350 cars from a certain GM factory in Detroit was taken to determine whether horsepower, engine size, or weight has the most to do with miles per gallon. Define the properties of simple random sample. Identify the sample and the population.

Level II: Medium Level of Challenge

Typical characteristics of Level II questions

- It is not a passive process and the student is expected to change the original information
- Understanding the meaning of the concept, translating, and paraphrasing it in one's own words.
- Connecting different topics

Typical verbs used for making knowledge questions

Compare, comprehend, contrast, explain, extend, generalize, interpolate, interpret, paraphrase, summarize, understand.

An example of a Level II question in which the student is expected to rewrite the original material in his own words and within context

After a study was done on a simple random sample of 350 cars in a GM factory in Detroit, the analyst finds the coefficient of correlation between miles per gallon and horsepower, engine size, and weight to be 0.25, 0.37, and 0.69 respectively. Based on these findings, explain which features increase fuel efficiency.

Level III: High level of challenge: Questions that require upper level thinking

Typical characteristics of level III questions

Application: Using old information to solve new problems

Analysis: Taking apart the different components of a complex problem to reach a conclusion(s)

Synthesis: Combining information to build and/or formulate a new structure or design

Evaluation: Assessing and/or judging the validity of the conclusions drawn from a study, analyzing the strengths and weaknesses of a study and offering solutions on how the study can be improved.

Overlap of classifications: The boundaries that separate application, analysis, synthesis and evaluation are not very sharp. Thus, it is not always easy to classify a question into a single category.

Upper level thinking: Given that the boundaries that separate the four classifications are not very sharp, and given that level III questions require critical thinking, we will refer to level III questions as questions that require upper level thinking.

Multiple right answers: Level III questions do not have a single right answer and there is usually more than one way to approach the problem. That is level III questions allow the teacher to enhance the creativity of the students and test them for critical thinking.

One of the best ways to kill creativity is to emphasize the single right answer. (Sternberg, R. J. 1987)

Typical verbs used to make level III questions

Application: apply, construct, develop, implement, relate, use, and utilize.

Analysis: analyze, break down, differentiate, distinguish, explain, infer, separate, support, take apart

Synthesis: build, combine, construct, create, design, devise, form, generate, incorporate, integrate, synthesize

Evaluation: assess, conclude, critique, defend, evaluate, judge, justify, recommend, select, support, validate

Typical level III question:

A newspaper reports that a meditation technique lowered the anxiety of the participants. The experimenter interviewed the subjects and assessed their level of anxiety. The subjects then learned how to meditate and did so regularly for a month. The experimenter re-interviewed them at the end of the month and assessed whether their anxiety level had decreased or not. Is the conclusion warranted from the way the study is designed? If yes, explain why and if not propose a design that would help to assess the effects of meditation on anxiety.

Table 1. Analysis of the total number of questions by the level of challenge

	Level of challenge			Total
	Level I	Level II	Level III	
Number of questions	99	130	15	244
Percentage	41%	53%	6%	100%

Of the total number of questions:

- 41% were at the low level of challenge
- **53% were at the medium level of challenge**
- **6% were at the high level of challenge**

Procedure:

- A total of 18 exams, 12 midterms and 6 finals, were collected from six professors who taught statistics 10 in 2001 to 2003.
- All six professors used Freedman, Pisani, and purves
- Topics discussed included: design of experiments, exploratory data analysis, correlation and regression, probability, chance variability, sampling, chance models, and tests of significance.
- Due to time limitations and the overall goal of statistics 10 (helping students develop a better understanding of how statistics is used and presented in the media and their discipline), it was decided to limit the study to the analysis of questions in experimental design, exploratory data analysis, sampling, and tests of significance.
- The major criteria used for the analysis of the questions included: 1) the type of question, 2) the level of challenge, and 3) whether the question was stated within a real world context or not.
- There were a total of 374 questions in the 18 exams of which 244 were analyzed. Some questions are classified under more than one type and that is why sometimes the sum is more than 244.

Table 2. Analysis of multiple-choice questions by the level of challenge

	Level of challenge			Total
	Level I	Level II	Level III	
Number of questions	38	42	0	80
Percentage	47%	53%	0%	100%

Of the total number of multiple-choice questions:

- 47% were at low level of challenge
- 53% were at medium level of challenge
- 0% were at high level of challenge

Table 3. Analysis of multiple choice questions by level of challenge and context

		Context	No Context	Total
Level I	Number	34	4	38
	Percentage	89.5%	10.5%	100%
Level II	Number	31	11	42
	Percentage	73.8%	26.2%	100%

- Of the multiple choice questions at level 1, 89.5% have a context
- Of the multiple choice questions at level II, 73.8% have a context

Table 4. Analysis of true-false questions by the level of challenge

	Level of challenge			Total
	Level I	Level II	Level III	
Number of questions	28	6	0	34
Percentage	82%	18%	0%	100%

Of the total number of true-false questions:
- 82% were at low level of challenge
- 18% were at medium level of challenge
- 0% were at high level of challenge

Table 5. Analysis of true-false questions by level of challenge and context

		Context	No Context	Total
Level I	Number	8	20	28
	Percentage	29%	71%	100%
Level II	Number	6	0	6
	Percentage	100%	0%	100%

- Of the true-false questions at level 1, 29% have a context
- Of the true-false questions at level II, 100% have a context

Table 6. Analysis of the word problems by the level of challenge

	Level of challenge			Total
	Level I	Level II	Level III	
Number of questions	28	45	15	88
Percentage	32%	51%	17%	100%

Of the total number of word problems:

- 32% were at low level of challenge
- 51% were at medium level of challenge
- 17% were at high level of challenge

Table 7. Analysis of word problems by level of challenge and context

		Context	No Context	Total
Level I	Number	26	2	28
	Percentage	93%	7%	100%
Level II	Number	42	3	45
	Percentage	93%	7%	100%
Level III	Number	15	0	15
	Percentage	100%	0%	100%

- Of the word problems at level 1, 93% have a context
- Of the word problems at level 1I, 93% have a context
- Of the word problems at level 1II, 100% have a context

Analysis of calculation problems

Table 8. Analysis of calculation problems by the level of challenge

	Level of challenge			Total
	Level I	Level II	Level III	
Number of questions	6	58	5	69
Percentage	9%	84%	7%	100%

Of the total number of calculation problems:

- 9% were at low level of challenge
- 84% were at medium level of challenge
- 7% were at high level of challenge

Table 9. Analysis of calculation problems by level of challenge and context

		Context	No Context	Total
Level I	Number	5	1	6
	Percentage	83%	17%	100%
Level II	Number	58	0	58
	Percentage	100%	0%	100%
Level III	Number	5	0	5
	Percentage	100%	0%	100%

- **Of the calculation problems at level 1, 83% have a context**
- **Of the calculation problems at level 1I, 100% have a context**
- **Of the calculation problems at level III, 100% have a context**

Table 10. The type of questions on the "design of experiments" by percentage and level of challenge

Type of question	Level of Challenge Level 1 N(%)	Level II N (%)	Level III N (%)	Total N(%)
Multiple Choice	11 (58%)	8 (42%)	0 (0%)	19 (100%)
True/False	12(100%)	0 (0%)	0 (0%)	12(100%)
Word problems	20 (65%)	7 (22%)	4 (13%)	31(100%)
Calculation problems	0 (0%)	0 (0%)	0 (0%)	0 (0%)
All questions	43 (69%)	15(24%)	4 (7%)	62(100%)

With respect to the design of experiments:

- The use of true/false questions limited the teachers to the lowest level of challenge, such that 100% of the questions were at this level.
- 42% of the multiple-choice questions and 22% of the word problems were at level II.
- Word problems were the only type of question that allowed the teacher to test students at level III. 13% of the word problems were at level III.

Table 11. The type of questions on "exploratory data analysis" by percentage and level of challenge

Type of question	Level of Challenge Level 1 N (%)	Level II N (%)	Level III N (%)	Total N(%)
Multiple Choice	15 (50%)	15 (50%)	0 (0%)	30(100%)
True/False	7(100%)	0 (0%)	0 (0%)	7(100%)
Word problems	1 (6%)	10 (63%)	5 (31%)	16(100%)
Calculation problems	4 (15%)	21 (78%)	2 (7%)	27(00%)
All questions	28 (38%)	41 (55%)	5 (7%)	74(100%)

With respect to question on exploratory data analysis:

- The use of true/false questions limited the teachers to testing the students at the lowest level of challenge, such that 100% of the questions were at this level.
- The the use of word problems, makes it more possible to test students at the highest level of challenge.
- With the exception of true-false questions, 50% to 94% of the multiple choice, word, and calculation problems were at the levels of II and III.

Table 12. The type of questions on "sampling" by percentage and level of challenge

Type of question	Level of Challenge			Total
	Level 1 N (%)	Level II N (%)	Level III N (%)	N(%)
Multiple Choice	15 (50%)	15 (50%)	0 (0%)	30(100%)
True/False	8 (89%)	1 (11%)	0 (0%)	9(100%)
Word problems	6 (30%)	10 (50%)	4 (20%)	20(100%)
Calculation problems	2 (9%)	19 (82%)	2 (9%)	23(00%)
All questions	24 (40%)	32 (53%)	4 (7%)	60(100%)

With respect to questions on sampling:

- The use of true/false questions limits the teachers to testing the students at the lowest level of challenge such that 89% of the questions were at this level.
- The percentage of word problems at level III was the highest. Thus, the use of world problems, makes it more possible to test students at the highest level of challenge.
- With the exception of true/false questions, 50% to 91% of the multiple choice, word, and calculation problems were at the levels of II and III.

Table 13. The type of questions on "hypothesis testing" by percentage and level of challenge

Type of question	Level of Challenge Level 1 N (%)	Level II N (%)	Level III N (%)	Total N(%)
Multiple Choice	4 (27%)	11 (73%)	0 (0%)	15(100%)
True/False	1 (17%)	5 (83%)	0 (0%)	6(100%)
Word problems	1 (5%)	18 (85%)	2 (10%)	20(100%)
Calculation problems	0 (0%)	18 (95%)	1 (5%)	23(00%)
All questions	10 (40%)	86 (53%)	4 (7%)	60(100%)

With respect to questions on "hypothesis testing":

- Contrary to the results obtained in the design of experiments, exploratory data analysis, and sampling, the use of true/false questions on hypothesis testing made it possible to make questions at level II. This could be due to the fact that the questions had a real-world context and the students were asked to explain and defend their answers. So, we may have to include these questions under word problems.

- 73% to 100% of all four types of questions were at levels II and III.

Overall conclusions:

•Except for the true-false questions, 75% to 100% of the questions analyzed were stated in a real world context, Given that exams are a valid reflection of the style of instruction, it can be concluded that we are in compliance with the new trend of using real world problems and real context in the teaching of introductory statistics.

•Overall, 59% of the questions asked were at levels II and III. The percentage of questions at level II and III were 35% to 42% in the design of experiments, 50% to 94% in exploratory data analysis, 50% to 92% in sampling, and 73% to 100% in hypothesis testing. Thus, we are not emphasizing recall of information and we are enticing the students to engage in comprehension and upper level thinking skills. Thus, in this respect we are also in compliance with emphasis on critical thinking.

•Word problems provided the best opportunity for asking questions that required upper level thinking.

•True-false questions limited the teachers to addressing the lowest level of challenge. In design, exploratory data analysis, and sampling 100%, 100%, and 89% of the true-false questions were at the lowest level of challenge. It is suggested to either avoid using true-false questions. If they are used, it is suggested to have the students defend their responses and explain why they chose the answer they did.

Future plans: Using technology and partial credit questions to test for upper level thinking:

- Avoid asking true-false questions that have no real context and do not require the students to defend their answers.
- Given that a high percentage of the questions have a real world context, it is recommended to ask more questions that require the students to engage in upper level thinking.
- Given that grading word problems is very time consuming, we should work toward asking multiple choice questions that branch off and require the students to engage in upper level thinking.
- Ask short word problems at the end of the discussion section and train the teaching assistants to grade written problems.
- Use computers (quiz tool) as a mean of formative evaluation and helping the students play an active role in their own learning and moving toward upper level thinking (An objective of the BICS case study).
- Use computerized testing to help the teaching assistants and the instructors re-teach the concepts and methods that need to be revisited (An objective of the BICS case study).
- Use computers to help the teachers design quizzes and exams that test the students at higher levels of thinking (An objective of the BICS case study).

How we plan to use partial credit questions:

- Present the students with a dilemma or a situation with a real context and the relevant, lots, numbers, printouts, etc.
- Ask the students an open-ended question the correct answer to which involves at least two components.
- Present the students with a number of alternative answers to the question such that
    - one of which is completely wrong,
    - one of which is completely right, and
    - Several alternatives which are partially right.

Generic examples of partial credit questions:

Suppose you pose a question the answer to which involves two correct components, in that sense the alternative answers would be:

- Right, Right (two points)
- Wrong, Wrong (zero points)
- Right, Wrong (one point)
- Wrong, Right (one points)

Suppose you ask a question the answer to which involves three correct components, in that sense the alternative answers would be:

- Right, Right, Right (three points)
- Wrong, Wrong, Wrong (zero points)
- Three combinations of Right, Right, Wrong (two points)
- Three combinations of Wrong, Wrong, Right (one points)

Sample Question 1:

The objective of a study was to examine whether the average SATQ scores of students who were admitted to a particular university was higher than the general population. Suppose that you are hired to analyze the relevant data and report the results to the registrars office. Which of the following options would you pick to explain the findings?

Xbar = 523

Mue = 500

Sigma = 100

N = 225

Answer to sample question 1

I. Since the risk associated with rejecting the true null is more than 5%, we the null hypothesis and conclude that the SAT scores of the students in this u similar to the general population. (WW = 0)

II. Since the confidence interval does no include the hypothesized value un the null, we reject the null and conclude that the SAT scores of the students i university should not be compared to the general population.
(R /W= 1)

Answer to sample question 1

III. On the average the students who are admitted to this university score 2 higher than the general population on SATQ, but, this difference is not statisti significant. (R/W = 1)

IV. We reject the null and we are 95% confident that the students who are admitted to this university score between 10 points to 36 points higher than the general population (R/R = 2)

Sample question 2:
What is the major message conveyed by the comparison of the following tables:

Tests of Between-Subjects Effects

Dependent Variable: posstest on authority

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	6069.915[a]	3	2023.305	16.255	.000
Intercept	3841054.8	1	3841054.8	30858.827	.000
GENDER	5073.352	1	5073.352	40.759	.000
GROUP	956.908	1	956.908	7.688	.006
GENDER * GROUP	473.275	1	473.275	3.802	.051
Error	146254.408	1175	124.472		
Total	4320908.8	1179			
Corrected Total	152324.323	1178			

a. R Squared = .040 (Adjusted R Squared = .037)

Tests of Between-Subjects Effects

Dependent Variable: posstest on authority

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	357.409[a]	3	119.136	1.135	.342
Intercept	216709.003	1	216709.003	2064.430	.000
GENDER	4.148E-05	1	4.148E-05	.000	1.000
GROUP	275.032	1	275.032	2.620	.111
GENDER * GROUP	67.218	1	67.218	.640	.427
Error	6508.314	62	104.973		
Total	239859.694	66			
Corrected Total	6865.724	65			

a. R Squared = .052 (Adjusted R Squared = .006)

Answer to sample question 2:

The comparison of tables A and B indicate that:

a) As the sample size increases, you get results that are more significant and more reliable. So, the recommendation is to use as large of a sample as possible to get wider confidence intervals and thus more confidence in your results. (R,W) = 1 point

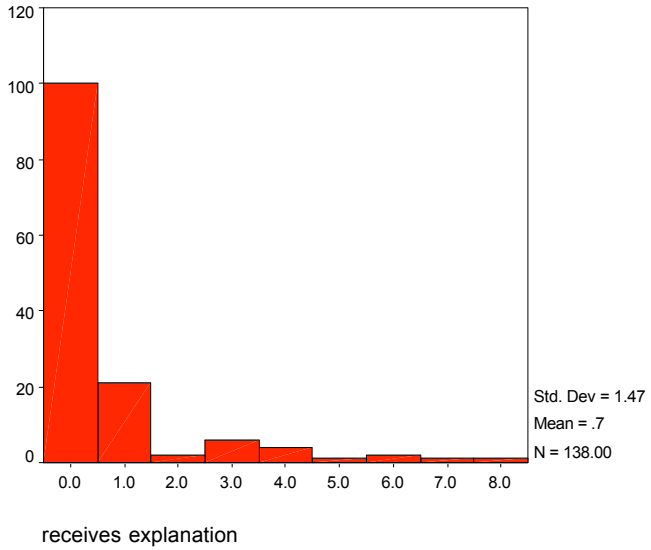
b) As the sample size increases, you are more likely to reject a false null. Thus, the recommendation is to use as large a sample as possible in order to guarantee statistical and practical significance. (R,W) = 1

c) As the sample size increases, error variance decreases, and the risk of rejecting a true null is decreased. In spite of getting statistical significance with large samples, practical significance is not guaranteed. (R,R) = 2

d) In order to reject the null, we need large F, low mean square error, high N, and small P value. That is why choosing as large a sample as possible and establishing statistical significance should be the major focus of any statistician. (R,W) = 1

Sample question 3:

A math teacher uses cooperative groups to teach her class. You are given the histogram and the frequency data for the percentage of explanations received by students from their team mates during group work.



receives explanation

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	100	72.5	72.5	72.5
	1	21	15.2	15.2	87.7
	2	2	1.4	1.4	89.1
	3	6	4.3	4.3	93.5
	4	4	2.9	2.9	96.4
	5	1	.7	.7	97.1
	6	2	1.4	1.4	98.6
	7	1	.7	.7	99.3
	8	1	.7	.7	100.0
	Total	138	100.0	100.0	

Answer to sample question 3

Given the above plots, what is the best conclusion?

a) The graph is negatively skewed and the majority of students (more than 70%) are receiving explanations during group work. (Wrong, Wrong = 0).

b) The graph is positively skewed and the majority of the students (more than 70%) are receiving explanations during group work (Right, Wrong = 1).

c) The graph is negatively skewed and the mean is less than the median of the explanations received during group work. (Wrong, Wrong = 0)

d) The graph is positively skewed and the majority of the students (more than 70%) are not receiving explanations during group work. (Right, Right = 1)