The objective-format question in statistics: Dead horse, old bath water, or overlooked baby?

George W. Cobb
Department of Mathematics, Statistics and Computer Science
Mount Holyoke College
South Hadley, MA 01075
GCobb@MtHolyoke.edu

Invited paper, annual meeting of the
American Educational Research Association,
San Diego, CA April 13, 1998

1. Introduction

My thesis in this paper is pseudo-contrarian. It is contrarian, in that I shall be saying some positive things about objective-format questions at a time when those of my statistician-colleagues who are most active in the reform of statistics education have long ago embraced authentic assessment.¹ It is pseudo-contrarian in that nothing I have to say is intended as a criticism of projects, writing assignments, oral presentations, etc., all of which I use and applaud. No one would argue that when you get right down to it, answering multiple-choice questions comes closer than data-gathering-and-analysis projects can to what it is that statisticians actually do. Nevertheless, I suggest we have been too quick to dismiss the objective-format question as a horse so long dead that only a zealot would still take the trouble to beat him. Of course, it is only in the context of authentic assessment that the horse's obituary has been published. If you consider the new advanced placement test in statistics, taken last spring by more than 7000 high school students, it appears that the objective-format question has merely gone through a near-death experience, only to emerge alive and kicking. Indeed, the economic advantage of an assessment method that makes it possible to assign defensible grades to large numbers of students without requiring a lot of expensive faculty time suggests that objective format questions are here to stay. All the

¹ Indeed, I think the rapid success of authentic assessment in statistics is due, perhaps to a greater extent than we realize, to a happy coincidence. Scholars in education were providing a theoretical frame and research justification for the new assessment at just the time when statisticians in the vanguard of the reform of the teaching of their subject were independently experimenting with, and giving talks about, the practical value of projects, writing, group discussions and the like, as vehicles for learning and assessment.
same, though, the economic advantage derives from features that encourage people who care about teaching to associate such questions with superficial pedagogy (e.g., large lecture classes) and superficial assessment (e.g., machine scoring of standardized tests). Is there any potential educational value to using these questions?

My question for you, myself, and others, is not whether it is easy to write bad objective-format questions -- this is the dead horse hypothesis -- nor whether objective format questions are on balance more limited and less effective than, say, data analysis projects -- another point that I accept without data. Rather, my question is this, “Have we been so distracted by the bath water of bad questions that we haven’t made sufficient effort to check the tub for sign of life? Might there be a baby in there? Is it possible to create objective-format questions that are effective tools for learning and assessing statistical thinking (as opposed to mechanical skills)?” In what follows I offer a qualified Yes. I begin with a discussion of the nature of statistical thinking, as I see it, then introduce five suggestions for increasing the richness and complexity of objective format questions in statistics, and devote the bulk of the paper to examples that illustrate how these suggestions work.

Two disclaimers before I get down to business. First, I am here in the role of a person with experience as a statistician and teacher but one who is comparatively naive about research on teaching and learning. My paper presents no research, either of the library variety, or the kind that deals with live subjects. In a sense, I offer my remarks more as an extended editorial than a piece of conventional scholarship. Second, I am by no means alone in wanting another look at objective format questions. Members of the committee that planned the advanced placement test in statistics\(^2\) include many of the leaders in statistics education, and if these people have worked hard on a test that uses objective format questions, they deserve to be taken seriously.

Moreover, standardized tests are not the only possible use of objective format questions. Setting aside the all-important question of content, these questions offer the advantage that they lend themselves naturally to self-assessment and computer-based learning. They are no substitute for the free give-and-take of the classroom, or the gestalt experience of a project, but once objective-format questions have been written, whatever substance they offer can be provided cheaply and easily, without a teacher’s involvement. One implementation of such

computer-based review and self-assessment is *An Electronic Companion to Statistics*, and the examples I use for illustration are taken from my work on that project. However, the point I wish to make here is that there are many such computer-based guides that rely on objective-format questions -- such questions and computers go naturally together -- but precisely because the technical challenges have been met, it is important to look at content. The value of the questions, if there is any to be found, will depend largely on whether they are helpful in teaching and assessing more than skill with simple algorithms and recall of isolated facts. Is it possible to use such questions for learning and assessing anything that even remotely resembles true statistical thinking?

2. Statistical thinking

Statistics has long been regarded as a branch of mathematics, sometimes by people who see only fancy arithmetic and equate arithmetic with mathematics, and sometimes by mathematicians who think the only substance to statistics is to be found in the theorems and proofs of its deductive structure. Both views are wrong. Statistics is the science of data production and data analysis, and data analysis is an interpretive activity, albeit one that seeks to orient itself within a rigorous deductive framework. To better understand the nature of statistical thinking, one might start with Pascal, who reduced the entire spectrum of the intellectual enterprise to just two pure forms of reasoning, which he called *esprit de geometrie* and *esprit de finesse*:

The geometric, or scientific mind works according to principles that are plain and few, but remote from common life. It may be hard to find these principles and grasp them from the very first, but once they are grasped one would have to be mentally deficient to confuse or misapply them. On the contrary, the subtle mind (*esprit de finesse*) works with principles that reside in the midst of life for everybody to see. But one needs a good

---


5 If mathematicians and statisticians were required to earn their livelihoods in an English department, the statisticians would do textual exegesis, and the mathematicians would do theory of criticism.
eye to detect them because they are so numerous and so fine. It is almost impossible that
some of them should not escape notice.\(^6\)

In statistics, the first kind of thinking corresponds roughly to the continuum that runs from
computational/algorithmic thinking at one extreme to logical/deductive thinking at the other.
(If you accept the values of the mathematicians, you should imagine this axis as vertically
oriented, with logic at the high end and arithmetic at the low end.) In statistics, Pascal's
second kind of thinking involves the so-called "soft" aspects of the subject, dealing with
pattern and interpretation. By interpretation I intend a largely verbal kind of thinking that
starts with patterns and gives them meaning by relating them to their applied context, to the
"story of the data." "Pattern" refers ultimately to patterns in the numbers, but recognizing
patterns in statistics has come to rely heavily on graphics, and more recently on dynamic
graphics, so that I regard the pattern-searching component of statistical thinking as a largely
visual enterprise. Thus, with some admitted oversimplification, I regard Pascal's second
kind of thinking, in statistics, as a second continuum, running from the verbal/interpretive to
the graphical/dynamic. (If you want to do obeisance to the theory of hemispheric
specialization, you can imagine this second axis as oriented horizontally, with the verbal
thinking on the left, and the visual thinking on the right.)

How does all this relate to objective format questions? I suggest that the archetypes
that have led us to dismiss such questions as bad educationally and bad statistically have for
the most part been written by people who misunderstand statistics as confined to the one-
dimensional world of the first axis, and who see a natural affinity between that kind of
thinking and objective-format questions. Objective-format questions

- tend to be (and some would say "ought to be") clear-cut;
- tend to focus (some would say ought to focus) on one feature at a time; and
- tend either to be independent of context, or to involve first a choice that can be made
  independent of context, followed by a rather mechanical translation into an applied
  context.

Interpretive work in statistics is not like that. It

- involves judgment (e.g., how important is the asymmetry in the data?);
- involves simultaneous attention to a variety of features (e.g., the usefulness of a
  summary statistic depends on the shape of the data); and
- involves an interweaving of several different kinds of thinking: logical/deductive,
  computational/algorithmic, visual/dynamic, and verbal interpretive.

The challenge then, is to find ways to put this kind of thinking into a format that has traditionally been associated with a much more mechanical view of statistics. Toward that end, I offer five suggestions.

3. Five principles for objective format questions in statistics

1. Increase the number of choices.
A true-false format offers two choices, multiple choice offers typically four or five, but a matching question offers \( n^2 \) choices if \( n = 4 \), \( 25 \) if \( n = 5 \). It is transparent that more choices will generally make a question harder, which may (but may not) be a virtue. As I see it, the greatest advantage of the matching format goes in a different direction. The two simpler formats can encourage an artificial strategy of ruling out wrong answers. When you answer a multiple-choice question, you end up with one match and three or four answers that don’t fit. When you answer a matching question you end up with a handful of correct matches -- four or five simultaneous and interrelated illustrations of appropriate cognitive connections all rolled into a single answer. Consider, for example, the following two questions.

**Question 1. McClesky versus Georgia.**
In McClesky vs. Georgia, lawyers presented data showing that for convicted murderers, a death sentence was more likely if the victim was white than if the victim was black. For each case, they tabulated the race of the victim and the sentence (death or life in prison). Which of A - E best describes their data?

- A. Response is quantitative; explanatory variable is quantitative.
- B. Response is quantitative; explanatory variable is categorical.
- C. Response is categorical; explanatory variable is quantitative.
- D. Response is categorical; explanatory variable is categorical.
- E. Neither is an explanatory variable.

I consider Question 1 a reasonable “beginner” question. In addition to testing basic concepts (response versus explanatory variables, categorical versus quantitative variables), it asks for a substantive connection between abstract categories and an applied context, and it implicitly calls attention to a 2x2 structure that is useful for thinking about association and relationships between pairs of variables. All the same, a similar question that uses the matching format can be much richer:
Question 2. Employment discrimination.

In studies of employment discrimination, several variables are often relevant: an employee's age, sex, race, years of experience, salary, whether promoted, and whether laid off. Place each question in its proper place in the table that follows them.

A. Are men paid more than women?
B. On average, how much extra salary is a year of experience worth?
C. Are whites more likely than blacks to be promoted?
D. Are older employees more likely to be laid off than younger ones?

<table>
<thead>
<tr>
<th>Response Y</th>
<th>Categorical</th>
<th>Quantitative</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Explanatory X</th>
<th>Categorical</th>
<th>Quantitative</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Question 2 asks you to think simultaneously about all four kinds of two-variable associations, so that, unavoidably, you end up making two sets of comparisons: one set in the context of the application, and a parallel set in abstract language used to label the 2x2 table. Thus Question 2 gives you a kind of cognitive practice that you don't get with the multiple choice Question 1.

The content of Questions 1 and 2 could be addressed by questions written in either of the two formats. For other important kinds of statistical understanding, matching is so natural that it is almost built into the nature of the thinking required. For example to understand the basics of experimental design, you need to learn to move comfortably back and forth between the concrete (and often messy) story of an experiment and the cleaner, abstract, but often-difficult categories of the theory:

Question 3: Carpets.

Hospital floors are usually covered by bare tiles. Carpets would cut down on noise, but might be more likely to harbor germs. To study this possibility, investigators randomly assigned 8 of 16 available hospital rooms to have carpet installed. The others were left bare. Later, air from each room was pumped over a dish of agar. The dish was incubated for a fixed period, and the number of bacteria colonies were counted. Match each item on the left with its category on the right.

A. The 16 hospital rooms 1. Experimental units
B. The 8 rooms left bare  
C. Carpeting and bare floors  
D. Number of colonies in a dish

2. Treatments  
3. Response  
4. Control group

Although Questions 2 and 3 encourage a limited kind of comparative thinking, the basic task is nevertheless one of category matching. Each variable in Question 2 is either a response or explanatory, and is either categorical or quantitative; each feature of the experiment in Question 3 belongs to exactly one of the abstract categories. Much of statistical thinking is not such a simple matter of yes/no, but involves some kind of quantitative comparison.

2. Ask for comparative judgments, not just category matching.

For example, we talk of association between variables being strong or weak, and the language is both meaningful and an important element in understanding quantitative relationships, but “strong” and “weak” are comparative rather than absolute, and don’t lend themselves to category matching. Comparisons are natural, though:

Question 4. Factors related to the death sentence

Which two of the four factors in the tables below are most strongly associated with whether a convicted murderer is sentenced to death?

<table>
<thead>
<tr>
<th>Race of the Defendant</th>
<th>Death Sentence?</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Black</td>
<td>10</td>
<td>24</td>
</tr>
<tr>
<td>White</td>
<td>4</td>
<td>17</td>
</tr>
<tr>
<td>Total</td>
<td>14</td>
<td>41</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Race of the Victim</th>
<th>Death Sentence?</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Black</td>
<td>1</td>
<td>14</td>
</tr>
<tr>
<td>White</td>
<td>13</td>
<td>26</td>
</tr>
<tr>
<td>Total</td>
<td>14</td>
<td>40</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Prior Record</th>
<th>Death Sentence?</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Prior record</td>
<td>12</td>
<td>21</td>
</tr>
<tr>
<td>No record</td>
<td>2</td>
<td>20</td>
</tr>
<tr>
<td>Total</td>
<td>14</td>
<td>41</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Type of Entry</th>
<th>Death Sentence?</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Authorized</td>
<td>6</td>
<td>10</td>
</tr>
<tr>
<td>Unauthorized</td>
<td>8</td>
<td>27</td>
</tr>
<tr>
<td>Total</td>
<td>14</td>
<td>37</td>
</tr>
</tbody>
</table>

Similarly, judging the strength of interaction from a graph is more naturally comparative than absolute.
**Question 5: Interaction**

The plots below summarize two biology experiments designed to measure interaction:

- **“Worms.”** Do hornworms raised on a poor diet get so used to eating extra to compensate that they eat extra even if given a regular diet?
- **“Birds.”** Does a hormone supplement have the same effect on plasma calcium levels in male and female birds?

Which of A - D best fits the pattern?

A. “Worms” shows much stronger interaction than “Birds.”
B. “Birds” shows much stronger interaction than “Worms.”
C. Both show strong patterns of interaction.
D. Neither shows a strong pattern of interaction.

![Box plots for Worms and Birds experiments](image)

There is nothing artificial about the kind of judgment these questions ask for. The thinking is quite typical of one sort that is central to statistical practice. Nevertheless, Questions 4 and 5 each involve just one feature of the data -- the strength of association between the row and column categories, or the strength of an interaction. Good data analysis nearly always requires -- usually sooner rather than later -- that you pay attention to *multiple* features of the data.

3. **Involve several features of the data simultaneously.**

Good data analysis typically starts with a plot designed to show several features at once. Here’s a fairly easy example based on time series data:

*Objective format questions in statistics* 2/22/98 DRAFT 10:40 AM
Question 6. Carbon dioxide levels
Plants convert carbon dioxide into oxygen, and many scientists worry that clearing forests may increase carbon dioxide to unacceptable levels. The graph shows carbon dioxide levels between 1974 and 1991. Which of the following is false?

A. Carbon dioxide levels are increasing.
B. The size of seasonal variation is constant, varying up and down by a little more than 5 ppm.
C. The average rate of change between 1975 and 1990 is between 0.5 and 1.5 ppm per year.
D. Carbon dioxide levels show a cyclical pattern, with two cycles per year.
E. The average rate of increase in carbon dioxide levels was greater between 1985 and 1990 than it was between 1975 and 1980.

Not only does this question ask about several features of the plot, it also, in a gentle way, involves more than one of the four modes of statistical thinking, another route to richness and complexity. In addition, the correct answer -- D is false -- is tied to the applied context in a way that thoughtful students may recognize: There is only one cycle per year, linked to the annual cycle for leaves of deciduous trees.

2. Involve two or more of the four modes of statistical thinking.
Here is a simple question that involves both visual and verbal/interpretive thinking:

Question 7. Land areas
Which histogram shows the distribution of land areas for the 50 US. states?
To answer the question correctly you must be able to match the shape of the histogram to what you know about the applied context. (Perhaps the easiest features to exploit are the two outliers that correspond to Alaska and Texas.)

In the next example, interweaving the visual, verbal, and implied numerical thinking is more intricate:

**Question 8. Pigs & vitamins**

You might expect giving vitamin B₁₂ to baby pigs would help them gain weight faster. However, intestinal bacteria can keep a pig from making use of the B₁₂, and a controlled experiment showed that it works better to give antibiotics along with the vitamin: Adding B₁₂ without antibiotics had little effect. Adding antibiotics without B₁₂ was worse than nothing. But both together were quite effective.

Which two of the four graphs below correctly show the interaction?
Note that while both Questions 7 and 8 involve more than one kind of thinking, they each focus on just one data feature — outliers in Question 7, interaction in Question 8. One standard strategy for analyzing distributions and relationships is to begin with a plot (visual), describe the main features of its shape (visual to verbal), and relate these to the applied context (verbal/interpretive), perhaps using numerical summaries (computational/algorithmic).

**Question 9. Emissions**

The four time plots below show annual emissions levels from 1980 through 1990 for four different pollutants. Match each plot with its verbal description.

- **SO₅ (sulfur oxides).** After a sharp 10% decline in the early 80s, later years showed more variation than trend.
- **Lead.** After a decline of more than 75%, emissions held steady, with little variation, for five years.
- **CO (carbon monoxide).** There is a linear pattern of decrease, apart from small year-to-year fluctuations.
- **Noₓ (nitrogen oxides).** Variation is larger, in relation to trend, than for the other graphs.

---

*Objective format questions in statistics 2/22/98 DRAFT 10:40 AM*
Question 10. Health Care
Match each verbal description with its plot:

A. People per nurse versus percent of population 65 or older. Two states have a far lower percentage of senior citizens than others. For the rest, states with a higher percentage of seniors have fewer people per nurse.

B. Percent of income spent on health versus number of doctors. The distribution for number of doctors is strongly skewed toward high values. (Converting to doctors per 1000 people, or people per MD, is indicated.)

C. Number of hospitals versus number of people per doctor. The distribution for number of hospitals is skewed toward large values. (Converting to people per hospital is indicated.

So far, all the examples have been "static" in the sense of asking only about what is there on the page. Questions can also ask students to compare what is there with possible changes, to ask "What if ... ?"

5. Ask for dynamic comparisons
Here are two examples of "What if ... ?" questions, one about experimental design, the other about influential points in regression.

Buttercups can grow either in damp woods or sunny fields. Suppose you want to see whether buttercups adapt in ways that make them grow larger in one environment than the other. You find one site of each kind, pick 10 plants from each, and randomly choose 5 of each 10 to transplant to the other site. Eight weeks later, you dig up all 20 plants and measure their dry weight. Evaluate each design change suggested below by matching it with one of the labels (1)–(4):


A. Use two sites of each kind
B. Repeat the study each year for five years.
C. Use each plant as a block -- assign it first to one environment, then the other, in a random order.
D. Dig up all 10 plants at each site; randomly transplant half to the other kind of environment; return the other half to their original environment.

**Question 12.**

All four scatter plots show 1990 vehicle-related deaths per 100,000 residents on the y-axis for 16 southern states (open circles) and the District of Columbia (x). Each plot shows a different predictor on the x-axis. Match each predictor and description with its plot.

A. Minimum age for a license. DC is an influential point. Eliminating DC changes the correlation from -.66 to -.17.
B. Number of cars per person. Eliminating DC changes the correlation from .16 to -.17.
C. Population density. Eliminating DC changes the fitted slope from -.0015 to -.018. (The shape of the plot suggests transforming the predictor variable to logarithms as a possible alternative to regarding DC as a separate case.)
D. Median family income. DC is an outlier, but only moderately influential. Eliminating DC improves the fit, and changes the slope from -.59 to -.49.
Before leaving this problem, note that it becomes a bit more challenging, but by no means impossible, if you redo the graphs using only open circles, so that DC is no longer automatically identified by its own plotting symbol. This change makes it harder to match the graphs to the descriptions purely on the basis of shape, and so encourages greater attention to the applied context. (Thus, for example, the graph in the lower right must be the one for the minimum age for a license because no other explanatory variable would be confined to just three values.) For a truly difficult variant, eliminate the pictures:

**Question 13.**

Imagine four scatter plots, each showing 1990 vehicle-related deaths per 100,000 residents on the y-axis and a different explanatory variable on the x-axis, for 16 southern states and the District of Columbia. Use the descriptions of the explanatory variables to visualize the likely relationship between DC and the 16 states; then use your imagined scatter plot to match each predictor with the appropriate description of the effect of removing DC from the plot.

A. Number of cars per person.
B. Population density.
C. Minimum age for a license.
D. Median family income.

1. Eliminating DC changes the fitted slope from -.0015 to -.018. (The shape of the plot suggests transforming the predictor variable to logarithms as a possible alternative to regarding DC as a separate case.)
2. DC is an outlier, but only moderately influential. Eliminating DC improves the fit, and changes the slope from -.59 to -.49.
3. DC is an influential point. Eliminating it changes the correlation from -.66 to -.17.
4. Eliminating DC changes the correlation from .16 to -.17.
4. Conclusion.

In my discussion of the examples, I have looked at each of the five principles one at a time, but that has been for the sake of exposition. In practice, I think of them as being used simultaneously. Thus, for example, Question 1 (McClesky versus Georgia) has 5 choices, asks for category matching rather than comparative judgments, involves one set of interrelated features of the data, relies almost exclusively on verbal thinking, and asks only about what is actually there. Judged by these criteria, it is a fairly simple and straightforward question. Question 10 (Health care) is richer and harder. It offers nine choices, involves some comparative thinking, asks about multiple data features involving three modes of thinking (verbal, visual, and numerical), but is static in the sense of asking only about what is present on the page. Question 12 is richer still. It offers 16 choices rather than 9, relies more on comparisons than category matching, asks about multiple data features using three modes of thinking, and is dynamic in that each verbal description asks about the effect of a change.

To my knowledge it remains an open question how and to what extent these principles, regarded as criteria for classifying questions, may relate to the difficulty of the questions and to other standard measures from the psychometric study of test items. It would be interesting to gather and analyze data on such relationships, and also to see whether and to what extent the five principles I have articulated here correlate with data-analytic skills and understanding of statistical concepts as measured by more “authentic” means of assessment.

Finally, I shall assert a naive outsider’s privilege and offer an exhortation: Statisticians who teach and researchers in the area of statistics education need each other to a greater extent than has been generally acknowledged. To make the teaching, learning, and assessment of learning in statistics as effective as possible, we need to think carefully about pedagogy and curriculum, about method and content, together. Regarded purely in terms of method, objective-format questions in statistics may appear to have little to offer, especially in this age of authentic assessment. I have tried to show, mainly through examples, that by thinking carefully about content, we may be able to create questions that not only take advantage of the format’s obvious affinity for computer implementation, self-assessment, and quick scoring, but also do some semblance of justice to the subject matter.