An Essay on the Statistical Analysis of Exam Questions

Prof. Patrick J. Davis, Associate Dean*

*Note: The original draft of this document was prepared by P4 student **Erin Reile** as part of her Fall-04 Academic Internship requirements.

Synopsis: The purpose of this handout is to help students gain a basic understanding of the statistical analysis that is performed by the Measurement and Evaluation Center (MEC) on exam questions. By explaining this process, students may better understand how final decisions are made regarding the validity of test questions and acceptable answers. At first glance this handout may seem to be very complicated but it only explains the four most commonly reviewed statistical values and their importance concerning exam questions. These four values, Split, Sum, **P value** and **R value** (the latter two being most important for purposes of this discussion) are explained in further detail below.

Detail: Here is an example of a multiple choice question (5 foils) with the resultant MEC analysis that suggests that this is a statistically 'good' question:

Question: Which of the following mechanisms serves as the genetic basis for the recently emergent strains of vancomycin-resistant Staph aureus?

A. These new strains produce penicillinases.
B. These new strains have altered PBP-2.
C. These new strains have acquired the VanA cassette from enterococci.
D. We don't yet know the basis of resistance in these strains (they're too new).
E. These new strains overproduce carboxypeptidases as a 'sponge' mechanism.

Keyed Response = C

Split	A	B	C	D	E	SUM
1	0	0	35	0	0	35
2	0	1	31	0	2	34
3	1	1	29	0	3	34
4	6	2	21	1	5	35
Sum	7	4	116	1	10	138

P value = 0.84 **R value** = 0.47

There are four items that are most commonly reviewed:

1. Split: The students are separated into quartiles (1-4).
   - For example, split 1 represents the top fourth of the class and split 4 represents the bottom fourth of the class in terms of overall performance on the exam (i.e., how did those that did well on the exam, i.e., split 1, answer this question?).
   - This allows faculty members to see how well students answered the question based on their ranking in the class on this exam.

- A "good" question (as shown above) would indicate that the top fourth of the class answered the question correctly and the number of correct answers decreased in each subsequent quartile.

2. Sum: This is the sum for each of the answers (i.e., individual sums for A,B,C,D,E).
   - This value becomes important when a majority of the class is split between two answers.
   - For example, if there were two (potentially) right answers or the question was misinterpreted by a majority of the class, we would see disproportionately high numbers in two of 'sum' columns.
   - This *may* lead to the faculty member giving credit for two answers, but would not justify throwing out the question; i.e., giving all students credit since there are still *clearly* three choices that are incorrect.

3. **P value**: This is the percentage of the class that answered the question correctly.
   - This value should ideally be higher than 0.62 for a multiple choice question (meaning that 62% of the class answered the question correctly) and 0.75 for a True/False question.
   - The higher the P value, the easier the question. A P value of 1.0 is not as desirable, because this means that 100% of the class answered the question correctly and the question does not discriminate student performance.

4. **R value**: This is the discrimination value, or the statistical difference between the students who scored high on the exam versus the students who scored low.
   - This value should be greater than 0.2; a value of 0.4-0.5 is very good!
   - This shows that there was a significant difference in the number of correct answers between the top fourth of the class versus the bottom fourth of the class.
   - Values less than 0.2 do not give much information about the differences among the quartiles. For example, if the *entire class* answered a question correctly there would be no difference between the quartiles, and the R value would be less than 0.2 (zero, in fact). The same would *probably* be true if a very large proportion of the students missed the question (a significant number of students would need to get the question right to be able to evaluate *who* got it right vs wrong).

Here is an example of results that would suggest a *potential* problem with the question:

Keyed Response = D

Split	A	B	C	D	E	SUM
1	8	0	0	27	0	35
2	18	0	1	15	0	34
3	19	2	2	11	0	34
4	25	0	2	8	0	35
Sum	70	2	5	61	0	138

P value = 0.44 **R value** = 0.39

- When evaluating this question you notice that the P value is 0.44 meaning that only 44% of the class answered this question correctly. However, that (in and of itself) does not make this a 'bad' question; the question that remains is "*Which* students got it right and *which* students got it wrong?") You also notice that 70 students answered A and only 61 students answered D (the correct answer). This tells you that something *may* be wrong with the question if half of the class thought the correct answer was A. *However*, you should also note that the R value is greater than 0.2; indeed, the R value of 0.39 suggests that the question did

discriminate well. That is, more students in top fourth of the class knew the correct answer and that the number of correct answers decreases in the subsequent quartiles. Thus, a reasonable interpretation would be that this was a challenging questions in that the majority of students missed it, but the question did discriminate well in terms of which students knew the material.

Final Analysis: It is important to understand that the faculty member authoring the question makes the final decision about the quality of a question; the statistical analysis simply provides information to help in making the decision. The faculty member may choose to ignore these results if they feel that the question was fair. For example:

- The faculty member switched the answers from last years exam and the majority of the class keyed the old answer because they simply studied from (memorized?) the old test. This would massively skew the statistical results, but the correct answer is indisputable.
- A chapter in the text was assigned but none of the students bothered to read it so the entire class missed the question. If it was clear that this assignment was "required" (rather than "optional"), there is no compelling reason to throw out this question, regardless of the statistics.
- The faculty member is confident that the material was addressed in class and its importance adequately emphasized and it doesn't matter how many students missed the question. The availability of recorded lectures facilitates this assessment.

If you are interested in a more extensive discussion concerning the MEC analysis of test questions, see http://www.utexas.edu/academic/mec/scan/scanitem.html