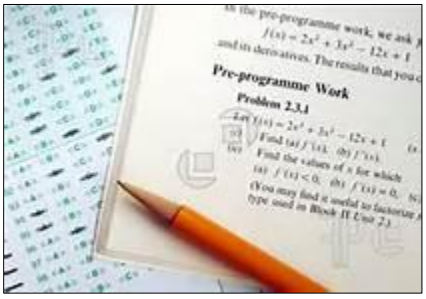
Interpreting Exam Performance: What Do Those Stats Mean Anyway?

"There are three kinds of lies: lies, damned lies, and statistics."
--Mark Twain



If you are utilizing selected response (i.e. multiple choice) exam items in your courses, Creighton University Grader Services can supply you with answer "bubble sheets." If you use these, Grader Services will scan the bubble sheets for you and provide you with detailed exam statistics and item analysis. The bubble sheets can be pre-filled with your students' names and net IDs from BlueLine or filled in manually. You will need to contact Grader Services well in advance to order the bubble sheets and the scanning service. Grader Services full description.

Some departments may use other programs to administer quizzes and exams to students, such as ExamSoft, Question Mark, LXR, or even BlueLine. Work with your program chair or mentor to determine what type of program you are using and how your exams will be administered.

What do those stats mean?

Regardless of what program you are using for quizzes or exams, you will receive a report containing exam statistics and individual item analysis. It is important to understand what the numbers on the report mean, as you will use these to make decisions about not only student learning, but also the quality of the exam.

Measures of Central Tendency: You will receive the **mean** exam score, the **median** score, the **range** of scores (lowest and highest scores), and the **standard deviation** (a measure of variability; the average distance from the mean score). It is up to you to determine acceptable benchmarks for these scores based on the level of the student and the difficulty of the material. You will also receive a report listing each student's score. You can use this to determine the number of As, Bs, Cs etc to determine how the exam scores are distributed.

Reliability Coefficient:

This is a measure of the likelihood of obtaining similar results if you re-administer the exam to another group of similar students. The most useful measure is generally the **Kuder-Richardson Formula 20 (KR-20):**
KR-20:

- Ranges from 0-1 (the higher the better)
- Greater than 0.5 can be considered good on a teacher-made test
- Best when a test measures a unified body of content
- Lots of very difficult items or poorly written items can skew this
- The higher the variability in scores, the higher the reliability (McGahee & Ball, 2009)

Tag Label: PtT-Exam Analysis

Individual Exam Item Analysis

For each item, you will receive a report on how many students selected each response, the **item difficulty**, and the **item discrimination**. Careful examination of each of these is critical, as you will use this information to determine the quality of the item. Sometimes, information from item analysis may be used to decide if you want to accept more than one item as correct, or discard an item all together (what Grader Services calls and "edit"). It should always be used to determine if the item should be re-written to improve it for future use. These are difficult decisions to make, especially on high-stakes exams. Your mentor or program chair can offer you guidance in this area.

Item Difficulty: p value: (Kehoe, as cited in McDonald, 2007)

- Percentage of correct responses
- Desired range on exam 0.3-0.8 (30% to 80% correct)
- Mean p value translates to mean % score on exam
- Very difficult items (less than 30% correct) or very easy items (more than 80% correct) contribute very little to the overall exam reliability
- Sometimes, you may be OK with 100% on an item, particularly if it is something that is critical for students to know.

Item Discrimination: Point Biserial Index (PBI) (McGahee & Ball, 2009)

- Correlation between score on an item and score on the exam
- Differentiates between those who have high or low test scores
- Range from -1 to +1
- Positive PBI indicates those who scored well on exam answered item correctly.
- PBI should be *positive* for correct answer
- PBI should be *negative* for distractors
- All options should be chosen
- Look at ***all*** PBIs; not just correct answer!

PBI: General rules **(**Penn, 2009; McGahee & Ball, 2009)

- Below 0.2: Poor; revise item
- 0.2-0.29: Fair
- 0.3-0.39: Good
- 0.4-0.7 Very good

Below are some examples of how you might use item difficulty and item discrimination to make decisions about individual exam item performance:

Tag Label: PtT-Exam Analysis

Sample Exam Statistics

On the examples below, a "*" indicates the correct answer (the "key"). Note: incorrect responses are referred to as "distracters."

Example #1

Response	Frequency	Percent	Point Biserial
A	4	4.88	-0.22
B	2	2.44	-0.21
C*	72	87.8	0.52
D	4	4.88	-0.42

This item performed well (discriminated between those who knew the content and those who did not). This is indicated by the high PBI (0.52) and the negative PBIs for each incorrect response. The item was challenging enough but not "too easy." (87.8% answered correctly)

Example #2

Response	Frequency	Percent	Point Biserial
A*	76	92.68	0.04
B	4	4.88	-0.08
C	0	0	-
D	2	2.44	0.04

This item was relatively easy (92.68% answered correctly). No one selected distracter C, so this should be revised (it didn't "distract" anybody). The PBI is positive, but very low. This item did not discriminate well. The positive PBI for distracter D indicates that the 2 students who selected D were high performers on the exam.

Example #3

Response	Frequency	Percent	Point Biserial
A	3	3.66	-0.21
B	0	0	-
C	2	2.44	-0.26
D*	77	93.9	0.34

The PBI for response D is acceptable at 0.34. Distracters have negative PBIs, so that is good; however this question was very easy (93.9% correct). No one selected distracter B, so it should be revised to be a more attractive response.

Example #4

Response	Frequency	Percent	Point Biserial
A	21	25.61	-0.02
B	13	15.85	0.11
C*	37	45.12	-0.05
D	11	13.41	-0.01

This item did not discriminate well. The negative PBI for the correct answer C indicates that students who performed poorly on this exam answered correctly; the positive PBI (0.11) on distracter B indicates that students who performed well on the exam selected it. This is problematic. Based on your expertise, you may consider accepting both B and C as correct or discarding this item. In any event, this question should be carefully examined and revised.

LINKS:

Tulane School of Medicine Interactive Item Analysis Module

SUBMITTER INFORMATION:

Anne Schoening, PhD, RN, CNE
School of Nursing

REFERENCES:

McDonald, M.E. (2007). *The nurse educator's guide to assessing learning outcomes*. Sudbury, MA: Jones and Bartlett.

McDonald, M.E. (2008). Developing trustworthy classroom tests. In Penn, B.K. (Ed.), *Mastering the teaching role: A guide for nurse educators* (pp. 275-286). Philadelphia: FA Davis

McGahee, T.W. & Ball, J. (2009). How to read and really use an item analysis. *Nurse Educator, 34*, 166-171.
Morrison, S., Nibert, A., & Flick, J. (2006). *Critical thinking and test item writing*. Houston: Health Education Systems, Inc.

Penn, B.K. (2009, August). Test item development and analysis. Presented at Creighton University School of Nursing Faculty Retreat, Omaha, NE.

Tag Label: PtT-Exam Analysis