**Sample Item Review Process**

**Item Analysis Process**

**\*** Evidence-Based Medicine sources used to verify accuracy, including UpToDate and Access Medicine textbooks.

**Item Correction Process**

**(at the instructor’s discretion)**

There are a variety of ways to obtain test item statistics:

1. Statistics are obtained through the Learning Management System (Pearson/Brady and Jones & Bartlett each have their own platforms).
2. Statistics are obtained through Scantron Report
3. Statistics are compiled by each instructor as exams are graded.

# **References:**

**Difficulty Index**

This statistic (also known as the *p* value) stands for the proportion of participants who got the item correct. The *p* value can take on any value between 0.00 and 1.00. Higher values denote easier items (more people answered the item correctly), and lower values denote harder items (fewer people answered the item correctly). The higher the value, the easier the question. (Dallas and Goodman, 2019).

A reasonable goal is to review items with *p* value less than 0.50.

**Standard Deviation**   
The standard deviation indicates how much scores vary from the average, ranging from 0% to 100%. A high standard deviation indicates that scores are spread out from the average, whereas a low standard deviation indicates that scores are close to the average. The item standard deviation is most meaningful when comparing items which have more than one correct alternative and when scale scoring is used. For this reason, it is not typically used to evaluate classroom tests. Standard deviation is generally utilized to look at the aggregate scores, not the individual items.

**Discrimination Index**   
The discrimination index indicates how well a question differentiates between high and low performers. It can range from -100% to 100%, with high values indicating a “good” question, and low values indicating a “bad” question. In practice, values of the discrimination index will seldom exceed 0.50 because of the differing shapes of item and total score distributions.

A general goal for item discrimination is “good” if the index is above 0.30; “fair” if it is between 0.10 and 0.30; and “poor” if it is below 0.10.

**Point Biserial Correlation Coefficient**   
The point biserial correlation coefficient is an analysis only applied to multiple choice and true/false question types that have only one answer with weight 100%, and all others with weight 0%.   
Similarly to the discrimination index, the point biserial correlation coefficient relates individuals’ quiz scores to whether or not they got a question correct. Another way to phrase this is that it correlates score on an item and score on the exam. It ranges from -1.00 to 1.00, with high values indicating a “good” question, and low values indicating a “bad” question.

Evaluate all point biserial index values, not just the correct answer. A reasonable threshold is that answers with a point biserial value below 0.2 are considered poor and should be revised. Items with a point biserial value between 0.2-0.29 are considered fair. Items with a point biserial value between 0.3-0.39 are considered good. Items with a point biserial value between 0.4-0.7 are considered very good (Schoening, 2019).  
  
**References:**

Dallas, D. and Goodman, J. (2019). *Just what you've always wanted: making large-scale testing practices accessible at the local level*.

Schoening, A. (2019). [online] Creighton.edu. Available at: [https://www.creighton.edu/sites/www12.creighton.edu/files/PtT-Exam%20Analysis.pdf](https://www.creighton.edu/sites/www12.creighton.edu/files/PtT-Exam%20Analysis.pdf%20) [Accessed 4 May 2019].