Evaluating the Psychometric Data

Users can evaluate the performance of assessments, questions, and distractors by utilizing the available ExamSoft reports to gather specific statistics. This allows faculty to identify how their questions are performing and make any necessary changes.

Psychometrics are the mathematical statistics of each question and assessment and can be interpreted very differently based on the purpose of the question. No single statistic can give you the entire picture of how an item should be interpreted. Likewise, there are no ideal psychometrics that are accurate for every item. The best practice is always to evaluate all pieces of information available about an item while accounting for the intention of the question. A question that was created by the faculty member as an "easy" question may serve that specific purpose, but it has very different psychometric results than that of a question that was created to be discriminatory. Additionally, take into account any outside factors that could be influencing the statistics (including content delivery method, conflicting information given to the students, testing environment, etc.) Lastly, always keep in mind how many students took the exam. If the number is very low then the statistics are significantly less reliable than if it is a large group of students.

Exam Statistics

Exam Statistics are statistics that are determined based on the performance of all students on all questions of the exam. This data can be found on a number of reports (including the Item Analysis and Summary Reports).

Mean: The mean is the average score of all exam takers who took the exam. It is found by dividing the sum of the scores by the total number of exam takers who took the exam.

$$\frac{S(x)}{N}$$

x = score of the exam
N = total number of exam takers

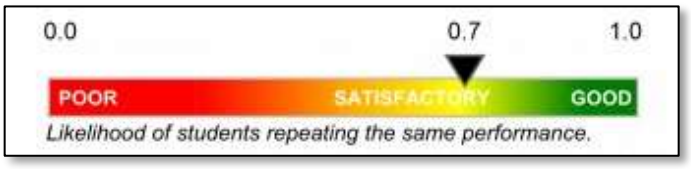
Median: The median is the score that marks the midpoint of all exam takers' scores. It Is the score that is halfway between the highest and lowest scores.

Standard Deviation: The standard deviation indicates the variation of exam scores. A low standard deviation indicates that exam taker's score were all close to the average, while a high standard deviation indicates that there was a large variation in scores.

$$\sqrt{\frac{\Sigma(x-\bar{x})^2}{N}}$$

x = exam score
$\bar{x}$ = average score
N = total number of exam takers

Reliability KR-20 (Kuder-Richardson Formula) *(0.00-1.00):* The KR-20 measures internal consistency reliability. It takes into account all dichotomous questions and how many exam takers answered each question correctly. A high KR-20 indicates that if the same exam takers took the same assessment there is a higher chance that the results would be the same. A low KR-20 means that the results would be more likely to be different.



$$\frac{k}{k-1}\left(1-\frac{\Sigma_j^k p_j q_j}{\sigma^2}\right)$$

k = number of questions
p_j = proportion of exam takers who answered question *j* correctly
q_j = proportion of exam takers who didn't answer question *j* correctly
σ^2 = variance of the total scores of all the people taking the test

Question Statistics

Question Statistics are statistics that assess a single question. These can be found on the item analysis as well as in the question history for each question. These can be calculated based on question performance from a single assessment or across the life of the question.

Difficulty Index *(0.00-1.00):* The difficulty index measures the proportion of Exam Takers who answered an item correctly. A higher value indicates a greater proportion of Exam Takers responded to an item correctly. A lower value indicates that less exam takers got the question correct.
In addition to being provided the overall difficulty index, there is an Upper Difficulty Index and Lower Difficulty Index. These follow the same format as above but only take into account the top 27% of the class and the lower 27% of the class respectively. Thus the Upper Difficulty Index/Lower Difficulty index reflects what percentage of the top 27%/lower 27% of scorers on an exam answered the question correctly.

Question #	Correct Responses			Disc. Index	Point Biserial
	Diff(p)	Upper	Lower		
1	0.90	100.00%	77.78%	0.22	0.40

Discrimination Index *(-1.00-1.00)*: The discrimination index of a question shows the difference in performance between the upper 27% and the lower 27%. It is determined by subtracting the difficulty index of the lower 27% from the difficulty index of the upper 27%. A score close to 0 indicates that the upper exam takers and the lower exam takers performed similarly on this question. As a discrimination index becomes negative, this indicates that more of the lower performers got this question correct than the upper performers. As it becomes more positive, more of the upper performers got this question correct.

$$\frac{\#\ Correct\ by\ Upper\ Group}{\#\ Exam\ Taker's\ in\ Upper\ Group} - \frac{\#\ Correct\ by\ Lower\ Group}{\#\ Exam\ Taker's\ in\ Lower\ Group}$$

Evaluating the Psychometric Data

Determining an acceptable item discrimination score depends on the intention of the item. For example, if it is intended to be a mastery-level item, then a score as low as 0 to .2 is acceptable. If it is intended to be a highly discriminating item, target a score of .25 to .5.

Point Bi-Serial *(-1.00-1.00):* The point bi-serial measures the correlation between an ET's response on a given item and how the ET performed on the overall exam.
A point bi-serial that is close to 1 indicates a positive correlation between the performance on the item and the performance on the exam. Students who did well on the exam also did well on this question and students who did poorly on the item did poorly on the exam. A negative point bi-serial indicates a negative correlation between the two. Students that did well on the item did not do well on the exam and students who did not do well on the item did do well on the exam. This may be something to review. A point bi-serial close to 0 indicates that there was little correlation between the performance of this item and performance on the test as a whole. This may indicate that the question tested on material outside of the other learning outcomes assessed on the exam or that it was a mastery item where all or most of the class got the question correct.

$$\frac{M_p - M_q}{S}\sqrt{pq}$$

M_p = whole–test mean for ETs answering question correctly

M_q = whole–test mean for ETs answering question incorrectly

S = standard deviation

p = proportion of students answering correctly

q = proportion of students answering incorrectly

Response Frequencies: This details the percentage of students that selected each answer choice. If there is an incorrect distractor that is receiving a very large portion of the answers, you may need to assess if that was the intention for this question or if something in that answer choice is causing confusion. Additionally, an answer choice with very low proportions of responses may need to be reviewed as well.

When reviewing response frequencies you may wish to also examine the distribution of responses from your top 27% and lower 27%. If a large portion of your top 27% picked the same incorrect answer choice it could indicate the need for further review.

Point Biserial	Correct Answer	Resp A	B	C	D
0.40	A,B	*0	*87	1	9
-	% Selected	0.00	89.69	1.03	9.28
	Point Biserial (rpb)	0.00	0.40	-0.03	-0.41
-	Disc. Index	0.00	0.22	0.00	-0.22
-	Upper 27%	0.00	1.00	0.00	0.00
-	Lower 27%	0.00	0.78	0.00	0.22

The Upper and Lower Groups of ETs are based on the top 27% and bottom 27% of performers respectively. 27% is an industry standard in item analyses.