University Center for
Assessment, Teaching
& Technology

Improving Multiple Choice Questions

This guide explains how to question statistics can be used to assess test questions and student learning so as to inform teaching decisions. D2L BrightSpace Quizzes are used as an example. This guide offers a resource for how to write good multiple-choice-type test questions. Depending on the circumstances, it may be best to re-write a question, remove a question, or re-think how a topic is addressed in the course.

Question statistics = Item Analysis

“Question statistics” are statistical data automatically gathered by an instructional tool, e.g. D2L BrightSpace’s Quizzes tool or Gradescope Bubblesheet Statistics. The process of examining student responses to multiple-choice-type questions is also known as “item analysis.”

What kinds of knowledge can be tested?

Factual Knowledge Terminology, Facts, Figures
Conceptual Knowledge Classification, Principles, Theories, Structures, Frameworks
Procedural Knowledge Processes, Techniques, and Methods and When and How to use them.
Metacognitive Knowledge Strategy, Overview, Self Knowledge, Knowing how one knows.

Writing Good Test Questions

For a basic guide to writing and revising questions, please see Cynthia Brame’s Writing Good Multiple Choice Test Questions. Test validity increases if each question aims at specific knowledge. Bloom’s Taxonomy can assist with precision and accuracy in writing each question’s objective and then crafting the question itself.

It is not an easy process to write a good question--so why do it? Tests and quizzes can help students study, practice, and retain new information. Online objective questions, e.g. multiple-choice-type questions, have additional advantages of efficiency. Automated grading saves instructional labor and returns results quickly to each student.

Why use question statistics or item analysis?

Question statistics can help you determine:

- Is a question difficult, easy, or confusing?
- Does a question do a good job separating students who know from those who are just guessing?
- Should a question be eliminated or revised?
- Which questions are good?

Updated 9/30/22

- With which learning objectives are students having difficulty?

The statistical data are best used together rather than separately. Most importantly, other contextual factors and the format of the written questions themselves should always be considered when making judgments about the quality of a question.

D2L Brightspace makes readily available to instructors the following information: Average Grade, Standard Deviation, Point Biserial, and Discrimination Index. *Note that in D2L only a student's first attempt on a question is included in question statistics.*

Gradescope also includes statistics; see Bubblesheet Statistics and Assignment and Question Statistics

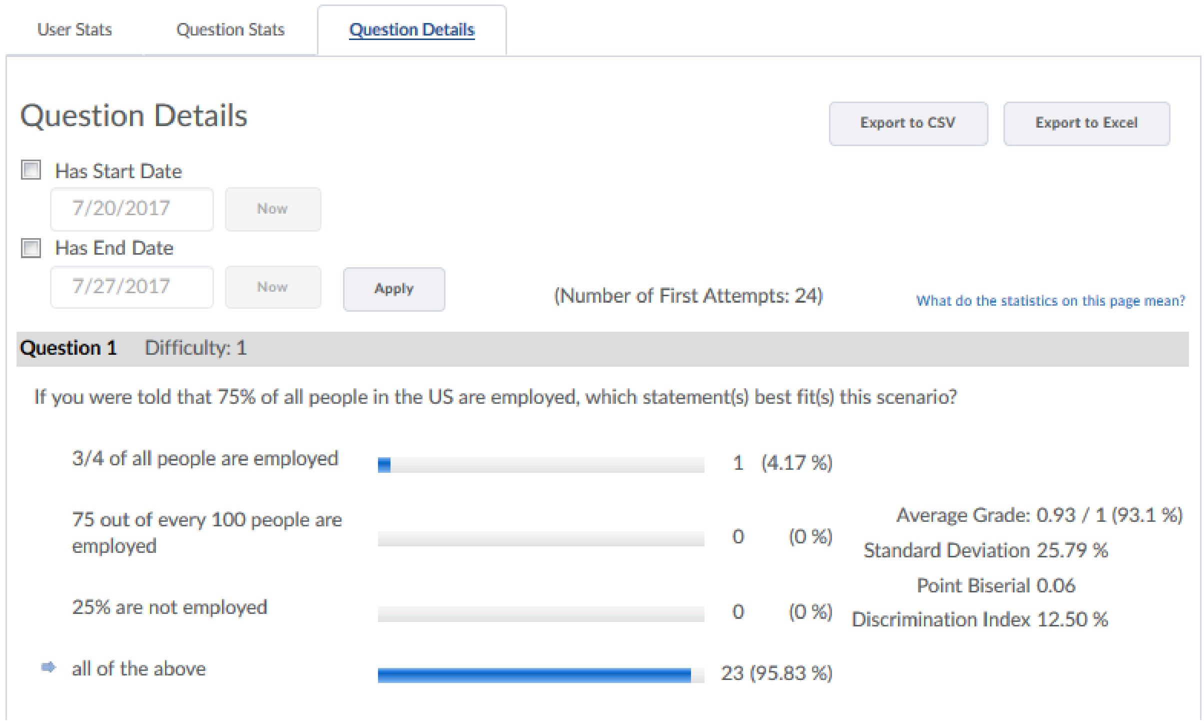
What is item analysis?

The goal of item analysis is to improve the reliability and validity of multiple-choice type questions. Item analysis is a process that views questions as tools used to discern those who know from those who don't. Metaphorically, item analysis is a way to sharpen these tools so that they work well; a sharp item is reliable and valid. **Reliability** is like consistency. If a question is highly reliable it yields consistent results under similar conditions. For example, a difficult question answered correctly only or mostly by students who scored high overall on the test is considered reliable. **Validity** refers to how well an item accurately measures what it is intended to measure. For example, each question should assess a specific learning objective. An ambiguous or confusing question will not yield valid measurements of student knowledge or skill.

Let's Look at an Example in D2L

Screenshot of D2L Quiz Question Details. <alt text: Question: If you were told that 75% of all people in the United States are employed, which statement(s) best fit this scenario? 23 of 24 students (95.83%) of respondents selected the answer option "all of the above." 1 of 24 students (4.17%) chose option "¾ of all people employed." The average grade is 0.93/1 (93.1%), with a Standard Deviation of 25.79%, a Point

Biserial 0.06, and a discrimination index of 12.50%. >



Question Details

☐ Has Start Date 7/20/2017

☐ Has End Date 7/27/2017

(Number of First Attempts: 24)

Question 1 Difficulty: 1

If you were told that 75% of all people in the US are employed, which statement(s) best fit(s) this scenario?

3/4 of all people are employed	1	(4.17 %)
75 out of every 100 people are employed	0	(0 %)
25% are not employed	0	(0 %)
➡ all of the above	23	(95.83 %)

Average Grade: 0.93 / 1 (93.1 %)
Standard Deviation 25.79 %
Point Biserial 0.06
Discrimination Index 12.50 %

Average Grade

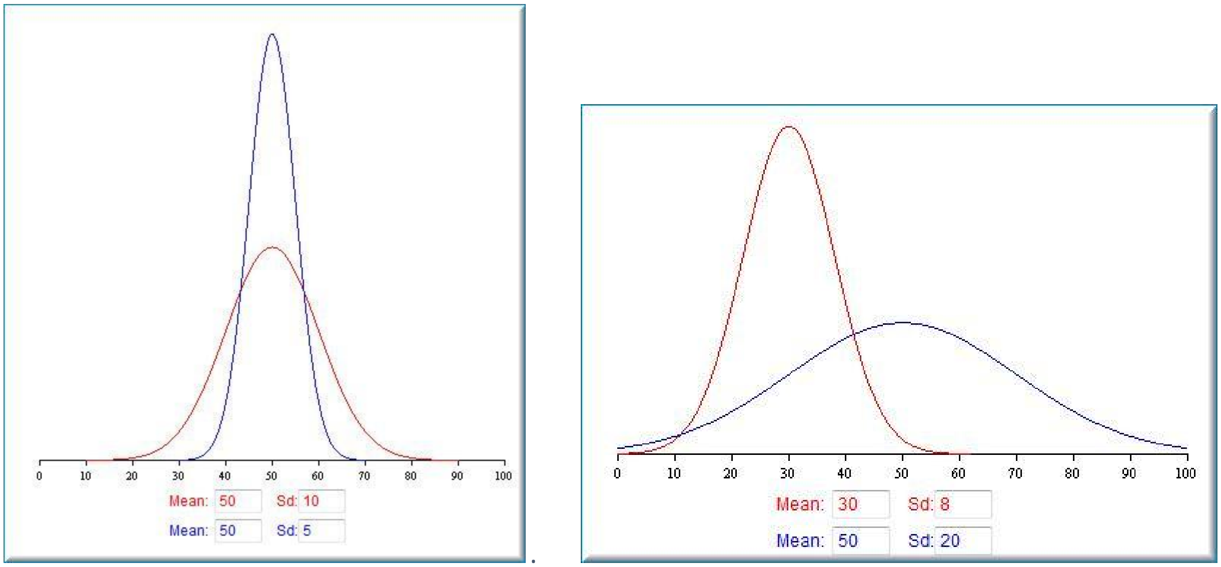
“Average grade” describes the percentage of test takers who answered this question correctly on the first attempt. A higher percentage may indicate an easier question. A lower percentage may indicate the question is difficult. However, you cannot draw this conclusion solely from this statistic. You should consider the other statistics for the question and examine the question itself. A low percentage may indicate a poorly written question.

The average Grade range in D2L is from 0% to 100%. It is important to note that in D2L only a student’s first attempt on a question is included in the question statistics.

On some of D2L’s display interfaces “Average Grade” is also used as a label for the percentage of students who selected each answer option.

Standard Deviation

The standard deviation tells you about the variation of scores. A high standard deviation indicates that scores are spread out from the mean, in other words, student performance varied widely. A low standard deviation indicates that scores cluster close to the mean, i.e. student performance did not vary a lot. Think of a bell curve; typically results in a cluster close to the mean.



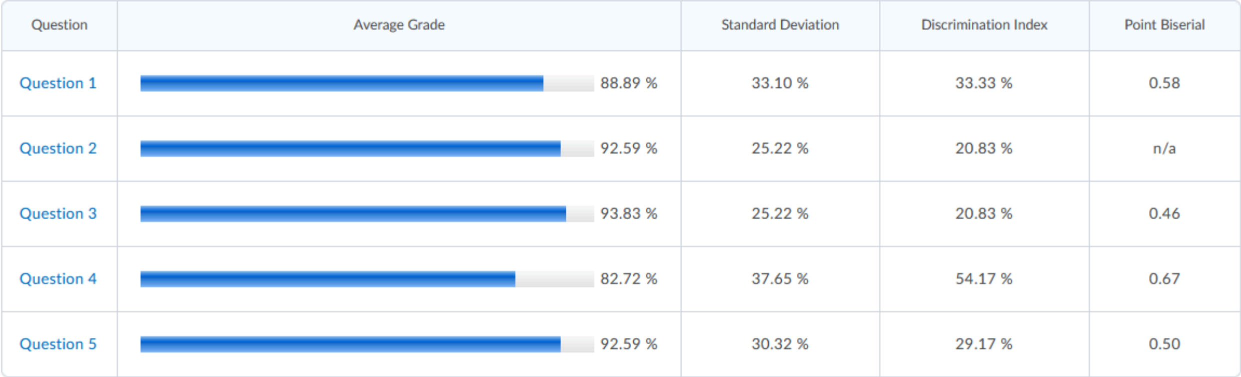
Graphical representation of differences in mean and standard deviation. The X-axis is scored; Y-axis is the number of learners. Greater standard deviation results in a graph with a lower peak and broader base. <alt text: Graph on the left shows: Bell curve graph of results of two tests with Mean of 50 are shown. The red graph line has SD=10; it has a lower peak and the base extends from 10 to 90. The blue graph line has SD=5; its peak is approximately twice as high and the base extends from 30 to 70.>

If there is a lot of variation in student performance, the standard deviation is greater.

The standard deviation range in D2L is from 0% to 100%. In D2L Brightspace one standard deviation from the mean is 34.13%.

PRACTICE ACTIVITY 1

In this example below, considering both average grade and standard deviation, would you eliminate or revise any of these questions?



Question	Average Grade	Standard Deviation	Discrimination Index	Point Biserial
Question 1	88.89 %	33.10 %	33.33 %	0.58
Question 2	92.59 %	25.22 %	20.83 %	n/a
Question 3	93.83 %	25.22 %	20.83 %	0.46
Question 4	82.72 %	37.65 %	54.17 %	0.67
Question 5	92.59 %	30.32 %	29.17 %	0.50

Discussion of Practice Activity 1

Performance on all questions clusters around the mean. Q4 has the widest spread.
The average score in Q3 is high (93.83%) and there is very little variation as indicated by a standard deviation of less than 34%. Question 3 might be too easy, but you'd have to examine the written question to determine that. When examining a question item, pay particular attention to whether or not all answer options are plausible.

Discrimination Index

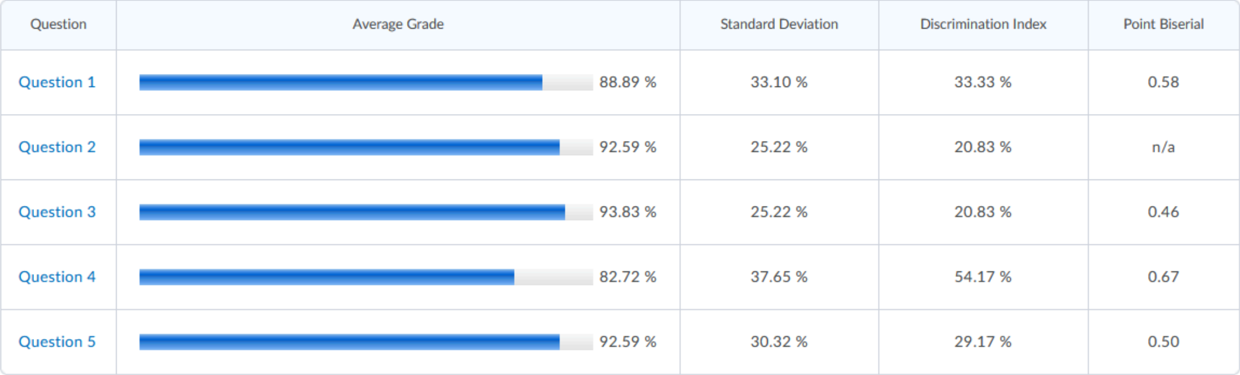
The discrimination index indicates how well a question differentiates between high and low performers on the question. A low discrimination index often indicates an ambiguously worded or confusing question, so this item should be examined. An item with a negative index should be examined, as it indicates that no student who performed in the upper range on the test overall answered this question correctly.

The discrimination index is calculated by subtracting the question's lower-performing 27% from its upper 27%. The higher the difference, the better the question discriminates between high and low.

Discrimination indices above 30% are considered “good;” between 10% and 30% are considered “fair.” Below 10% is “poor.

PRACTICE ACTIVITY 2

Which question appears to be best at discerning high and low-performing students?



Question	Average Grade	Standard Deviation	Discrimination Index	Point Biserial
Question 1	88.89 %	33.10 %	33.33 %	0.58
Question 2	92.59 %	25.22 %	20.83 %	n/a
Question 3	93.83 %	25.22 %	20.83 %	0.46
Question 4	82.72 %	37.65 %	54.17 %	0.67
Question 5	92.59 %	30.32 %	29.17 %	0.50

Discussion of practice activity 2

Question 4 because it has the highest discrimination index and an average grade which indicates it is not super easy.

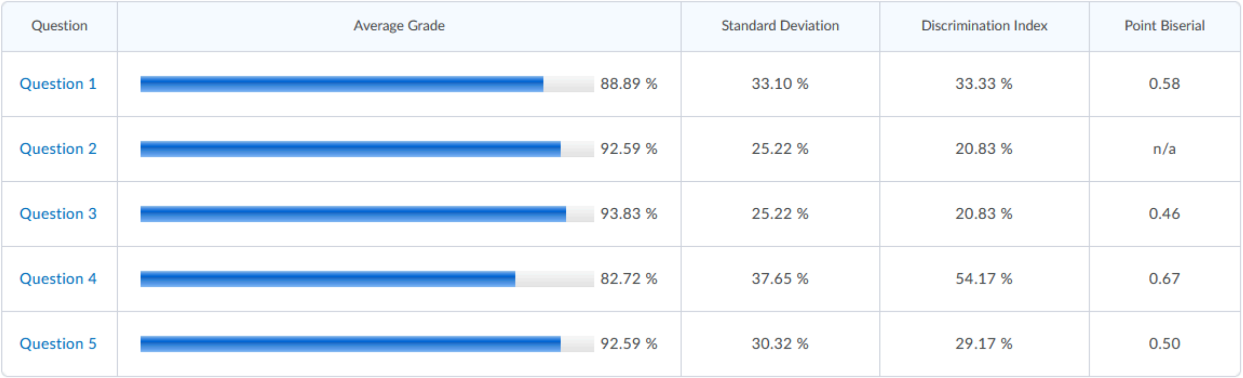
Point Biserial

Point Biserial is similar to the Discrimination Index as it measures the correlation between the question score and overall quiz score. In D2L Brightspace, the point biserial correlation coefficient is an analysis only applied to multiple choice and true/false question types that have only one answer with a weight of 100% and all other answer options weigh 0%. A negative score indicates that high-performing students got this question wrong. A value of zero means that all test-takers got the item correct.

Range in D2L is from -1.00 to +1.00. A value of “n/a” is displayed when all learners answer the question correctly or when there is no single 100% correct answer option.

PRACTICE ACTIVITY 3

Why does Question 2 not have a point biserial value?



Question	Average Grade	Standard Deviation	Discrimination Index	Point Biserial
Question 1	88.89 %	33.10 %	33.33 %	0.58
Question 2	92.59 %	25.22 %	20.83 %	n/a
Question 3	93.83 %	25.22 %	20.83 %	0.46
Question 4	82.72 %	37.65 %	54.17 %	0.67
Question 5	92.59 %	30.32 %	29.17 %	0.50

Other useful statistics

Difficulty index = the number of students who responded correctly divided by the total number of students who responded.

Below 0.20 (20%) indicates that the question is very difficult. Above 0.90 (90%) indicates that a question is very easy. The ideal difficulty index value is slightly higher than midway between chance and a perfect score for the item. Chance is equal to 1.0 (or 100) divided by the number of choices. A perfect score equals 100. Very difficult and very easy questions should probably be eliminated from a test. However, there may be good reasons to purposefully include an easy item. For example, to settle student test anxiety or affirm student knowledge.

Example: Given a question with five answer options, the chance is 20%. So a difficulty index of 65-70% is appropriate.

Note: The number associated with “Difficulty” in D2L is not a result of statistical analysis; it is assigned by the person authoring the question.

References and Resources

D2L Brightspace Help: Using the Assessment Quality Dashboard

Brame, C. (2013) Writing good multiple choice test questions. Retrieved July 19, 2021, from https://cft.vanderbilt.edu/guides-sub-pages/writing-good-multiple-choice-test-questions/

University of Washington Office of Educational Assessment https://www.washington.edu/assessment/scanning-scoring__trashed/scoring/reports/item-analysis/

Tobin, M. A. Guide to Item Analysis. Schreyer Institute. Pennsylvania State University. Retrieved July 20, 2021, from http://www.schreyerinstitute.psu.edu/pdf/GuideToItemAnalysis.pdf

How Do I Create a Test for MyStudents? https://www.depts.ttu.edu/tlpdc/Resources/Teaching_resources/TLPDC_teaching_resources/Documents/HowdoICreateaTestforMyStudentswhitepaper.pdf

Acknowledgements: Thanks to Erin Dokter and Georgia Davis for sharing their knowledge of test-design and D2L Brightspace.

Please contact this guide's current curator, Gretchen Gibbs, with any questions about multiple-choice questions.

Updated 9/30/22