# ITEM STATISTICS AND ANALYSIS

Item statistics are useful when reviewing items on an assessment. They provide information about each item to help reviewers improve the quality of the assessment.

The purpose of item statistics for item review is to provide information. These data are quantitative indicators for individual test questions. These metrics help reviewers objectively evaluate how well each item functions within the assessment. The initial review of an item is done by expert opinion. A more rigorous review is done after testing how well the items perform in practice.

## OBTAINING DATA

When reviewing items, it is best to pilot-test them with actual students. This gives us data that we can use to help evaluate each item.  A minimum of 100-200 respondents is often cited as a recommended sample size for basic (classical) item analysis. This allows for preliminary calculations of item difficulty and discrimination. However, the specific requirements can vary based on the purpose of the assessment, the stakes involved, and the particular statistical methods being used. For high-stakes or large-scale assessments, you will need a larger sample, one that is representative of the population you are testing ([sampling](https://edtechbooks.org/designing_surveys/sampling)). You can calculate the item statistics for a classroom test (low-stakes situations), but be cautious of the interpretation. Small sample sizes may results in a biased outcome because the characteristics of those in the class do not represent students in the larger population. If a test is reused, the combined item statistics from multiple exams will better estimate how well individual items perform.

## ITEM STATISTICS

Item statistics provide empirical evidence we can use to make informed judgments about items. This information guides decisions on whether to retain, revise, or remove specific items, ultimately improving the overall quality and effectiveness of the assessment. These statistics are not direct indicators of item quality, but together and in context, they help with the review process. Several metrics are used to review the effectiveness of test items. Item difficulty, the discrimination index, and a distractor analysis (for multiple-choice items) are commonly used in a basic (classic) item review process.

### ITEM DIFFICULTY

This statistic is represented by the letter P, it indicates the percentage of students who got the item correct, but in mathematical terms, item difficulty represents the proportion of students who got the item correct. Each item on the test will have an item difficulty. The numbers range from 0 to 1. Zero means no one (0%) got the item right, and a P = 1 (100%) indicates that everyone got this item correct. The lower the percentage, the more difficult the test item. This information is not an indicator of the quality of the item. It simply indicates how difficult the item was for those taking the test. This statistic is calculated using the following formula:

### Item Difficulty Formula

P=R/T

**Where**:

P is the Item Difficulty

R is the number of students who answered a question correctly.

T is the total number of students in the class who answered the question.

Consider the following example:

### Item Difficulty Example

Let R=49 individuals with correct answers and T=62 individuals responding

P=49/62

P=0.79

Interpreting this statistic is a simple descriptive process. Seventy-nine percent (79%) of those attempting this item got it correct. However, categorizing the difficulty of an item qualitatively is somewhat arbitrary. For example, you might use the following ranges to describe the difficulty of an item.

* Easy 80-100%
* Moderately Easy 50-79%
* Moderately Difficult 30-49%
* Difficult 0-29%

It is important to remember that the result should not be understood as an indicator of the quality of the item. The item difficulty is simply an indicator of how well individuals did on that item overall. For example, if everyone taking a test got a specific item correct, we can say it was an easy item, but we should not assume it was a good item. The item may only be testing a student's ability to recall a trivial fact, or it may have a hint to the correct answer, leading everyone to get the item correct even if they didn't know the answer. Likewise, an item with a P of 0 means everyone got the item wrong -- a difficult item.  The item may have identified something those taking the test did not know, were not taught, or developmentally were unable to accomplish, or perhaps there was a flaw in the item, leading people to choose the answer wrong. A teacher should use this data to inform potential changes to the instruction or to revise the item based on their understanding of the situation.

A decision about whether the level of difficulty for an item is acceptable must take into account the purpose of the test and why the item was included in the test. The difficulty level (P) is simply a piece of information that can be used, along with other data, to evaluate the item. For instance, if the item was part of a mastery test and the item measured an essential piece of knowledge, a difficulty level between 0.80 and 1.00 would be acceptable and might be expected. However, if the item was part of a norm-referenced test, even if the item was well written, you may choose to exclude the item as it may not help discriminate differences in achievement between students.

You may wish to review items identified as easy items as well as difficult items. An easy item, one that almost everyone gets correct, may be too easy for the intended audience, test trivial material, or be written in a way that the correct answer is obvious. A difficult item may be unclear or contain more than one correct answer. These item writing mistakes lead to measurement errors and diminish the validity of the assessment results. We call this measurement error. To better understand the statistics you have calculated, you may wish to conduct an interview with students about what they were thinking when they read and answered the question. Remember, the purpose of testing students isn't to trick them, it is to measure the degree to which they have accomplished the intended learning objectives.

### DISCRIMINATION INDEX

This statistic, also known as Discriminating Power, is a measure of how well an assessment item can distinguish between high-performing and low-performing students. It is a statistical measure that indicates the relationship (i.e., correlation) between the overall score on a test and how well individuals answered a specific test item. It is expected that those who do better on the overall test will tend to do better on individual items. If not, there may be a problem.

Each item on a test will have a discriminating power. A positive discriminating power indicates that students who do well on the overall test tend to do well on the test item. Conversely, a negative discriminating power indicates that students who do well on the overall test tend to do poorly on that test item. This means that a high discriminating index indicates that the item effectively discriminates between high and low performers on the overall test. Conversely, a low discriminating index suggests that the item is less effective in differentiating between individuals and, in the case of a norm-reference test, may not contribute as much to the overall purpose of the test. These statistics may not be that useful when analyzing items on a criterion-referenced test. They are typically used for norm-referenced tests where differentiating between students is the goal. Very easy and very hard items will have little or no discriminating power. In a criterion-referend test, you may have several easy questions that test essential knowledge, as item selection is based on the importance of the material or skills being tested. The discriminating power for easy items (and difficult items) would be close to zero, but the item might still provide valuable information about the student's overall competence. In norm-referenced tests, discriminating power is important. Items with low discriminating power are typically excluded unless they test essential learning objectives. When reviewing items, questions with a negative discriminating index should be reviewed. A discriminating index around zero suggests it is equally likely that a student who did well on the overall test will get this item correct as a student who does poorly on the test.

We will discuss two ways to calculate discriminating power.

### 

### Method 1

**Step 1**: After administering and scoring the test, rank the students based on their total test scores (highest to lowest), then divide the students into three groups (upper, middle, and lower). We only use the upper and lower groups for this calculation.

Some suggest the upper and lower groups should each represent about 27% of the students. This choice is based on classical test theory, which suggests a 27%, 46%, 27% split is an optimal balance for the size of the groups to provide reliable discrimination indices. However, others argue that this method excludes important information from the middle group. In practice, different group sizes can be used depending on the context. the only thing you should do is make sure the upper and lower groups are the same size. For a small class (<30 students), using 10 in the upper group and 10 in the lower group makes for an easy calculation -- one you should be able to do in your head. You will probably only use this method for low-stakes, simple estimates of discriminating power as the results obtained from small samples need to be interpreted cautiously.

**Step 2**: Calculate the proportion of students in the upper and lower groups who got the item correct. Fun fact: This is the same calculation as the item difficult, only for the upper and lower groups by themselves. Fun fact: This is the same calculation as the item difficult, only for the upper and lower groups by themselves.

P(Upper)=R(Upper)/T(Upper)

P(Lower)=R(Lower)/T(Lower)

**Step 3**: The discrimination index is the difference between the two proportions.

D=P(Upper)-P(Lower)

Where D = Discriminating Power

Consider the following example: Suppose you had 120 students take a test. You score the assessment and divide the group into three. Suppose you decided to compare the top 50 and the bottom 50 students.  In the upper group, 42 students got the answer to a specific question correct.  In the lower group, only 35 students got the item correct. In this situation the discriminating power of the item is calculated as follows:

$$ P\_{upper} = \frac {42} {50}\; =\; .84 \;\;\;\;\;\;\;\;\;\ P\_{lower} = \frac {35} {50}\; =\;70; $$

The discrimination index is the difference between the two proportions.

$$ Discriminating \;Power (D) = .84\;-\;.72\;=\;.14 $$

This means the difference between the upper and lower groups was 14%. This is positive but not a large difference if this item was to be used in a norm reference assessment. Only 14% more of the upper group got the item correct than in the lower group.

### Method 2

This method for calculating the discrimination index uses a point biserial correlation. It measures the correlation between the item score and the overall test score.

A point biserial correlation is a statistical measure used to determine the relationship between a binary variable (e.g., correct/incorrect answers on a test item) and a continuous variable (e.g., overall test scores). In practice, the Person correlation is used because it works for dichotomously scored items (right and wrong) as well as those that might be awarded partial marks. This is important for situations where an individual item is out of more than 1 point or can be awarded partial credit. The resulting correlation is the discrimination index for the item. You likely won't want to do this calculation by hand. You could, but most statistical packages and spreadsheets will do them for you easily enough.

Most often, you will use this method to calculate discriminating power. While the upper and lower group comparison method is straightforward and widely used, this statistic measures item discrimination using all the available data. This method is more statistically robust than the simple discrimination index as it is less affected by item difficulty and uses all the available data. It provides a more precise measure of item discrimination and can be used with smaller sample sizes. However, care should still be taken when interpreting the discriminating power of an item is based on a small sample.

**INTERPRET THE DISCRIMINATION INDEX STATISTIC**

Mathematically, a correlation is represented by the Greek letter rho ("ρ") for an entire population, or "r" for a sampling. A correlation will range from -1 to +1. The interpretation is as follows:

* Positive Values (e.g., rpb>0r\_{pb} > 0r ​>0): Indicate that students who answered the item correctly tend to have higher overall test scores, suggesting the item is discriminating well.
* Negative Values (e.g., rpb<0r\_{pb} < 0r <0): Indicate that students who answered the item correctly tend to have lower overall test scores, which may suggest a problem with the item.
* Values Close to Zero: Indicate little or no discrimination, meaning the item does not effectively distinguish between high and low performers. Any difference is likely be due to random chance.

As it was with the item difficulty, the discriminating power does not describe the quality of the item. A well-written, perfectly good item can have a D of zero. Discriminating power is highly affected by the item's difficulty. When an item is extremely easy (P=1) or really hard (P=0), the discriminating power will be zero because everyone got the same score, and there is nothing to compare. In practice, we are only interested in the discrimination index when we are creating items for a norm reference test (and the results will be used to differentiate between students), or if an item has a negative discriminating power.

In situations where an item's discriminating power is negative, we have a concern. By this, we mean items with a discriminating power of negative 0.2 or lower.  Remember, a discriminating power around zero should be interpreted as zero. A slightly negative discriminating power could be the result of a small sample or the item's difficulty. In situations where the discrimination index is obtained from a small sample, you can get a large negative value for a number of simple reasons. We are more concerned when the item has a negative value that was obtained from a larger sample of student scores. A negative discrimination index suggests that the more knowledgeable the student is, the less likely it is that they will get the item correct—the opposite of what we would expect. This result would indicate that something is wrong with this item. It is not functioning as expected and should be examined.

We are also concerned with discriminating power when selecting items for a norm-referenced test. In this case, the item needs to be able to discriminate. Assuming the items are well written and measure important learning objectives, items with higher discriminating power are desirable for norm-reference tests. In practice, a D of 0.25 or higher would be acceptable. Items with a D < 0.20 would be considered unsuitable as they have poor discrimination power. Items with negative discriminating power are not suitable and highly questionable in terms of quality.

**DISTRACTOR ANALYSIS FOR MULTIPLE-CHOICE QUESTIONS**

Distractor analysis involves evaluating the incorrect options (distractors) in multiple-choice (MC) questions to determine their effectiveness and learn more about students (i.e., what they don't know and why). This process examines how often each distractor is chosen by students (response frequency) and identifies any patterns that might indicate problems with the question or the distractors themselves. Effective distractors should be plausible enough to be chosen by students who do not know the correct answer but not so confusing that they mislead students who understand the material. By analyzing distractor performance, educators can refine questions to improve their diagnostic value and ensure they accurately measure student knowledge and understanding.

In multiple-choice items, distractors play a crucial role. Distractors are designed to "distract" less capable test-takers (those who don't know the answer) from guessing the correct answer. The effectiveness of these distractors is key to accurately assessing student knowledge. The difficulty of a multiple-choice item is dependent on the quality of the distractors. Having plausible distractors makes an item more difficult.

Consider a scenario where a multiple-choice question offers four possible answers. If two of these options are blatantly incorrect and easily dismissed, students can narrow their choices to just two plausible options. This makes the item easier because their chances of guessing correctly changes from 25% to 50%, which may undermines the question's ability to assess true understanding. Poorly crafted distractors are those that are obviously incorrect, making them ineffective in evaluating student knowledge. In contrast, well-designed distractors challenge students to engage in critical thinking to determine the correct answer.

Effective distractors should attract more low-scoring students than high-scoring ones. This characteristic helps differentiate between students who have a solid grasp of the material and those who don't, thereby enhancing the assessment's discriminatory power. In essence, the quality of item distractors significantly impacts the overall effectiveness of multiple-choice assessments in measuring student knowledge and understanding.

Good distractors also have diagnostic value. Looking at response frequency—as well as the other item analysis components listed above—and noting the pattern of student errors can help teachers identify student misconceptions or knowledge gaps. This information can be used to design future instruction or remediation.

### Chapter Summary

* Item statistics help reviewers objectively evaluate how well each item functions within the assessment.
* These statistics are not an indicator of the item's quality. They provide descriptive data to help evaluate an item's effectiveness and can be used to make decisions about whether to keep, revise, or discard an item.
* In order to get these statistics, you should pilot-test the assessment with a minimum of 100-200 respondents. Small samples in a pilot test will often produce biased results.
* Item Difficulty is calculated as the proportion of correct responses, indicating how easy or difficult an item is.
* Discrimination Indexes (or discriminating power) measures how well an item differentiates between high and low performers.
* Discriminating power is required for items used in a norm-referenced test.
* Easy and hard items will have little or no discriminating power.
* A negative discrimination index may be an indication of a flawed item because more knowledgeable students were less likely to get the item correct.
* A distractor Analysis assesses the effectiveness of incorrect options in multiple-choice questions.
* The distractors in a multiple-choice item determine the effectiveness and difficulty of the item.
* Distractor analysis can expose student response patterns, revealing misconceptions or learning gaps that can inform future instruction and the need for remediation.

### Discussion Questions

1. Discuss the implications of poorly crafted distractors on the reliability and validity of assessment results. How can educators ensure that their distractors enhance the discriminatory power of test items?

Read this online at <https://edtechbooks.org/Assessment_Basics/item_analsyis>