8 Multiple-Choice Items

Always Pick Answer C and You'll Be Right About 25% of the Time

Difficulty Index ☺☺☺☺ (a bit difficult because of the math—but don't worry too much)

WHEN WE USE 'EM AND WHAT THEY LOOK LIKE

Multiple-choice items are the ones that you see all the time—most often as items on achievement tests used to assess some area of knowledge such as introductory chemistry, advanced biology, the written part of the Red Cross CPR test, the national boards in nursing, automotive mechanics, internal medicine, and so on. They are so easy to score, easy to analyze, and so easily tied to learning outcomes that multiple-choice items are, by far, the preferred way of testing achievement-oriented outcomes.

But beyond all the other great things about multiple-choice items, they are hugely flexible. And by that we mean that it is very easy to create an item that exactly matches a learning outcome.

In Chapter 13, we will emphasize the importance of taxonomies (such as Benjamin Bloom's taxonomy, which we will cover there) and how these hierarchical systems can be used to help you define the level at which a question should be written. Well, what multiple-choice items allow you to do is to write an item at any one of these levels. You can do this, for example, with short answer and completion items as well, but it is much more difficult. Why? Because multiple-choice items provide you with much more flexibility.

THINGS TO REMEMBER

Multiple-choice items are most often used to assess knowledge of a particular topic, and they typically appear on achievement tests.

For example, such items can target simple memorization, like the following multiple-choice item from an undergraduate research methods course:

1. Another name for research that occurs "post hoc" is
   a. experimental.
   b. correlational.
   c. quasi-experimental.
   d. historical.

Choice C is the correct answer.

There's nothing more required here than memorizing that another name for post hoc research is quasi-experimental. On the other hand, the following item taken from the tests for the same course taps a bit of higher-level thinking.

1. A nondirectional research hypothesis is similar to a directional hypothesis in that both
   a. specify the direction of the difference between groups.
   b. reflect differences between groups.
   c. are nonspecific regarding the direction of group differences.
   d. make no allusion to group differences.

Choice B is the right one here.

In this example, the test taker has to understand the similarities and dissimilarities of directional and nondirectional research hypotheses and select the ones that both types of hypotheses have in common—much more thinking is involved.

You can see in both of the above examples how a multiple-choice item can easily handle simple or complex ideas.

MULTIPLE-CHOICE ANATOMY 101

Multiple-choice items consist of three distinct parts: a stem and a set of responses, which in turn consists of the correct answer and

alternatives. Let's look at the following items from a high school geology class and identify these parts.

1. If the hanging wall has moved down, the fault is
   a. reversed.
   b. normal.
   c. strike-slip.
   d. indeterminate.

Choice B is it.

The **stem** sets the premise for the question and comes before any of the alternatives appear (in this example, the stem is "If the hanging wall has moved down, the fault is").

An **alternative** is a sentence or a part of a sentence that is used to complete the stem or answer the question being asked. For example, responses a through d above are all alternatives. And, there is only one **correct alternative,** which in this example is b. Sometimes, the correct alternative is called the **key** or the **key alternative.**

Distracters are a special type of alternative. These are the incorrect alternatives whose job it is to distract the test taker into thinking that the incorrect alternative is correct, but not to be so distracting that if the test taker knows the correct answer, he or she cannot identify it.

This is a pretty cool idea. Come up with a set of alternative answers to a question, one of which is correct and some of which are not. But, make sure that the ones that are not correct sound pretty good so that if someone knows only a bit about the topic, then the incorrect ones may seem attractive enough to select. However, for the test taker who is knowledgeable and understands the material, even though the distracters may appear to have some merit, they have less merit than the one correct answer. That's what you get for being smart.

THINGS TO REMEMBER

The key to a great multiple-choice question is a set of terrific distracters—those alternatives that could be, but are not, correct.

Why such a cool idea? Because the primary job of achievement tests that use multiple-choice items is to discriminate between those who know the material and those who do not. A really easy test and everyone gets 100%? (Oops—test didn't do a very good job of discriminating.) A really hard test and no one gets any right? (Oops—same problem.) Much more about this later under item analysis, because it is a really important aspect of judging how well (or poorly) multiple-choice items work.

Your job is to

- write terrific multiple-choice items that have clear, well-written stems that reflect the learning objective you want tested,
- create adequate alternatives that are all appealing, and
- make sure that within those alternatives, there are distracters that are almost as appealing as the correct answer.

Let's find out how to do this.

HOW TO WRITE MULTIPLE-CHOICE ITEMS: THE GUIDELINES

As the Great Oz said, the best place to start is at the beginning.

1. *List alternatives on separate lines and one right after the other.* It makes the items easier to read and is much less confusing, especially if the alternatives are of any length.

Here's the way it should not be done:

> 1. The first president of the United States was
>
> a. Washington b. Jefferson c. Lincoln d. Kennedy

And the way it should be is as follows:

> 1. The first president of the United States was
>
> a. Washington.
> b. Jefferson.
> c. Lincoln.
> d. Kennedy.

2. *Be sure that each item reflects a clearly defined learning outcome.* This is a biggie for sure. The key here is to be sure that your defined

learning objective (and, for example, its level of complexity) is reflected in the items you write.

3. *Be sure that the position of the correct alternative varies such that each position is represented an equal number of times.* Let's say you have created a 100-question multiple-choice test. By chance alone the test taker will get a score of 25% correct. And indeed, when those bubble scoring sheets are used, some test taker who did not prepare just goes right down the sheet, marking off whatever column he or she wants. If you denote a correct response for each item (A, B, C, or D) such that there is an equal chance of any one of them being correct by chance alone, you reduce the impact that guessing can have.

4. *Use correct grammar and be consistent.* This may sound like a no-brainer, but many multiple-choice items are poorly constructed in such a way that the test taker can easily figure out what the correct answer is or at least can eliminate some of the distracters. For example (and for those of you who are aspiring cooks),

1. Mirepoix is a
   a. Mixture of onions, carrots, and celery
   b. Ingredients for fondant icing
   c. Entrée served in France
   d. Alternative to flour used in baking

The only reasonable answer based on grammar alone is a. All the other alternatives are grammatically incorrect where a vowel "a" is followed by another vowel (such as "i," "e," or "a"). See, you don't even have to know a whisk from crème anglaise to answer the above question correctly.

Other Species of Multiple-Choice Items

In this chapter, we are concentrating on only one type of multiple-choice item—the one that has only one correct answer—but there are several other types of more complex multiple-choice items that you may want to consider. Some multiple-choice items are *context dependent,* where the questions can be answered only within the context in which they are asked, such as when the test taker is asked to read a passage and then answer a multiple-choice item about that passage. Then, there's the *best-answer* type of multiple-choice item, where there may be more than one correct answer, but only one that

is best. *Danger, Will Robinson!*—both of these may work quite well, but they should be used only with additional training and experience. If you're just starting, stick to the basic type of multiple-choice items where there is only one correct answer.

5. *The stem of the item should be self-contained and written in clear and precise language.* Everything the test taker needs to know about the question's topic should be contained in the stem so that he or she does not have to read redundant information in each of the alternatives. Remember, you want to know if someone has the knowledge that the question is tapping, not if they can read the same material over again quickly. For example, here's an item that doesn't contain enough in the stem and too much in each alternative.

1. New York City is the site of the next Olympics and
   a. has a new stadium for track and field.
   b. is building a new stadium for track and field.
   c. will be using only the Jets stadium in New Jersey.
   d. hasn't yet completed plans for where stadium events will take place.

The following question sets a better set of conditions for anyone who knows the correct response:

1. The current population of New York City is
   a. More than 15,000,000
   b. Less than 15,000,000
   c. More than 25,000,000
   d. Indeterminate

6. *Negatives, absolutes, and qualifiers in question stems are no-nos.* Not only are negatives confusing, but there also is rarely an absolute case of anything, so test takers can easily get confused. And qualifiers (such as *only, although, perhaps,* etc.) drive test takers nuts. In addition, almost any good test taker who might miss a question where an absolute is involved could probably argue for his or her answer. For example, here's a confusing multiple-choice item that contains negatives. This is so confusing that all or none of these answers could be correct.

1. Not only do cicadas come every 17 years, but they also never arrive
   a. during the rainy season.
   b. only if the temperature is sufficiently warm.
   c. whenever the ground is just about soft enough for them to merge.
   d. after June 1.

7. *Be sure that all distracters are plausible.* If a distracter does not even seem possibly correct, it is eliminated, and the value of the item decreases substantially because the likelihood of guessing correctly increases substantially. For example, in the following item, two of the four alternatives (a and c) are implausible, leaving b and d as possibilities. So, rather than a 25% chance of getting the item correct by guessing, the odds are now 50–50.

1. The square root of 64 is
   a. 64
   b. 8
   c. 642
   d. Infinity

8. *Items need to be independent of one another.* Multiple-choice items need to stand alone, and the answer on one item should not inform the test taker as to what the correct answer might be on another item. For example, an item early in a test may provide a clue or even an answer to an item that comes later.

THE GOOD AND THE BAD

Multiple-choice items have their advantages and disadvantages—let's review them, and they are summarized in Table 8.1.

Why Multiple-Choice Items Are Good

1. *Multiple-choice items can be used to measure learning outcomes at almost any level.* This is the big one, and we have mentioned it before. This allows multiple-choice items to be very flexible and to

Table 8.1 The Advantages and Disadvantages of Multiple-Choice Questions

Advantages of Multiple-Choice Items	Disadvantages of Multiple-Choice Items
• They can be used to measure learning outcomes at almost any level.	• They take a long time to write.
• They are easy to understand (if well written, that is).	• Good ones are difficult to write.
• They deemphasize writing skills.	• They limit creativity.
• They minimize guessing.	• They may have more than one correct answer.
• They are easy to score.	
• They can be easily analyzed for their effectiveness.	

be useful anytime you are sure that test takers can adequately read, and understand, the content of the question.

2. *They are clear and straightforward.* Well-written multiple-choice items are very clear, and what is expected of the test taker is clear as well. There's usually no ambiguity (how many pages should I write, can I use personal experiences, etc.) about answering the test questions.

3. *No writing needed.* Well, not very much anyway, and that has two distinct advantages. First, it eliminates any differences between test takers based on their writing skills. And, second, it allows for responses to be completed fairly quickly, leaving more time for more questions. You should allot about 60 seconds per multiple-choice question when designing your test.

4. *The effects of guessing are minimized, especially when compared to true-false items.* With four or five options, the likelihood of getting a well-written item correct by chance alone (and that's exactly what guessing is) is anywhere between 20% and 25%.

5. *Multiple-choice items are easy to score, and the scoring is reliable as well.* If this is the case, and you have a choice of what kind of items to use, why not use these? Being able to bring 200 bubble scoring sheets to your office's scoring machine and having the results back in 5 minutes sure makes life a lot easier. And, when the scoring system is more reliable and more accurate, the reliability of the entire test increases.

6. *Multiple-choice items lend themselves to item analysis.* We'll talk shortly about item analysis, including how to do it and what it does. For now, it's enough to understand that this technique allows you to further refine multiple-choice items so that they perform better and give you a clearer picture of how this or that item performed and if it did what it was supposed to do. For this reason, multiple-choice items can be diagnostic tools to tell you what test takers understand and what they do not.

Why Multiple-Choice Items Are Not So Good

Those just-mentioned advantages sound pretty rosy, but there's a down side as well.

Guess What? No, Don't Guess . . .

You already know from the reading in this chapter that guessing may have a significant impact on a test taker's score, which is why we need to consider some kind of correction for guessing. Remember that the standard four-alternative multiple-choice item already has a probability of being correct one out of four times by chance alone, or 25%, or 0.25. So, we'd like to even the playing field and get a more accurate picture as to what's going on.

The correction for guessing looks like this:

$$CS = R - \frac{W}{n-1}$$

where

CS = corrected score

R = number of correct responses

W = number of wrong responses

n = the number of choices for each item

For example, on a 60-item multiple-choice test with four alternatives, you can expect a score of 15 by chance alone, right? ($0.25 \times 60 = 15$). Russ gets 15 correct on this 60-item test, and Sheldon gets 40 correct. How can we adjust these scores such that Sheldon's performance is encouraged (because it is way above chance) and that Russ is gently punished for lots of guessing? When we are correcting scores (using the

above formula), it turns out that Russ's new score is 15 – (45/3), or 0!, and Shel's is 40 – (20/3), or about 33. Russ is clearly "punished" for guessing.

1. *Multiple-choice items take a long time to write.* You can figure on anywhere between 10 and 20 minutes to write a decent first draft of a multiple-choice item. Now, you may be able to use this same item in many different settings, and perhaps for many different years, but nonetheless it's a lot of work. And, once these new items are administered and after their performance analyzed, count on a few more minutes for revision.

2. *Good multiple-choice items are not easy to write.* Not only do they take a long time, but unless you have very good distracters (written well, focused, etc.), and you include one correct answer, then you will get test takers who can argue for any of the alternatives as being correct (even though you think they are not), and they can sometimes do this pretty persuasively.

3. *Multiple-choice items do not allow for creative or unique responses.* Test takers have no choice as to how to respond (A or B or C or D). So, if there is anything more they would like to add or show what they know beyond what is present in the individual item, they are out of luck!

4. *The best test takers may know more than you!* Multiple-choice items operate on the assumption that there is only one correct alternative. Although the person who designs the test might believe this is true, the brightest (student and) test taker may indeed find something about every alternative, including the correct one, that is flawed.

Multiple-Choice Items: More Than Just "Which One Is Correct"

Throughout this chapter, we are emphasizing the importance of the type of multiple-choice item that has only one correct alternative. But, there are others with which you should be familiar.

- *Best-answer multiple-choice items.* These are multiple-choice items where there may be more than one correct answer, but only one of them is the best of all the correct ones.
- *Rearrangement multiple-choice items.* Here's where the test taker arranges a set of items in sequential order, be it steps in a process or the temporal sequence in which something might have occurred or should occur.

- *Interpretive multiple-choice items*. Here, the test taker reads through a passage and then selects a response where the alternatives (and the correct answer) all are based on the same passage. Keep in mind that although this appears to be an attractive format, it does place a premium on reading and comprehension skills.
- *Substitution multiple-choice items*. This is something like a short answer or completion item (see Chapter 6), but there are alternatives from which to select. The test taker selects those responses from a set of responses that he or she thinks answers the question correctly.

But in spite of these many choices, the more sophisticated the form of the question becomes (such as any of those described previously) the more cautious and careful you have to be to make sure that you are creating a question that is good and works as intended.

ANALYZING MULTIPLE-CHOICE ITEMS

OK—you've got your multiple-choice items created, you understand very well the advantages and disadvantages and feel pretty comfortable that the test you just gave did what it should have—separate those who know the material from those who do not. Now's the time to find out if the questions you created and are using "work."

Multiple-choice items (as do other types) allow for an in-depth analysis of whether the item did what it was supposed to—discriminate between those who know the material and those who do not. This can be done through the completion of an **item analysis,** which consists of generating two different indexes for each item: a difficulty index and a discrimination index.

By looking at how well the distracters work, the multiple-choice question format can be diagnostic in nature. For example, if all the people who did very well on the test selected the incorrect distracter, there's got to be a good reason and one well worth exploring.

Let's explore both of these using the following item (for all you baseball lovers) as an example. Let's say this is Item 14 on a test that contains 25 multiple-choice items, and 50 people took the test.

1. Who was the first president to dedicate a new baseball stadium?

   a. John F. Kennedy
   b. Lyndon B. Johnson
   c. Calvin Coolidge
   d. Chester Arthur

It was Lyndon Johnson.

First, we are going to show you the final scores for all 50 people as shown here.

Individual	Score	Individual	Score	Individual	Score	Individual	Score	Individual	Score
#1	21	#11	25	#21	23	#31	24	#41	25
#2	24	#12	21	#22	24	#32	21	#42	20
#3	23	#13	21	#23	13	#33	21	#43	12
#4	18	#14	12	#24	16	#34	11	#44	14
#5	15	#15	9	#25	14	#35	9	#45	14
#6	21	#16	23	#26	13	#36	19	#46	9
#7	24	#17	7	#27	4	#37	18	#47	23
#8	15	#18	21	#28	13	#38	3	#48	21
#9	18	#19	22	#29	21	#39	18	#49	24
#10	19	#20	3	#30	24	#40	5	#50	8

So, individual 12 got 21 correct out of 25, and individual 40 got 5 correct out of 25.

Now, here are the results for (only) item 14 that we are using as an example. You can't see these individual results in the above listing—you can only see what's there, which is the score for each individual test taker.

Alternative	Total Times Selected
a	4
b	26
c	8
d	12
Total Responses	50

So, of all 50 people who took the test, 26 selected the correct response B, and 24 (4 + 8 + 12) selected one of the alternatives.

As you learned earlier, the rationale behind doing an item analysis is to find out how well an item discriminates between those who know the material and those who do not. In order to do this, we will create two groups of test takers—yes—those who know the material and those who don't.

How do we do this? Simple. We order all the test scores from highest to lowest and take a percentage of the top performers and call that our high group and take a percentage of low performers and call that our low group. What percentage of the group do we take? 27%. Why 27? Because various studies have shown that this percentage maximizes the difference between those who know the material and those who do not—just what we want to find out.

27%? 50% Could Do Just Fine

OK—you've seen the recommendation—take the top and bottom 27% from the groups that are used in an item analysis. But, for all practical purposes, and especially for the classroom teacher, splitting the entire group in half (so you have the high group and the low group, each consisting of 50% of the entire group) is just fine, especially because classrooms often have relatively small numbers of students, and 27% would result in just too small a number to be useful in the analysis.

Here's how we create these two groups.

1. Place all of the scores for all test takers in descending order (from highest to lowest).
2. Select the top 27% of scores and identify that as the high group. In this case, 27% of 50 is 14, so the highest 14 test takers constitute the high group.
3. Select the bottom 27% of scores and identify that as the low group. In this case, 27% of 50 is 14, so the lowest 14 test takers constitute the low group.
4. Here are the final scores of the high and the low groups.

#11	25	High Group
#41	25	High Group
#2	24	High Group
#7	24	High Group
#22	24	High Group
#30	24	High Group
#31	24	High Group
#49	24	High Group
#3	23	High Group
#16	23	High Group
#21	23	High Group
#47	23	High Group
#19	22	High Group
#1	21	High Group
#26	13	Low Group
#28	13	Low Group
#14	12	Low Group
#43	12	Low Group
#34	11	Low Group
#15	9	Low Group
#35	9	Low Group
#46	9	Low Group
#50	8	Low Group
#17	7	Low Group
#40	5	Low Group
#27	4	Low Group
#20	3	Low Group
#38	3	Low Group

5. Finally, for these 28 different test takers (and remember, we are doing an item analysis, so we are looking at just one item at a time, which is #14 in our example), we get the following results, again for that one item only.

	Alternative				
Item 14	a	b*	c	d	Total
High Group	2	21	2	4	29
Low Group	2	5	6	8	21
Total	4	26	8	12	50

So, on Item 14 (and only Item 14), a total of 26 test takers selected Alternative B (which is correct); 21 of those were in the high group, and 5 of them were in the low group. Similarly, a total of 12 test takers selected Alternative D (which is an incorrect alternative); 4 of those were in the high group, and 8 of them were in the low group. For our upcoming item analysis, all we care about are the responses to Alternative B.

OK, let's get to the analysis.

THINGS TO REMEMBER

One of the most common mistakes made by new users of item analysis is that they believe the analysis is of the *entire test* and not individual items. Indeed, the analysis is for each item, one at a time. So, if you have a 50-item test, then a difficulty index and a discrimination index can be computed for *each* item on that test. One might average these two indexes across all items to get an overall set of numbers, but the purpose of this item analysis is to get feedback on how good an individual item is, then to revise that item so that it will be even better next time you use it.

Computing the Difficulty Index

The **difficulty index** tells us how many people got the item correct. It is a percentage reflecting the number of correct responses in both the high and low groups. Here's the formula:

$$D = \frac{N_h + N_l}{T}$$

where

D = difficulty level

N_h = the number of correct responses in the high group

N_l = the number of correct responses in the low group

T = the total number of responses to the item

If we plug in the number to the preceding formula, we find that the difficulty level for Item 14 is 52%.

$$D = \frac{21+5}{50} = 52\%$$

Fifty-two percent (52%) of the responses to Item 14 were correct. If the difficulty level were equal to 100%, it would mean that everyone got the item correct (far too easy), and if the difficulty level were equal to 0%, it would mean that everyone got the item incorrect (far too hard).

And, that's the story of the difficulty index—but only half the entire story of item discrimination. Let's move on to the discrimination index.

Computing the Discrimination Index

Where the difficulty index is a measure of the percentage of responses in the high and low groups that are correct, the **discrimination index** is a measure of how effectively an item discriminates between the high and low groups.

Here's the formula:

$$d = \frac{N_h - N_l}{(.5)T}$$

where

d = discrimination level

N_h = the number of correct responses in the high group

N_l = the number of correct responses in the low group

T = the total number of responses to the item

If we plug in the numbers to the preceding formula, we find that the discrimination level for Item 14 is 0.64.

$$d = \frac{21 - 5}{25} = .64$$

As you can see, the discrimination index is expressed as a decimal.

Right off the bat, you should recognize that when *d* is positive, more people in the high group than in the low group got the item correct (the item is doing what it should). And, if *d* is negative, it means that more people in the low group than in the high group got the item correct (the item is not doing what it should).

Another Way to Compute the Discrimination Index

The method we show you in this chapter to compute *d* (the discrimination index) for any one item is straightforward and pretty easy to calculate. There's another way you may see mentioned and one that is especially good for longer tests. It's called the point biserial method, which is the correlation between a dichotomous variable (right or wrong for the item under analysis) and a continuous variable (the test score). Here's the formula:

$$d = \left[\frac{\bar{X}_1 - \bar{X}}{S_y}\right]\sqrt{\frac{P_X}{1 - P_X}}$$

where

d = discrimination index

$\bar{X}_1$ = the mean score on the test for all those test takers who got the item correct

$\bar{X}$ = the mean score for the test

S_y = the standard deviation for the test

P_X = the proportion of people who got the item correct

So, for example, suppose that 67% of test takers got Item 17 correct, and the mean score on the test was 83 and the standard deviation 7.5. For those who got Item 17 correct, also suppose that the mean score for their tests was 79. The formula for *d* would then look like this:

$$d = \left[\frac{79 - 83}{7.5}\right]\sqrt{\frac{.67}{.33}} = -.757$$

Not a great item because the negative value indicates that those who knew less overall did better on this item than those who knew more overall.

THINGS TO REMEMBER

Multiple-choice questions are used for many reasons, but among the most important is that they lend themselves to a thorough quantitative analysis using the item analysis we talk about in this chapter.

A fun and useful little applet (kind of a baby application) is where you can use Excel to compute *D* and *d*. Here are the data, which you saw on page 147 in a spreadsheet:

	Alternative				
	a	**b***	**c**	**d**	**Total**
High Group	2	21	2	4	29
Low Group	2	5	6	8	21
Total	4	26	8	12	50

And here are the same data with *D* and *d*.

	Alternative				
	a	**b***	**c**	**d**	**Total**
High Group	2	21	2	4	29
Low Group	2	5	6	8	21
Total	4	26	8	12	50
D	52%				
d	0.64				

And now the formulas used to compute *D* and *d,* and you can just copy them and plug in your data for each item in almost any spreadsheet application.

$D = (C3 + C4)/50$

$d = (C3 - C4)/(0.5*F5)$

How the Difficulty and Discrimination Index Get Along: Quite Well, Thank You

As we move to understanding the relationship between the difficulty and discrimination indexes and what that means for any one item, remember that the perfect item has a discrimination index of 1.00, which means that 100% of the test takers in the high group got it correct and 100% of the test takers in the low group got it incorrect. And in order for this to occur, guess what? Read on . . .

The perfect item has two characteristics. First, the item is sufficiently difficult such that 50% of all the test takers get it correct and 50% of all the test takers get it incorrect. Guess which 50% get it correct?

That's right. The second condition is that all the test takers in the high-scoring group (the "smarter" folks) get the item correct and all the test takers in the low-scoring group get the item incorrect.

Now for the prize question. If the two conditions are met, then what is the value of *D* and *d*? That's right . . . *D* = 50% (half got it right and half got it wrong), and *d* = 1.00 (the half that got it right scored best on the test). And that's how they get along.

But there's more—in fact, the *only* possible way for an item to have perfect outcomes for the item analysis is if these two conditions are met, because *discrimination level is constrained by the value of the difficulty level*. That's right, the only way you can have perfect discrimination is if the difficulty level is equal to 50%. The more the difficulty level varies from 50%, the less well the item discriminates.

For example, the only way that the highest half of the total number of test takers can get the item right is if 50% of all the test takers get it right and the other 50% get it wrong. Similarly, if the item doesn't perfectly discriminate between the two groups, it means that some test takers in the high group got the item incorrect and some of the test takers in the low group got the item correct—not what we want.

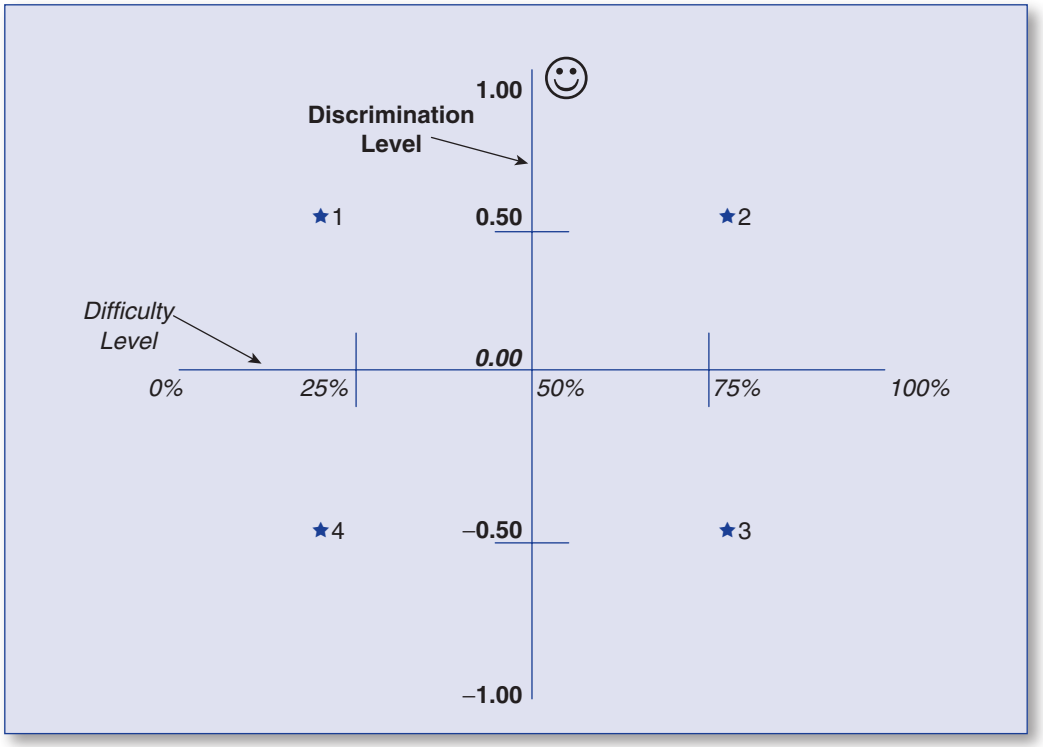
Now, we all know that nothing's perfect, so let's look at four different items (1, 2, 3, and 4) with different difficulty and discrimination indexes and try to understand what they are like.

You can see each of them in Figure 8.1 represented by a star (★) and an item number (★1 through ★ 4). And, to make everyone's day just perfect, there's the happy face ☺, indicating the absolute perfect item with a difficulty level of 50% and a discriminating level of +1.00. Isn't that cute?

Item ★1 has a difficulty level of about 25% and a discrimination level of about 50%. Not bad—about 25% of the test takers got it correct and it positively discriminates, meaning that more in the group that got better scores got it correct than those in the lower group. This item is doing a pretty good job.

Item ★2 has a difficulty level of about 75% and a discrimination level of about 50%. Like Item ★1, this item positively discriminates, but, with a difficulty level of 75%, it's getting a bit too easy.

Figure 8.1 Difficulty and Discrimination Indexes



Item ★3 isn't one that we particularly want to take home and show our friends and relatives—nor, of course, use again on a test. With a negative discrimination level around –.50, which means that more test takers in the low group got it correct than did those test takers in the high group, and a pretty high difficulty index of 75%, indicating that it may be a bit too easy—it's time to go back to the drawing board.

THINGS TO REMEMBER

Remember that the only way to maximize an item's discrimination is to have 50% of the test takers get it correct, and those 50% be the ones who have the highest performance.

Finally, Item ★4 does not discriminate very well either (it's also negative), and it's too hard—only 25% of the entire group got it correct. Yuk.

For a perfect item, then, we shoot for $D = 50\%$ and $d = 1.00$. Do we reach those levels of perfection? Not usually, but we surely can come close. But how do we come closer? Read on.

The reason why an item fails to discriminate and has an unacceptable level of difficulty is because one or more of the rules for writing items (discussed earlier in this chapter) has been violated. For example, if you do not have one clearly correct answer, it's unlikely that there will be uniform agreement among the test takers in the high group as to which is correct. Similarly, if you have distracters that are not effective and are all implausible, the item becomes too easy, and if it is too easy (or too hard, for that matter), then discrimination is restricted (look at Figure 8.1 and you can see that if difficulty level is around 0% or 100%, then there can be little discrimination).

But most of all, if you want great items, create clearly correct answers and have terrific distracters. That ensures that the folks who know the material show such on a test and the folks who don't know it show their skills (or lack thereof) as well!

Each time you create a new multiple-choice item, place the stem and alternatives on the front of an index card. Each time you use the item in a test, enter the difficulty and discrimination indexes on the back of the card along with the date the item was used and any other special circumstances (pop quiz, final exam, pretest, etc.). Then, the next time you want to use that item again, you know what it is that needs to be changed. If the difficulty and discrimination indexes are not what you want them to be, you can fine tune the item more and more. With multiple uses of the same item that is constantly modified, you can eventually come up with very well-written and effective multiple-choice items.

SUMMARY

There's nothing like a dip in the lake on a hot and humid summer day or that favorite dessert of yours on Thanksgiving. And, there's nothing like a multiple-choice item (or a set of them, actually) to give you a fairly accurate and unbiased assessment of someone's level of knowledge about a particular area. Multiple-choice items have been used for decades with great success, and the simple and straightforward tools of item analysis have led to refinements that make these types of items simply the best for finding out what others know.

TIME TO PRACTICE

1. What are the three parts of a multiple-choice question, and what purpose does each serve?
2. In your area of expertise (your area of professional training, a hobby, or just personal interest), write five multiple-choice questions. Then, share these five with a class colleague, and go through the criteria presented in this chapter to evaluate how well each question meets the set of criteria on our cheat sheet for writing multiple-choice items.
3. Evaluate the following item based on the difficulty and discrimination indexes.

	Alternative			
	a	b	c*	d
High Group	12	6	25	7
Low Group	18	15	5	12

4. Jim is pretty smart, but he didn't prepare very well for his multiple-choice test and guessed his way to 23 out of 50 items correct with five alternatives to each item. Chris is also pretty smart as well, and she studied and got 43 correct. What are their corrected scores for guessing?
5. In order for an item to be perfect (discriminate efficiently and have the appropriate level of difficulty), two conditions have to be met. What are those conditions?
6. About how much time should you leave yourself to write a 50-item multiple choice test?
7. What can you do to better an item's discrimination and difficulty indexes?

8. Write whether you would approve as is or need to revise the following multiple-choice questions. If you need to revise a question, state why.

   A. On average, a newborn sleeps for how many hours each day?
      1. 2
      2. 6
      3. 16
      4. 25

   B. The cochlea is located in which part of the ear?
      1. inner ear
      2. middle ear
      3. outer ear

   C. The minority of nutritionists suggest having fewer than three
      1. tablespoons of sugar per day
      2. desserts per week
      3. salt in your diet
      4. meals each day

   D. The D-Day landings began on June 6 of which year?
      1. 1941
      2. 1942
      3. 1943
      4. 1944

WANT TO KNOW MORE?

Further Readings

- Elstein, A. S. (1993). Beyond multiple-choice questions and essays: The need for a new way to assess clinical competence. *Academic Medicine, 68,* 244–249.

You may think that multiple-choice questions are great, but not everyone shares that perspective. This article talks about alternatives and why other techniques might be a better choice for assessing knowledge than multiple-choice questions.

- McG, H. R., Brown, R. A., Biran, L. A., Ross, W. P., & Wakeford, R. E. (1976). Multiple choice questions: To guess or not to guess. *Medical Education, 10*(1), 27–32.

This is a pretty old study, but one that raises a question that every multiple-choice test giver and test taker should ask (and answer). These researchers found that the "don't know" option in multiple-choice question papers favors the bold and test-wise student.

And, on the Internet

- From the good people at the University of Wisconsin, Eau Claire at http://www.uwec.edu/geography/Ivogeler/multiple.htm, lots of ways to help increase your score on a multiple-choice exam—strategies for taking them and doing well.
- And more of the same from George Washington University at http://gwired.gwu.edu/counsel/asc/index.gw/Site_ID/46/Page_ID/ 14561/.

And, in the Real (Testing) World . . .

Real World 1

The type of item used in a test can actually reflect, quite accurately, what is learned. This study compared the effect of multiple-choice items against that of constructed-response items. A direct comparison of the responses to the items in the two tests showed that only 26% of the responses were the same, suggesting that most of what the multiple-choice items measured was directly dependent on the item format. The study also found consistency in the response patterns between the tests among experimental groups of participants, who sat different option number formats of the multiple-choice test, pointing to the possibility of a general effect of multiple-choice items in testing the learning of structure in second and foreign languages.

Want to know more? Currie, M., & Thanyapa Chiramanee, T. (2010). The effect of the multiple-choice item format on the measurement of knowledge of language structure. *Language Testing, 27*, 471–491.

Real World 2

An analysis was conducted on the effects of multiple-choice and open-ended formats on reading and listening test performance for several primary (L1) and secondary (L2) language learners.

Multiple-choice formats are easier than open-ended formats in L1 reading and L2 listening, with the degree of format effect ranging from small to large in L1 reading and medium to large in L2 listening. Format effects favoring multiple-choice formats were observed consistently.

Want to know more? In'nami, Y., & Koizumi, R. (2009). A meta-analysis of test format effects on reading and listening test performance: Focus on multiple-choice and open-ended formats. *Language Testing, 26,* 219–244.

Real World 3

This article delves into the reasoning behind popular methods for analyzing the raw data generated by multiple-choice question (MCQ) tests and why this reasoning is not always appreciated. The authors discuss three options for processing the raw data produced by multiple-choice test items and discusses the pros and cons.

Want to know more? Scharf, E. M., & Baldwin, L. P. (2007). Assessing multiple choice question (MCQ) tests—A mathematical perspective. *Active Learning in Higher Education, 8,* 31–47.