Phillips Associates
on M&E
3rd in the series

Writing Test Questions That Actually Measure Something

"Just because someone has taken tests doesn't make him or her a good test item writer."

Writing good test questions is both an art and a science. Moreover, just because someone has taken tests doesn't make him or her a good test item writer. However, writing good test items is a critical skill for L&D professionals whose personal credibility and ability to prove the value of a training program depend upon it.

Unfortunately, writing good test questions is not something many L&D professionals do well, and the result often is the creation of test items that contain clues as to the correct answer or are overly difficult or tricky and discourage test takers from getting the right answer. In either case, the result is an invalid test – one that doesn't measure what it is supposed to and is unfair either to the test taker or the test taker's organization.

Also, invalid tests put L&D professionals at risk by creating situations where it appears:

1. That learning took place when it didn't (the test items contained clues as to the correct answer).

   OR

2. That learning didn't take place when it did (the test items were tricky or overly difficult and discouraged test takers from getting the correct answer).

In the first situation, business executives may question why participant job behavior didn't change (Level 3), or business results didn't improve (Level 4) if learning improved. In the second situation, business executives may be upset that time and money was wasted on training where participants didn't learn anything. In either case, your reputation and credibility as an L&D professional are on the line and sure to suffer.

However, both of these situations are avoidable by conducting a test item analysis to determine whether or not each of your questions is "good."

There are three test item statistics you can use to evaluate the quality of your test items: difficulty index, p-values, and point-biserial correlation.* To apply these statistics, you first need to create your test and then administer it to a group of at least 25–30 program participants.

"*Good* test questions generally have a difficulty score between .30 and .70…"

DIFFICULTY INDEX

The difficulty index, as the name implies, is a measure of how many program participants answer a particular test question correctly. The statistic, expressed as a percentage such as .60 or 60%, indicates the percent of program participants who responded to the question correctly. "Good" test questions generally have a difficulty score between .30 and .70, where the range is from .00 (no one answered the item correctly) to 1.00 (everyone responded to the question correctly). Test items that fall outside the 30/70 range are candidates for possible revision as being either too easy or too difficult. One exception is if you created a mastery test in which case you would be looking for difficulty index scores in the .90 or 90% range.

P-VALUE

"The value of p-value data is that it enables you to conduct a response option analysis when using multiple-choice test questions, to see if any of the responses are being over or under selected. "

The p-value is similar to the difficulty index but indicates what percent of test takers chose each of the incorrect response options rather than the percent of test takers who answered the item correctly. For example, in the case of a multiple-choice test question with four response options, each of the three incorrect responses would have its own a p-value.

For illustration, imagine a multiple-choice test question with a difficulty index of .60 and the following p-value scores for each of the incorrect responses: .10, .15, and .15. (Note: that the p-values plus the difficulty index sum to 1.00 or 100% of the test takers.) The value of p-value data is that it enables you to conduct a response option analysis when using multiple-choice test questions, to see if any of the responses are being over or under selected. An over selected response alternative, when the option is the correct answer, indicates that the question either is too easy or that none of the other response options are considered plausible. An over selected incorrect response option demonstrates that either the item is misleading or that the

response option needs rewording so that it is less attractive (less like the correct response). The existence of under selected response options also increases the odds of a test taker guessing the right answer. For example, the chances of guessing the correct answer to a multiple-choice test question with four-response choices go from 25% to 33% with the existence of one under selected response option and 50% with two under selected response options.

POINT–BISERIAL CORRELATION

"Most test creation experts regard the point-biserial correlation as the single most useful test item analysis statistic "

Most test creation experts regard the point-biserial correlation as the single most useful test item analysis statistic. The statistic correlates test-takers' performance on a single test item with their overall test scores. In short, the statistic shows if test-takers who scored high on the test overall also answered the particular test question correctly. Each test item will have a point-biserial correlation score ranging from +1.00 to -1.00. Of specific concern are negative point-biserial scores because they indicate that test-takers who generally scored high on the test overall missed the item and test-takers who generally did poorly on the test overall got the item right. Any test items with a negative point-biserial score should be investigated immediately to determine the source of the problem and then rewritten.

If you're creating a Level 2 knowledge test, writing test items is only half your job. The other half is being sure the questions you have written are "good" and actually measure something. By using the three test item statistics described above to analyze the test item you've written, you can ensure that you are successfully performing the second part of your job.

*Adapted from Shrock & Coscarelli, Criterion-Referenced Test Development, John Wiley & Sons, 2007.

202006022

Share Phillips Associates *with your colleagues*

Corporate Workshops

OFFER THESE WORKSHOPS TO YOUR ENTIRE LEARNING & DEVELOPMENT TEAM

Mastering M&E
2-Day Workshop

Provide your L&D team with the latest guidelines and hands-on techniques for creating valid, scientifically sound Level 1, 2, 3, and 4 evaluations that produce data perceived by business executives as both credible and valuable.

Boost Training Transfer using Predictive Learning Analytics
2-Day Workshop

Equip your L&D team with a systematic, credible and repeatable process for maximizing the value of your learning investments by boosting training transfer.

Crack the Code of Test Question Design
1-Day Workshop

Equip your L&D team with practical tips and specific techniques for creating quizzes and tests that actually measure something.

Survey Magic: Capturing Level 3 Evaluation Data
1-Day Workshop

Equip your L&D team with a five-step process for creating Level 3 surveys that capture on the job behavior change.

Presentations for Professional Meetings & Industry Events

IDEAL FOR CORPORATE L&D TEAMS AND INTERNAL LEARNING CONFERENCES

Ken Phillips is available to present on the following topics. All include the valuable, "how-to" tips and hands-on measurement and evaluation techniques that L&D professionals crave—and can't find anywhere else! All topics can be delivered as 75-90 minute programs or webinars.

- **Power up your Level 1 Evaluations** and Gain Surprisingly Useful, Valued Data
- **Take Your Level 2s Up a Notch:** The Magic of Well-written Multiple Choice Test Questions
- **Capture Elusive Level 3 Data:** The Secrets of Survey Design
- **Business Results Made Visible:** Design Proof Positive Level 4 Evaluations
- **Boost Training Transfer Using Predictive Learning Analytics™ (PLA)**
- **Going The Distance: Making Sense Out of Level 1–4 Evaluation Data**

Contact Ken Phillips at 847.231.6068 or ken@phillipsassociates.com

Ken Phillips, CPLP, delivers all programs and workshops in his signature style: professional, engaging, and approachable.

Ken is founder and CEO of Phillips Associates, and the creator and chief architect of the Predictive Learning Analytics™ (PLA) learning evaluation methodology. He has more than 30 years experience designing learning instruments and assessments and has authored more that a dozen published learning instruments. He regularly speaks to Association for Talent Development (ATD) groups, university classes, and corporate L&D groups. Since 2008, he has spoken at the annual ATD International Conference on topics related to measurement and evaluation of learning.

Phillips Associates

34137 N. Wooded Glen Drive | Grayslake, IL 60030

847.231.6068 or ken@phillipsassociates.com