Developing Valid Level 2 Evaluations*

"…many L&D professionals … often develop questions that contain clues as to the correct answer or items that are overly difficult and discourage participants from getting the right answer "

Collecting anecdotal information regarding training effectiveness is a start. However, to establish real credibility and prove value, L&D professionals, need to measure whether or not participants learned something – a Level 2 evaluation.

Unfortunately, conducting Level 2 evaluations is not something many L&D professionals do well because they are uninformed in the art and science of test creation. As a result, they often develop questions that contain clues as to the correct answer or items that are overly difficult and discourage participants from getting the right answer. In either case, the result is an invalid Level 2 evaluation – one that doesn't measure what it is purported to and is either unfair to the learner or the organization.

Invalid Level 2 evaluations also put L&D professionals at risk by creating situations where it appears that either:

1. Learning took place when it didn't (the assessment contained clues as to the correct answers). OR

2. Learning didn't take place when it did (the evaluation was difficult or tricky and discouraged learners from getting the right answers).

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 occurred. In the second situation, executives may question why time and money was wasted on training if 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 in the eyes of company executives. However, both of these situations can be avoided by merely following a set of proven test creation guidelines and tips that result in the creation of valid Level 2 evaluations.

PHILLIPS ASSOCIATES

"Correctly answering application questions not only assesses whether participants truly understand the material taught but also do they know how to apply it and, after all, isn't this what a Level 2 evaluation should be measuring?"

TOP TEN TEST CREATION GUIDELINES

There are ten guidelines to follow when developing Level 2 evaluations to ensure that the test items you create are fair to both the learner and the organization. These guidelines apply regardless of what types of questions you create -- multiple choice, true/false, matching or fill-in-the-blank.

1. **Focus on creating test items that test for understanding, <u>not</u> just knowledge or recall.**

(This guideline is courtesy of Matt Allen, an I/O consultant with HumRRO a Washington D. C. based human and organizational performance research-consulting company.) For example, consider the following test questions:

QUESTION	COMMENTS
• *What do the letters TV stand for?* • *What is the primary function of a TV?*	Recall focused questions such as these test for knowledge, but not understanding. Moreover, if learners merely "know" something, but don't understand it, they're also likely to forget it shortly after the learning program is over.
• *What physical principle is used to display images on a TV?*	While choosing the correct answer to this question requires more in-depth knowledge of TVs, it's still a recall-based question.
• *Your TV is not working. What's the most likely cause of the problem given the following symptoms...?*	Testing for real understanding requires the use of application-focused questions such this one.

As you can see, correctly answering each of these questions requires an increasingly greater mastery of the topic. Correctly answering application questions not only assesses whether participants truly understand the material taught but also do they know how to apply it and, after all, isn't this what a Level 2 evaluation should be measuring?

PHILLIPS ASSOCIATES

"If you're having difficulty gaining business executive support for your evaluation efforts, position it as reinforcement, and not just evaluation."

2. Where appropriate, use Level 2 evaluations for reinforcement as well as evaluation.

Administering a knowledge test a week or two following the conclusion of a learning program positions it to serve as both reinforcement and evaluation. It also increases the credibility of the results. Specifically, business executives expect participant knowledge test scores to be high immediately following a learning program. However, if high scores are obtained a week or two after the training, this strengthens the case that participants retained what they learned and adds credibility to the results. Also, if you're having difficulty gaining business executive support for your evaluation efforts, position it as reinforcement, and not just evaluation. This slight reframing often makes it easier to secure their buy-in.

3. Group test questions by topic for scoring, but randomize for administration.

"Sometimes one question on a program topic provides a clue as to the correct answer to another question on that topic."

The reason for this is because sometimes one question on a program topic provides a clue as to the correct answer to another question on that topic. However, randomizing all the items related to the same topic area eliminates this possibility. Just remember, when scoring the test, you'll want to re-group the questions into their respective topic areas to identify any trends such as having all or most of the items in a particular area answered incorrectly. This outcome suggests that either the topic was not presented correctly or that additional insight about the topic is required.

4. All test items should discriminate between participants who know the material from those who don't.

"Check all multiple-choice questions to be sure none of the response choices are over or under selected."

· Check all multiple-choice questions to be sure none of the response choices are over or under selected. Response alternatives that are over or under selected should be revised either to make them less attractive or more attractive.

· True/false questions should be rewritten if everyone (or nearly everyone) chooses either True or False and it's the wrong answer. Remember the goal is to create test items that discriminate between learners who know the material taught from those who don't.

"It's not fair to test participants on material that wasn't covered or only covered in a cursory way."

5. Avoid compound questions that ask for more than one thing.

(This guideline is also courtesy of Matt Allen.) Learners find compound questions confusing and view them as unfair. Also, avoid questions with compound answers. For example:

QUESTION	COMMENTS
What do the letters in the acronym ADDIE stand for? A *Analyze, Design, Develop, Implement and Evaluate* B *Analyze, Design, Develop, Integrate and Evaluate* C *Analyze, Design, Develop, Implement and Execute* D *Analyze, Design, Deploy, Implement and Evaluate*	Savvy test takers will pick the first option, which is the correct answer, because Integrate, Execute and Deploy only appear in one option and Analyze, Design, Develop, Implement and Evaluate appear in multiple options.

The solution: make both the question and the options short and to the point.

6. Don't test participants on concepts or material not covered in the learning program.

(This guideline, along with the next four, are adapted from Nanette Miner in an article published in T + D magazine titled "The Art of Test Creation.")

While not testing on material not taught may seem obvious, it happens far more often than you might imagine. For example, how many times have you delivered a training program and not covered all the material or only covered some of it in a cursory fashion because you ran short of time? The point here is that it's not fair to test participants on material that wasn't covered or only covered in a cursory way.

"Unclear test instructions can cause participants to view the test as unfair."

7. Write all test questions the same way the program material was taught.

Don't ask "null" or negatively worded test questions such as: "Which of the following is not one of the steps in the ADDIE model?" Null questions are often viewed as tricky and unfair. Also, why reinforce something you don't want participants to remember?

8. Provide clear test instructions.

If you're administering the knowledge test in person, have participants read through the instructions first to be sure they understand what to do. If the test is not going to be administered face-to-face, ask one or two colleagues to read through the instructions to be sure they are clear. Unclear test instructions can cause participants to view the test as unfair.

9. Allow participants to use job aids while taking the test, if they use job aids when performing their work.

Don't ask participants to take the test from memory if they don't have to recall the information from memory while performing their job.

10. Avoid creating test items that contain irrelevant information.

Irrelevant information is anything included in the test item that isn't needed to understand the question. For example, take the following question:

> *Chris is an internal L&D consultant. Her boss, Larry, the VP of HR, has asked her to design and deliver a new hire orientation program at four company locations across the U.S., Boston, Dayton, Omaha, and Oakland. Larry has also requested that the program not be longer than four hours. What approach should Chris use to design the training?*

The names Chris and Larry, Larry's title and the identification of the four company locations all constitute irrelevant information. A more concise version of the question might read something like this:

> *You are an internal L&D consultant and have been asked by your boss to design and deliver a four-hour new hire orientation training program. The schedule calls for conducting sessions at four company locations across the U. S. What approach would you use to design the training?*

> "Multiple-choice questions are… easy to grade and produce the most valid data, but they are challenging to write."

TEST ITEM CREATION TIPS

In addition to the general test creation guidelines described above, creating valid Level 2 knowledge tests also means paying attention to the following specific tips when creating multiple choice, true/false, matching and fill-in-the-blank test questions. (Many of these tips also are described in Nanette Miner's article.)

Multiple-choice questions are the most popular type of Level 2 test question. They are easy to grade and produce the most valid data, but they are challenging to write. They are hard to write because they must be void of any clues as to the correct answer, can only have one right answer and must strike a balance between being too easy and too difficult. Following is a list of common errors made when creating multiple-choice test questions and a tip on how to avoid each one.

COMMON ERRORS IN MULTIPLE CHOICE QUESTIONS	TIPS
A tendency for the correct answer to be the longest and to sound like a definition.	Solution: make sure all the response alternatives contain about the same number of words and sound similar.
The wording of the question reveals the correct answer. For example: *The ADDIE model is used primarily as an:* 1. *Instructional design tool* 2. *Measurement and evaluation tool* 3. *Change management tool* 4. *Process improvement tool*	Savvy test takers know that the correct answer to this question has to be "A" because it's the only response alternative that begins with a vowel and is grammatically correct with "an" at the end of the stem (the test item). Solution: if the correct answer begins with a vowel, end the question with a(n). Placing the "n" in parentheses enables any of the response choices to be correct.

"…the probability of guessing the correct answer, even without having mastered the program material taught increases from 25% to 50%, thus reducing question validity."

COMMON ERRORS IN MULTIPLE CHOICE QUESTIONS	TIPS
Some response alternatives are not viewed as plausible. For example: *Which of these communication techniques is used to consolidate a discussion and to move the focus of the conversation to another topic?* *A. Arguing* *B. Interrupting* *C. Summarizing* *D. Initiating*	Savvy test takers (and most other people for that matter) know that response choices A and B are incorrect. As a result, the probability of guessing the correct answer, even without having mastered the program material taught increases from 25% to 50%, thus reducing question validity. Solution: develop only plausible response alternatives. Some techniques for developing credible alternatives also courtesy of Matt Allen include: • Use common misunderstandings or confusions about the program content • Use other familiar, but incorrect, phrases or concepts • Use common errors made with the program content • Skip a step in a multi-step process.

"When "*All the above*" or "*None of the above*" appear as response options, they frequently are also the correct answer."

COMMON ERRORS IN MULTIPLE CHOICE QUESTIONS	TIPS
The use of "All the above" or "None of the above" as a response option. When "All the above" or "None of the above" appear as response options, they frequently are also the correct answer.	Savvy test takers know this, and when they see one of these response options used with a test question where they don't know the correct answer, they will select it because they know there is a high probability that it is the right answer. Solution: if you need to use "All the above" or "None of the above" as a response option, be sure to include some test items where it is an incorrect answer.

True/False test items the second most common type of question used in Level 2 evaluations. They are easy to write, but also tend to be the least valid -- learners have a 50/50 chance of guessing the correct answer. Following is a list of typical errors made when creating true/false questions and corresponding tips on how to overcome each one.

"When creating True/False questions, keep a balance between the number of each."

COMMON ERRORS IN TRUE/FALSE QUESTIONS	TIPS
The tendency to create more True questions than False questions. True questions are easier to write than False questions because of familiarity with the program content which is why many L&D professionals develop more True questions when creating Level 2 evaluations.	Savvy test takers know this and when they're not sure of the correct answer, choose True because they know it gives them a better chance of guessing the right answer. Solution: when creating True/False questions, keep a balance between the number of each.

"Savvy test takers know that while it's possible a statement might always or never be true; usually that is not the case."

COMMON ERRORS IN TRUE/FALSE QUESTIONS	TIPS
The development of test questions that are not entirely True or False. For example: *February has 28 days.* ☐ *True* ☐ *False* It's true that February has 28 days, but every four years (leap year) it has 29. Thus, learners could argue, regardless of whether they chose True or False, that their answer is correct. Since only one answer can be right, learners who selected the other answer are going to see the question as unfair or tricky.	Solution: only create questions that are entirely True or False.
The inclusion of words like "never" and "always" in the test question. For example: *Open-ended questions are always preferred to close-ended questions.* ☐ *True* ☐ *False* (Note: false is the correct answer.)	Savvy test takers know that while it's possible a statement might always or never be true; usually that is not the case. Therefore, when in doubt about the correct answer, savvy test takers will choose False because they know it gives them the best chance of guessing the right answer. Solution: avoid the inclusion of "absolute determiners" like always and never in your true/false test items.

> "Matching questions with more than 10 "A" column questions or "B" column answers are likely to be perceived as overwhelming."

Matching questions, the third type of Level 2 evaluation test question, are easy to create because only one "B" column correct answer is needed for each "A" column question. In contrast, multiple choice questions require at least three or four plausible answers. However, as with both multiple choice and true/false questions, a few common errors are often made when creating matching questions. A description of these and tips on how to prevent them follows:

COMMON ERRORS IN MATCHING QUESTIONS	TIPS
Creating a matching question that contains more than 10 "A" column questions or "B" column answers. According to Cognitive Learning Science research, humans possess the mental capacity to work with seven (plus or minus two) different pieces of information at the same time. Therefore, matching questions with more than 10 "A" column questions or "B" column answers are likely to be perceived as overwhelming.	Solution: keep the number of items in both the "A" and "B" columns to 10 or fewer. If there are more than ten items on a particular topic that you want to include in a matching question, break them up into chunks of 10 or fewer.

"Having too many extra questions or answers is likely to be perceived by learners as tricky and unfair."

COMMON ERRORS IN MATCHING QUESTIONS	TIPS
Including more than three extra "A" column questions than "B" column answers or vice versa. While having more questions than answers or more answers than questions increases the difficulty level of a matching question, it's important to limit the number of extra questions or answers so as not to overwhelm the learner. Having too many extra questions or answers is likely to be perceived by learners as tricky and unfair.	Solution: limit the number of extra questions or answers to no more than three. Limiting the number to three or fewer ensures that your matching questions will be fair to both the learner and the organization.

"(Fill in the blank questions…) test for recall and not understanding."

Fill-in-the-Blank questions, the fourth type of Level 2 evaluation test question, are used to test for learner recall of key facts and concepts. They are easy to create but require more time to grade than multiple choice, true/false or matching questions. Also, they test for recall and not understanding. (Note: See overall guideline number one above for the difference between testing for recall and testing for understanding.)

COMMON ERRORS IN FILL-IN-THE-BLANK QUESTIONS	TIPS
Creating fill-in-the-blank questions that ask learners to recall obscure facts and concepts. When this occurs, learners view the question as tricky and unfair.	Solution: when creating fill-in-the-blank questions, be sure the facts and concepts you're asking learners to recall are essential to know.

In summary, Level 2 evaluations often miss the mark because they are created by someone who is uninformed in the art and science of test item creation. As a result, the test items often either contain clues as to the correct answer or they are overly difficult or tricky. In either case, the result is an invalid Level 2 evaluation. However, by following the guidelines and tips described above, you'll be able to create valid Level 2 evaluations that measure what they purport to and are fair to both the learner and the organization, which after all is the real purpose of a Level 2 evaluation.

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