

Eight Tips on Developing Valid Level 1 Evaluation Forms*

Level 1 evaluations, for better or worse, are a ubiquitous part of the workplace learning and performance landscape. In fact, according to research conducted by ATD (*Evaluating Learning: Getting to Measurements That Matter*, 2015), 88% of organizations evaluate at least some learning programs at Level 1 of the four level evaluation model (Reaction, Learning, Behavior Change and Business Results). However, despite their widespread use, do you ever wonder about the validity of the results you obtain? You should. According to the same ATD research study, only 44% of organizations view the data they collect from their Level 1 evaluations to have either high or very high value.

Moreover, according to research conducted by Richard Clark and Fred Estes and published in a book titled *Turning Research into Results: A Guide to Selecting the Right Performance Solutions*, Level 1 evaluation results generally demonstrate a negative or inverse correlation with actual on-the-job behavior (Level 3). In other words, Level 1 evaluations often indicate the opposite of what actually happened in a learning program – either rating an effective program poorly or an ineffective program highly. This then raises a key question: “How do you design Level 1 evaluation forms that capture valid, high value data?” The following eight tips are offered as suggestions.

*Originally published in Training Today, Fall 2007 (A quarterly magazine published by the Chicagoland Chapter of ATD) and updated in 2019

“Asking questions about things you can’t do anything about to improve the effectiveness of a learning program wastes participant time...”

TIP 1 | Only ask questions that lead to ACTIONABLE DATA.

Asking questions about things you can’t do anything about to improve the effectiveness of a learning program wastes participant time and may eventually lead to participant frustration. For example, asking a question about how effectively a training room contributes to participant learning when it has a post in the middle that obstructs participant view, but is the only training room available, is a waste of time. If the room can’t be modified to eliminate the post or you’re not collecting data to build a business case for getting a new training room, stop asking the question. Answering the same question over and over and not seeing anything change leads to participant frustration and a lack of interest in completing the evaluation form.

TIP 2 | Write LEARNER-CENTERED evaluation items not trainer-centered.

This: *I found the room comfortable and conducive to learning.*

Not This: *The room lighting and temperature were conducive to learning.*

“Most Level 1 evaluation items are written from a ‘trainer-centered’ rather and a ‘learner-centered perspective.”

In a 2008 article titled “The New World Level 1 Reaction Sheets,” Jim Kirkpatrick points out that most Level 1 evaluation items are written from a ‘trainer-centered’ rather than a ‘learner-centered’ perspective (see the examples above). Jim’s point is that instead of asking participants for their thoughts about us and how well we clarified the learning objectives, organized the program material, kept the program moving, responded to their questions, etc., we should be asking participants questions about them and how they experienced the learning program relative to their own needs. He makes a good point. After all, we go to great lengths to make sure our learning programs are participant-centered so why shouldn’t we follow the same model when developing our Level 1 evaluation forms?

“Using qualitative questions and quantitative measures to assess the same dimension is an effective way to cross validate item results.”

TIP 3 | Where appropriate, MATCH UP QUALITATIVE questions WITH QUANTITATIVE measures.

Example:

In a word, how would you describe this session? _____

Using a number, how would you describe this session?

NO VALUE						GREAT VALUE
1	2	3	4	5	6	7

Using qualitative questions and quantitative measures to assess the same dimension is an effective way to cross validate item results. Specifically, asking participants to rate a learning program in a word and then to rate it using a number enables you to see if the word descriptions match the numeric value.

For example, if participants use words like “Outstanding”, “Excellent”, and “Great” to describe a learning program and then numerically rate it a 3.2 on a seven-point scale, something is wrong – either the participants misunderstood the survey item or the learning program wasn’t as good as the words suggest. On the other hand, if the numeric rating associated with these same words was 6.5, you could feel very confident that the learning program was a success because of the high positive correlation between the words and the numeric rating.

However, when using qualitative questions, keep the following caution in mind: Level 1 evaluations are typically administered at the conclusion of a learning program when participants have psychologically “checked out” and are physically ready to leave. Therefore, administering a Level 1 evaluation with lots of open-ended questions is a sure fire way either to get no response or thoughtless responses. The solution: either keep the number of qualitative questions to a minimum or limit the required response to a word or two such as in the example above.

“However, both these problems, as well as others created by word labels, can be eliminated by creating scales with only two word labels anchoring either end and a continuum of numbers in between.”

TIP 4 | When collecting quantitative data using a Likert scale, create a response scale with numbers at regularly spaced intervals and WORDS ONLY AT EACH END.

Example:

I felt engaged during the session because the facilitator kept the program moving.

NOT AT ALL TRUE							COMPLETELY TRUE	
1	2	3	4	5	6	7		

Many Level 1 evaluation forms use words to describe all the points along a scale. For example, in the scale above, words like “Barely True,” “Occasionally True,” “Mostly True,” and “Frequently True” might be used to describe points 2, 3, 4, 5, and 6. However, as Palmer Morrel-Samuels points out in a 2002 Harvard Business Review article titled “Getting the Truth into Workplace Surveys,” “The results from this type of evaluation [scale] are notoriously unreliable.”

He goes on to point out that because different words are used to describe each of the scale points, even though they may be in a plausible order, the distance between each pair of descriptors is not necessarily the same. For example, many people see the distance between “Not At All True” and “Barely True” (points 1 and 2) closer together than “Frequently True” and “Completely True” (points 6 and 7) are to each other. Because of this, the response choices are no longer spread across an evenly spaced mathematical continuum thus making it difficult to conduct informative statistical tests on the results obtained.

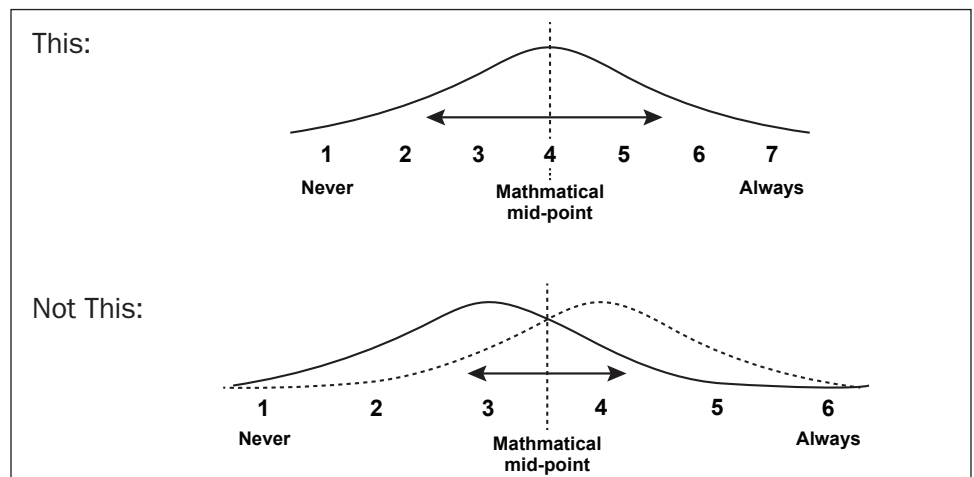
Another potential problem with labeling all the points on a scale identified by Morrel-Samuels is that often the descriptors overlap (“Occasionally True” and “Somewhat True”) and they may mean different things to different people thus making it difficult to compare results across groups.

However, both these problems, as well as others created by word labels, can be eliminated by creating scales with only two word labels anchoring either end and a continuum of numbers in between.

“Using an odd numbered scale with 7 to 11 response options is preferred over an even numbered scale of a similar length.”

TIP 5 | Use only one response scale with an ODD NUMBER of points (7, 9, and 11 point scales are best).

Again, according to Morrel-Samuels, single-scale evaluation forms, where the same two word labels are used to anchor either end of every evaluation item, are better than multiple-scale evaluation forms. Single scale evaluation forms take less time for participants to complete, provide more reliable data and make the comparison of results between different items easier. However, it may not always be possible to create a single scale Level 1 evaluation form and in these instances you should keep the number of different scales to a minimum as well as try to cluster the same scale items together.



Using an odd numbered scale with 7 to 11 response options is preferred over an even numbered scale of a similar length. Odd numbered scales allow participants the option of choosing a neutral response, which is a perfectly valid response. Odd numbered scales also more readily allow for the possibility of obtaining a normal bell shaped curve distribution of responses across the scale because it has an actual mid-point. Even numbered scales, on the other hand, increase the possibility of obtaining a skewed distribution of responses above or below the actual mathematical mid-point, such as in the example above, because participants aren't allowed to register a neutral response. The net result is that something that scored highly or poorly may not be as good or bad as the scores suggest.

“This runs counter to the way we normally count and can create problems when participants are in a hurry to complete the evaluation form.”

TIP 6 | Use SMALL NUMBERS AT THE LOW OR LEFT END of the scale and LARGER NUMBERS AT THE RIGHT OR HIGH END of the scale.

Example:

The learning activities used in this session helped me to achieve proficiency with the program material.

NOT AT ALL TRUE							COMPLETELY TRUE
1	2	3	4	5	6	7	

Sometimes you’ll see evaluation forms where the scale used runs in descending order or from high to low (e.g. 5, 4, 3, 2, 1) instead of low to high. However, this runs counter to the way we normally count and can create problems when participants are in a hurry to complete the evaluation form and mistakenly mark their responses at the right end of the scale thinking these are the better responses. The result is that good things about your learning event may come out looking bad and bad things may come out looking good.

Although not as common, another mistake occasionally made on evaluation forms is to create a scale where low numbers represent positive responses and high numbers represent negative responses (e.g. 1 = Completely True and 5 = Not at all True). Here again the scale is counter intuitive because we generally associate higher numbers with better and may create the same kind of problem described above where good things get rated low and bad things get rated high.

“While this may not have an adverse effect on item results, at a minimum it defies logic.”

TIP 7 | Write items either as a CONTINUUM or as a STATEMENT.

Examples:

This: *How effectively did the AV materials used during the session help reinforce your understanding of the program material?*

NOT EFFECTIVELY							VERY EFFECTIVELY
1	2	3	4	5	6	7	

Or This: *The AV materials used during the session helped reinforce my understanding of the program material.*

NOT AT ALL TRUE							COMPLETELY TRUE
1	2	3	4	5	6	7	

Not This: *Did the AV materials used during the session help reinforce your understanding of the program material?*

STRONGLY DISAGREE							STRONGLY AGREE
1	2	3	4	5	6	7	

Another mistake some people make when creating Level 1 evaluation items is to write an item that asks for a “yes/no” answer and then use a Likert scale for recording responses.

For example, the question “Did the AV materials used during the session help reinforce your understanding of the program material?” asks for a “yes” or “no” answer, not an answer that falls along a continuum such as the “This” and the “Or This” examples above. While this may not have an adverse effect on item results, at a minimum it defies logic.

“Rackham’s research suggests that not only does a relevance scale have a high positive correlation with learning (Level 2)...”

TIP 8 | Include at least one item asking participants how RELEVANT the learning event/material was TO THEM AND THEIR JOB.

Example:

How would you rate the overall relevance of this session to you and your job?

NOT AT ALL RELEVANT						COMPLETELY RELEVANT
1	2	3	4	5	6	7

Does this mean that every Level 1 evaluation form should include at least one question asking participants how relevant the training was? According to research conducted by Neil Rackham, author of SPIN Selling and Major Account Sales Strategy, and reported in Training magazine, the answer is a resounding “Yes!” In fact, Rackham’s research suggests that not only does a relevance scale have a high positive correlation with learning (Level 2), it also has a higher correlation with learning than an item evaluating participant learning.

In summary, Level 1 evaluations, while ubiquitous, often miss the mark because they are poorly designed and result in the capturing of misleading or invalid data. However, by following the tips described above you’ll be able to improve the validity and value of the data you collect and make better decisions regarding improvements needed in your learning programs, which after all is the real purpose of Level 1 evaluations.

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 Levels 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 than 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**