Phillips Associates
on M&E
5th in the series

Capturing Elusive Level 3 Data:
The Secrets of Survey Design

"…while only slightly more than half of all organizations evaluate some learning at Level 3, those that do place great value in the results."

Knowing whether or not participants apply back on the job what they learned in a training program is both a critical L&D and business issue. Demonstrating that learning is applied speaks directly to our ability as L&D professionals to be viewed by the business as a credible partner. This point is made clear in a 2009 research study by the ROI Institute where they found that employee application of what was learned in a training program back on the job is one of the top three metrics senior business executives are most interested in seeing.

Learning that is delivered but not applied on the job, i.e., "scrap learning," wastes time and money, both scarce organization resources. Unfortunately, measuring on-the-job behavior change is something few L&D professionals have much experience with nor know much about.

According to a 2015 ATD research study titled, "Evaluating Learning Getting to Measurements That Matter," only 60% of organizations evaluate some learning programs at Level 3: Behavior. Data regarding the percent of programs assessed at Level 3 is, even more, telling: only 33% of live classroom programs and 18% of technology-based programs. However, when asked about the value Level 3 evaluation data has for their organization, 75% of study respondents indicated it had either high or very high value. In short, while only slightly more than half of all organizations evaluate some learning at Level 3, those that do place great value in the results.

Also, according to the study, the most common method used by organizations to collect Level 3 evaluation data is to administer a survey. I now will address twelve specific tips for creating valid, scientifically sound Level 3 surveys. These tips and the issues they address fall into three

categories: content, format, and measurement. The tips also are based on recent education and behavioral science research and call into question many survey design principles formulated 60 or more years ago, but still in use today. These tips apply equally to surveys created for completion by program participants themselves, their managers, colleagues or direct reports*

"Responding to an item measuring thoughts or motives and not behavior requires speculation…"

TIPS ON **CONTENT**

TIP 1 | Focus Survey Items on OBSERVABLE BEHAVIORS Not Thoughts or Motives.

Often survey items that measure thoughts or motives produce invalid results. For example, if you're creating a survey that is going to be completed by others about a learner's behavior (e.g., employees completing a questionnaire about their supervisor who attended a training program), it's impossible for employees to know with certainty what's going on inside their supervisor's head. As a result, responding to an item measuring thoughts or motives and not behavior requires speculation by the employees as to what their supervisor was thinking or what motives the supervisor had in mind. Similarly, if you're creating a survey that's going to be completed by learners about their behavior, they are a lot less likely to recall the thoughts or motives behind their behavior than the action itself.

For example:

This: *When giving constructive feedback, my manager does it in private.*

Not This: *When giving constructive feedback, my manager considers whether it should be done privately or in the presence of others.*

TIP 2 | **Limit each item to a SINGLE DESCRIPTION of behavior.**

Combining two distinct behaviors into a single survey item, referred to as a "double-barreled" item, results in "muddied" data that is impossible to interpret accurately. Whether a score is high, low or neutral, it is impossible to draw an accurate conclusion from the result because there is no way to know whether the survey respondent considered both behaviors or one or the other when responding to the item.

For example:

This: *My manager provides feedback just as soon as possible after an event has happened.*

Not This: *My manager provides feedback just as soon as possible after an event has happened and avoids getting emotional or evaluative.*

"In surveys longer than just a few items, there is a tendency for respondents to start 'just checking boxes' without thoroughly reading each question."

TIP 3 | **Write about 1/3 of the survey items so that the desired answer is negative.**

While this tip may seem counterintuitive, it addresses two common response biases that often invalidate survey data. First, there is considerable evidence documenting the tendency of survey respondents to select response options on the positive side of the scale mid-point when evaluating individual behavior.

Second, in surveys longer than just a few items, there is a tendency for respondents to start "just checking boxes" without thoroughly reading each question. Negatively worded survey items help to mitigate both of these biases. However, to obtain maximum benefit from this technique, it is recommended that you point out in the survey instructions that negatively worded questions are in the survey and then have a negatively worded item appear as the second, third or fourth item in the questionnaire.

TIPS ON **FORMAT**

"Respondents tend to rate all the items in a section the same way they evaluate the first item."

TIP 4 | Keep the sections of the survey unlabeled.

The reason for this tip is straightforward: respondents tend to rate all the items in a section the same way they evaluate the first item. For example, if a respondent rates the first item in a section highly, he or she tends to assess all the items in that section highly.

To prevent this, do two things:

1) remove all section or topic headings from the survey and

2) randomly distribute the items from each section throughout the questionnaire.

Of course, to interpret the survey results you'll need to reassemble the questions back into their respective topic areas, but if the survey is electronic, this is relatively easy.

"To avoid business executive credibility questions and to ensure the validity of survey results, create participant surveys that contain approximately the same number of items in each section."

TIP 5 | Design the survey sections to contain a similar number of items and the items to include a similar number of words.

This tip addresses two common survey issues: the credibility and the validity of the survey results. If you create a survey that has significantly more items in some sections compared with others, you put yourself at risk that a business executive will call into question the credibility of the results in those sections with vastly fewer items. Also, research evidence shows that those sections containing significantly more items tend to have higher average scores than those sections with fewer items, which calls into question the validity of the results. This same phenomenon also occurs with survey items that contain vastly more words than others – the average score on these items tends to be higher. Therefore, to avoid business executive credibility questions and to ensure the validity of survey results, create participant surveys that contain approximately the same number of items in each section and the items to include a similar number of words.

"Moving demographic questions to the end of a survey improves response rates by about 8 percent."

TIP 6 | Place questions regarding participant demographics (e.g., name, title, department, and so forth) at the end of the survey, make completion optional and keep the number of questions to a minimum.

Placing demographic questions at the beginning of a survey has the potential dual effect of biasing responses towards the favorable and depressing response rates.

If respondents think their ratings are traceable back to them, it increases the probability that they will provide more positive responses so they won't be asked to explain their scores. Respondents who possess low levels of organizational trust, or who are not sure how the data will be summarized and used, or who have had a negative experience answering a previous survey are the most susceptible to this.

Similarly, placing demographic questions at the beginning of a participant survey can potentially depress response rates. For example, some respondents may decline to fill out the questionnaire if they think there is a chance their responses will not be anonymous. In short, research indicates that moving demographic questions to the end of a survey improves response rates by about 8 percent.

TIPS ON **MEASUREMENT**

TIP 7 | Collect data from multiple observers or a single observer multiple times.

The basis for this tip is that timing is critical when conducting Level 3 evaluations. Two things to consider are:

1) how long following a learning program will it be before participants can apply what they learned; and

2) how soon are others (boss, colleagues, direct reports) likely to recognize participant on-the-job behavior change?

While it's nearly impossible to know the precise answer to either of these questions, to give yourself the best chance of detecting behavior change if it occurs you should either collect data from multiple observers or collect data from a single observer multiple times. Using numerous observers improves your chances of detecting behavior change because if one observer didn't

see or recognize any change, perhaps one of the other observers did. This same logic would apply to collecting data from a single observer multiple times – if the observer didn't see any participant behavior change the first time perhaps he or she will the second time.

"While the word descriptors may appear in a plausible order, often the distance between each pair of descriptors is not the same."

TIP 8 | Create a response scale with numbers at regularly spaced intervals and words only at each end.

Many participant surveys use words to describe all the points along a response scale. For example, words like "Not at all True," "Barely True," "Occasionally True," "Somewhat True," "Mostly True," "Frequently True," and "Completely True" might be used to describe points 1–7 on a scale. However, research indicates that the results from this type of measurement scale are notoriously unreliable. While the word descriptors may appear in a plausible order, often the distance between each pair of descriptors is not 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). Because of this, the response choices are not spread across an evenly spaced mathematical continuum thus making it difficult to conduct informative statistical tests on the data collected. Another potential problem with labeling all the points on a scale is that often the descriptors overlap ("Occasionally True" and "Somewhat True") and the descriptors may mean different things to different people making it difficult to compare results across groups. However, these problems, as well as others created by word labels, can be eliminated by creating a scale with word descriptors only at each end and a continuum of numbers in between.

This:

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

Not This:

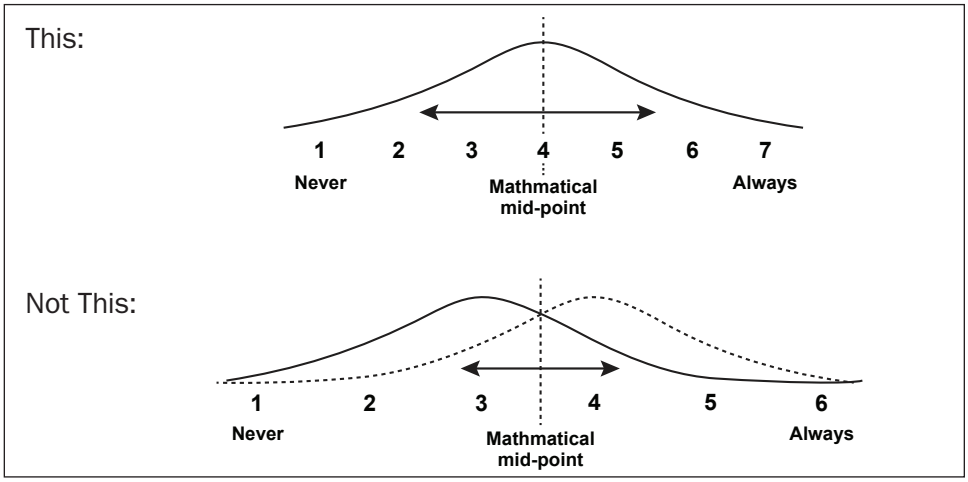
NOT AT ALL TRUE	RARELY TRUE	OCCASIONALLY TRUE	SOMEWHAT TRUE	MOSTLY TRUE	FREQUENTLY TURE	COMPLETELY TRUE
1	2	3	4	5	6	7

"Odd numbered scales allow participants the option of choosing a neutral or mid-point response, a perfectly valid answer."

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

Single-Scale surveys, where the same word descriptors appear with every survey item are better than questionnaires using multiple word descriptor scales because they take less time to complete, provide more reliable data and make the comparison of results between different survey items easier.

Using an odd-numbered scale with 7 to 11 response options also is preferred over an even-numbered scale of a similar length. Odd numbered scales allow participants the option of choosing a neutral or mid-point response, a perfectly valid answer. Odd-numbered scales also readily allow for the possibility of obtaining a normal bell-shaped curve distribution of responses because they have an actual mid-point. In contrast, even numbered scales increase the likelihood of getting a skewed distribution of responses above or below the mathematical mid-point (see example below) because participants aren't allowed to register a neutral reply. The result is something that scored poorly or highly may not be as bad or good as the scores suggest.



"Respondents can quite accurately recall whether a behavior happened frequently or not at all even if they weren't consciously keeping track of how often it occurred."

TIP 10 | Use a response scale that measures frequency, not agreement or effectiveness.

Research has demonstrated that a frequency scale provides more accurate and reliable data than either an agreement or effectiveness scale. Respondents can quite accurately recall whether a behavior happened frequently or not at all even if they weren't consciously keeping track of how often it occurred. In contrast, agreement scales often produce biased results where the majority of responses end up clustered at the high end of the scale. Effectiveness scales, on the other hand, often produce biased data because unless all the raters have gone thru calibration, what is considered effective by one respondent may be viewed as ineffective by another.

TIP 11 | Place small numbers at the left or low end of the scale and large numbers at the right or high end of the scale.

Occasionally you'll see surveys where the scale used runs in descending order or from high to low (e.g., 7, 6, 5, 4, 3, 2, 1) instead of low to high. High to low scales are contrary to how we usually count and can create problems when respondents are in a hurry to complete the survey and mistakenly mark their responses at the right end of the scale thinking these are the more positive responses. The result is that while behavior may have changed, the data suggest otherwise.

This:

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

Not This:

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

"We need to account for the fact that the respondent completing the questionnaire may not have been in a situation to observe any change in the behavior."

Although not as common, another mistake occasionally made on participant surveys is to create a scale where low numbers represent positive responses and high numbers represent negative responses (e.g., 1 = Completely True and 7 = 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 behavior changed, but the data indicate otherwise.

TIP 12 | Include a "Did Not Observe" response choice and make it different from the rest of the scale.

Because Level 3 surveys focus on measuring on-the-job behavior, we need to account for the fact that the respondent completing the questionnaire may not have been in a situation to observe any change in the behavior described by the item. Including a "Did Not Observe" response option, allows respondents to avoid having to choose one of the other scale response choices or to leave the item blank when they weren't in a position to observe any behavior change.

Capturing Level 3 on-the-job behavior change data is becoming increasingly important as business executives seek evidence that the employees they send to training are applying back on the job what they learned. Moreover, while it's essential to meet these expectations, it is equally important to be sure we provide evidence that is unbiased, credible and useful. Following these tips will enable you to create Level 3 surveys that give business executives the on-the-job behavior evidence they seek and gain recognition for the value you provide.

* Some of the following tips are adapted from "Getting the Truth into Workplace Surveys," Palmer Morrel-Samuels, *Harvard Business Review*, February 2002, pp. 111-118.

202006023

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