Improving Practitioner Training Evaluation Techniques

Brandi Goff

Independent Study

University of Maryland, Baltimore County

Winter 2020

Improving Practitioner Training Evaluation Techniques

In the field of education and training, evaluations are used to assess the effectiveness of learning interventions. Unfortunately, evaluations are oftentimes less than effective in determining knowledge retention and behavior change. The most prominent evaluation types include reaction surveys or smile-sheets, and pre-/post-test evaluations. While these tools serve a purpose, they provide little insight as to whether the student internalized what was being taught. In order to more effectively measure internalization, methods for measuring behavior change must be improved.

Trainings can teach a variety of subject matters and skills, and the evaluations tools and techniques used to measure each training's success must cater towards the unique needs of each. This paper will focus specifically on practitioner training and education (hereafter referred to as "practitioner training") having to do with knowledge acquisition, rather than soft or hard skill training or compliance training. Knowledge acquisition training, for the purposes of this paper, refers to educational experiences that improve understanding and knowledge of specific subject areas.

Kirkpatrick identifies four levels of evaluation: reaction, learning, behavior, and results, with a fifth level commonly agreed upon—return on investment. Most relevant to internalization is the level three behavior evaluation, providing insight post-training into how an individual's behavior has changed as a result of the training. For job-training, evaluators look to see if participants are internalizing what they learned in a training and applying it to their job tasks. However, in what this paper will refer to as "traditional behavior change evaluations," this typically requires for there to be follow-up surveys at some point after the training, access to supervisor reviews, or interviews for such things as job-task analyses—all of which can be difficult to obtain and validate data. This becomes more difficult for firms and evaluators that are contracted by other organizations to conduct their knowledge acquisition trainings, e.g. a government agency contracts a training and research center to provide subject-matter specific training to their employees. In these scenarios, oftentimes once the training is completed, the evaluators lose access to participants, do not have direct access to supervisors, and, outside of mandates, there is little incentive for participants to complete surveys after they have already completed the training and received certification. Simply put, there is a need for an evaluation tool that can speak to internalization while overcoming the obstacles that traditional behavior change evaluations face.

Innovative Techniques for Evaluations

Before dismantling evaluations in their entirety, it is helpful to first take a step back and assess the number of tools and techniques that already exist. The ultimate purpose is to create a behavior evaluation tool that produces usable and valid results. The goal here is not to reinvent the wheel if it is not necessary. Techniques utilized by instructional designers as well as social science researchers provide a starting point for more effective evaluation techniques.

Retrospective Assessments

First coming to prominence in the 1970s, retrospective assessments or post-then-pre evaluations utilize self-reporting measures after a learning intervention has taken place in order to reduce

response-shift bias. This assessment is "shown to reduce response-shift bias providing more accurate assessments of actual effect, is convenient to implement, provides comparison data in the absence of 'pre' data, and may be more appropriate in given situations" (Klatt and Taylor-Powell, 2005, p. 1). Response-shift bias is introduced when a participant has different internal standards between ratings (Klatt and Taylor-Powell, 2005). For instance, a participant receiving incorrect information prior to a training on what will be covered will gain a different understanding of the assessment question after the training, producing a response-shift and invalidating the data.

In a standard retrospective assessment, participants are first asked about their behavior as a result of the learning intervention. They are then asked to report on their behavior prior to the intervention. "This second question is really the pre[-assessment] question, but it's asked after the program when the participant has sufficient knowledge to answer the question validly" (Rockwell and Kohn, 1989, Correcting Problem, para. 1) and allows the participant to draw from the same internal standard (Sprangers and Hoogstraten, 1989). Program outcomes measurable by retrospective assessments "include changes in knowledge, skills, abilities, motivations, self-efficacy, and behaviors" (Klatt and Taylor-Powell, 2005, p.6). It is important to note here that as this assessment type is asking about behavior, i.e. what a participant did before or after a training, it falls closer in line with Kirkpatrick's level three evaluation, rather than level two knowledge evaluations which are seen as synonymous with the terms "pre-test" and "post-test" as often used in retrospective evaluation research. Researchers "argue that self-assessments are more appropriate for measuring self-efficacy … than as a measure of knowledge or learning" Klatt and Taylor-Powell, 2005, p. 6).

The format of the assessment can vary. It may be "as a single questionnaire or as two separate questionnaires; as a single question within a series of questions or as the total questionnaire; in vertical or horizontal layouts and various page designs" (Klatt and Taylor-Powell, 2005, p. 8). The order of questions may vary; however, it is commonly agreed upon for post-training questions to be asked first, immediately followed by the pre-training question. Ultimately, the format used may depend upon the participant characteristics and the skills of the designer developing the tool (Klatt and Taylor-Powell, 2005).

Research demonstrates that retrospective assessments are more effective than traditional pre- and post-assessments and can reduce or eliminate response-shift (Howard, Ralph, et al., 1979; Lam, 2003; Sprangers and Hoogstraten, 1989). "In most instances, greater program effects are found with [retrospective assessments] than with the traditional pre-post as participants tend to overestimate their preprogram performance on the pre[-assessment], thus displaying little or negative effect at post-[assessment]" (Klatt and Taylor Powell, 2005, p.3). However, there are critics that argue that retrospective assessments do not eliminate bias. Griner, Hill, and Betz (2005) point out that not unlike traditional pre- and post-assessments, there are validity concerns that arise with retrospective assessments, such as implicit theories of change and effort justification bias. What matters is controlling for the biases that may arise, as validity concerns can arise regardless of what form of assessment is used.

Proponents of retrospective assessments argue that the advantages to implementation outweigh the disadvantages. As it relates specifically to behavior change, Pohl (1982) found that there was

no statistically significant difference between objective behavioral change and self-reported change. Pohl (1982) also found a "significantly higher correlation between the retrospective rating and the objective pretest than between the pre rating and objective pretest, indicating that the retrospective rating is a more accurate estimate of actual pretest performance" (Klatt and Taylor-Powell, 2005, p. 4).

Survey Methodology

A number of survey methodologies, including feeling thermometer scores and survey experimental designs, which are commonly used in social science research, can be drawn on for improving evaluation tools.

Feeling Thermometer

Feeling thermometer scores allow respondents to use an imaginary scale to assign a numeric number to their feelings about a person, group, or issue (Nelson, 2008). Attitude scores correlate to temperatures, with 0 being very cold and 100 being very warm. Lower scores indicate negative attitudes towards the person, group, or issue, while higher scores indicate positive attitudes, and middle range scores indicating neutral attitudes. Research shows that providing respondents a greater range of categories from which to choose results in more reliable and valid data (Alwin, 1997). Thus, feeling thermometers "allow researchers to gather information about the direction, as well as the intensity of respondents' attitudes and feelings" (Nelson, 2008, p. 276).

The subjectivity of this evaluation technique lends to high levels of variation in responses (Nelson, 2008), which should not be overlooked. Research has found that some respondents have different tendencies when applying the scale, with some tending to be warmer or colder than others (Nelson 2008). Some respondents also "restrict their ratings to relatively small portions of the thermometer, whereas others are just more open to using the entire spectrum" (Nelson, 2008, p. 276). For this reason, this survey technique may be best applied to observing how subjective responses change over time.

List Experiment/Item count

List experiments or the item count technique provides a mechanism for evaluators to measure sensitive topics without requiring individuals to directly answer sensitive questions (Blair, Imai, and Lyall, 2014; Lavrakas, 2008). This technique can also be used "to estimate the proportion of people who have engaged in stigmatizing behavior" (Tsuchiya, Hirai, and Ono, 2007, p. 253). For the purposes of evaluating behavior, this method may be used to measure undesirable traits that may be impacted by individual biases or matters of privacy concerns.

With this technique, respondents are randomly assigned to either a control or a treatment group. Without knowing that the lists they are responding to are different, each group is asked to consider how many of the statements apply to them. Individuals are not asked to indicate which statements apply to them, but only a total item count number. The treatment group's list would include an additional statement about the sensitive behavior or issues that evaluators are attempting to measure. Evaluators can then compare the average answers in order to determine how many individuals the sensitive statement may apply to (Lavrakas, 2008).

One limitation of this experiment type is the potential for respondents in the treatment group to unintentionally reveal their responses by choosing either all items or none (Blair, Imai, and Lyall, 2008). Individual-level analysis is not possible with this technique, and analysis must remain at the group-level. However, that can still provide insight into sensitive areas that may otherwise be impacted by biases.

Integrating Techniques into Evaluation Tools

In the absence of the ability to conduct traditional behavior change evaluations, evaluators can instead look to self-reported behavior change and intent. The following adjustments can either be incorporated into a joint knowledge-level post-test and post-assessment survey, or a post-assessment survey can be disseminated separately from the knowledge-level post-test. Example post-assessment survey questions have been made available in *Appendix A.*

Behavioral Assessment

As demonstrated by Pohl (1982), self-reported behavioral change is not statistically significantly different from objective behavioral change. Retrospective assessments can therefore be used to obtain self-reported behavior change with confidence. Retrospective behavioral assessments should ask for participants to draw on the behavior at the conclusion of the training as well as prior to when the training began. Clarifying time periods is particularly important when asking retrospective questions (Klatt and Taylor-Powell, 2005).

List experiments can also be utilized to increase a participant's willingness to claim undesirable behaviors they possess or previously possessed.

In addition to asking questions regarding immediate behavior change following the training, participants can also be asked questions about behavioral intent. When asked in line with the retrospective behavior questions, this would provide a future point for participants to consider when responding about their prior, current, and future behavior. Should a traditional behavior change assessment be possible, these data can supplement those results.

Each of these mechanisms can also be used as a secondary data-point should validity of traditional behavior change evaluations come into question. For example, self-reported data may be used in place of traditional behavior change evaluations should the traditional method receive a low response rate, thus injecting bias and validity concerns into the results.

Satisfaction/Reaction Surveys

Satisfaction or reaction surveys can continue to provide insight into what a participant liked or disliked in a course. While participants are typically asked to respond using what is referred to as "smile sheets," the feeling thermometer technique could provide greater insight into a participant's feelings towards the training program, as it provides more response flexibility for the respondent. The greater range that feeling thermometer scores allow may also reduce the inclination of students to check all of the highest (or lowest) rating when completing the survey.

Questions pertaining to behavior assessment can be interspersed within the satisfaction survey in order to create one cohesive post-assessment evaluation.

Learning Assessment

Preview pre-assessments, commonly referred to as pre-tests, as well as traditional post-tests should continue to be implemented to measure knowledge growth. While knowledge growth does not necessitate behavioral change, participants are more likely to engage if they perceive that they are being "tested" on what they have learned. This can ultimately support behavior change.

Discussion

After reviewing literature and survey methodology, there lies a path forward for improving evaluation techniques related to behavior change. In an ideal world, traditional behavior change assessments can be implemented. However, this is not always possible nor practical. Instead, evaluators can implement assessment strategies including retrospective assessments and list experiments to determine immediate behavior change and future intent. These assessments can be used either on their own or in conjunction with traditional behavior change assessments in order to improve evaluation data collection.

This paper acknowledges a likely correlation between learning assessment implementation and behavior change. Further research should be conducted to confirm that perception and implementation of learning assessments supports positive behavior change at the conclusion of a training. This should specifically be explored with regards to self-reported behavior change and intent as proposed by this paper.

Appendix A

Note: The following evaluation example questions includes both reaction and behavior assessment questions. Learning assessment is not included.

The example does not represent a definitive list of evaluation questions and should be modified for the purpose and needs of each training. The "Evaluation Level" column should be removed prior to dissemination to participants.

Question	***Possible Answers/Responses***	***Evaluation Type***
What did you expect to achieve from this training?	*Open-ended*	*Reaction*
Considering your answer above, did the training meet your expectations?	*A) The training did not meet my expectations* *B) The training met my expectations* *C) The training exceeded my expectations*	*Reaction*
Please consider the level to which you agree with the following statement: *The training was easy to follow.*	*(Feeling Thermometer) Using a scale of 0-100, with 0 being strongly disagree, and 100 being strongly agree, please indicate your answer.*	*Reaction*
Please consider the level to which you agree with the following statement: *Class discussion was managed effectively.*	*(Feeling Thermometer) Using a scale of 0-100, with 0 being strongly disagree, and 100 being strongly agree, please indicate your answer.*	*Reaction*
Please consider the level to which you agree with the following statement: *Overall, the instructors were knowledgeable about the subject matter.*	*(Feeling Thermometer) Using a scale of 0-100, with 0 being strongly disagree, and 100 being strongly agree, please indicate your answer.*	*Reaction*
Please consider the level to which you agree with the following statement: *Overall, the instructors were well prepared.*	*(Feeling Thermometer) Using a scale of 0-100, with 0 being strongly disagree, and 100 being strongly agree, please indicate your answer.*	*Reaction*
Please select the most accurate response. ***After*** *completion of the training, I am able to [Behavior/skill/ability dictated by terminal objective(s) of training].*	*1) Almost never* *2) Seldom* *3) Often* *4) Always*	*Behavior*
Please select the most accurate response.	*1) Almost never* *2) Seldom*	*Behavior*

***Prior** to completing the training, I was able to [Behavior/skill/ability dictated by terminal objective(s) of training].*	*3) Often* *4) Always*	
Please select the most accurate response. ***In the future**, I intend to [Behavior/skill/ability dictated by terminal objective(s) of training].*	*1) Almost never* *2) Seldom* *3) Often* *4) Always*	*Behavior (intent)*
*You will now read a list with different statements on it. After you read the entire list, please record how many of these items you identify with **after** completing the training. Please do not record which items you identify with; only record the number of items.*	(*List Experiment – must be adjusted to fit the behaviors being measured by the training. A control group will only see options A-C).* A) I know how to conduct interviews during my assignment. B) I know where to look for background material related to my assignment. C) I feel the resources I may need are accessible to me. D) I am not fully prepared to deploy on my assignment. *[Experimental]*	*Behavior*
*You will now read a list with different statements on it. After you read the entire list, please record how many of these items you identified with **prior** to completing the training. Please do not record which items you identify with; only record the number of items.*	(*List Experiment – must be adjusted to fit the behaviors being measured by the training)* A) I know how to conduct interviews during my assignment. B) I know where to look for background material related to my assignment. C) I feel the resources I may need are accessible to me. D) I am not fully prepared to deploy on my assignment. *[Experimental]*	*Behavior*

Bibliography

Alwin, D. F. (1997). Feeling thermometers versus 7point scales: Which are better? *Sociological Methods & Research, 25*(3), 318-340. DOI: 10.1177/0049124197025003003

Blair, G., Imai, K., and Lyall, J. (2014). Comparing and combining list and endorsement experiments: Evidence from Afghanistan. *American Journal of Political Science, 58*(4), 1043-1063. DOI: 10.1111/ajps.12086

Griner Hill, L. and Betz, D. L. (2005). Revisiting the retrospective pretest. *American Journal of Evaluation, 26*(4), 501-517. DOI: 10.1177/1098214005281356

Howard, G. S., Ralph, K. M, Gulanick, N. A., Maxwell, S. E., Nance, D. W., and Gerber, S. K. (1979). Internal invalidity in pretest-posttest self-report evaluations and a re-evaluation of retrospective pretests. *Applied Psychological Measurement, 3*(1), 1–23. DOI: 10.1177/014662167900300101

Klatt, J. and Taylor-Powell, E. (2005). Synthesis of literature relative to the retrospective pretest design. Panel presentation for 2005 Joint CES/AEA Conference, Toronto, October 29, 2005. Retrieved from http://comm.eval.org/HigherLogic/System/DownloadDocumentFile.ashx?DocumentFileKey=31536e2f-4d71-4904-ae5d-056e3280c767

Lam, T. (2003). A comparison of three retrospective self-reporting methods of measuring change in instructional practice. *American Journal of Evaluation, 24*(1), 65-80. DOI: 10.1177/109821400302400106.

Lavrakas, P. J. (2008). List-experiment technique In P. J. Lavrakas (Ed.), *Encyclopedia of Survey Research Methods* (pp. 432-434). Thousand Oaks, CA: Sage Publications, Inc.

Nelson, S. C. (2008). Feeling thermometer In P. J. Lavrakas (Ed.), *Encyclopedia of Survey Research Methods* (p. 276). Thousand Oaks, CA: Sage Publications, Inc.

Pohl, N. F. (1982). Using retrospective pre-ratings to counteract response-shift confounding. *Journal of Experimental Education*, *50*(4), 211-214.

Rockwell, S. K. and Kohn, H. (1989). Post-then-pre evaluation. *Journal of Extension, 27*(2). Retrieved from https://www.joe.org/joe/1989summer/a5.php

Sprangers, M. and Hoogstraten, J. (1989). Pretesting effects in retrospective pretest-posttest designs. *Journal of Applied Psychology, 74*(2), 265-272. DOI: 10.1037/0021-9010.74.2.265

Tsuchiya, T., Hirai, Y., and Ono, S. (2007). A study of the properties of the item count technique. *The Public Opinion Quarterly, 71*(2), 253-272. DOI: 10.1093/poq/nfm012