

Taking Measure: Choose the right data for the job

A good evaluation measures the same thing from different angles

Robby Champion

Journal of Staff Development, Summer 2002 (Vol. 23, No. 3)

Copyright, National Staff Development Council, 2002. All rights reserved.

There are no Swiss army knives when it comes to data: No single data source or instrument can measure everything. We must customize each evaluation so it fits the program and choose the data we will collect very strategically.

A good evaluation includes several kinds of data that measure the same thing from different angles. For example, to track student achievement throughout an initiative, you might collect students' standardized test scores, student grades, and student work on major school projects. In analyzing the data, evaluators need to explore agreements and contradictions in the data gathered from different sources. Cross-checking various data measuring the same construct is known as triangulation.

There is no shortage of program evaluation models. Select a model to guide your work that fits best with your evaluation philosophy. Donald Kirkpatrick's four-level model, created more than 40 years ago, is the best known in the training field. It is simple, user-friendly, and many have used and expanded on it. The model advocates evaluating the major points of impact: (1) reaction, (2) learning, (3) use, and (4) results. The following adapts Kirkpatrick's model to education.

R
L
C
R

LEVEL 1: Data to evaluate reaction/awareness

Evaluating participants' reactions to a professional development program should be straightforward and not drain your evaluation resources.

You may want to measure participants' reactions to different aspects of the experience, such as content, process, and context. Data might be analyzed to distinguish one type of participant from another. For example, you may collect data on how beginning teachers react compared with veteran teachers.

Your guiding question: What was the typical (or average) participant reaction to the program?

Data to evaluate reaction are:

- Results of pencil-paper feedback surveys distributed as participants leave a seminar.
- Record of the participants who voluntarily enroll in the next level program.
- Record of the number of participants who voluntarily stay through the program.
- Record of participants' questions or comments.
- Participants' responses to questions on a program's content.
- Scores on pre- and post-program quizzes on the key content points.
- Requests for materials after participants complete the program.
- Observations of participants' behavior during program.
- Results of follow-up inquiries via telephone, e-mail, pencil-paper survey, focus group, or interview.
- Record of participants' follow-through in using a web site introduced in the program.

LEVEL 2: Data to evaluate trainee learning

Without data, we may wrongly assume participation equals learning. Evaluating learning involves gathering data to determine whether participants know how to apply the new learning to their work. Look for data that can do more than one job. For example, if teachers

lect students' work throughout a professional development program, those work samples may document teachers' learning (Level 2) and whether teachers are using what they are learning (Level 3).

Track individual growth in programs such as leadership academies where participants attain a set of competencies. Pre- and post-assessments address questions about what participants knew about the content before entering the program.

Your guiding question: To what extent has each participant learned what this program intended?

Data to evaluate participant learning are:

- Scores on pre- and post-assessments given to each participant.
- Quality of learning products/artifacts which require participants to apply what was learned during program.
- Demonstration of understandings/skills on performance tasks throughout the program (such as role play exercises or presentations).
- Workplace observations to determine participants' progress in applying the new learning to the classroom or workplace.
- Reports from principals and supervisors on participants' application of the new learning in the classroom/workplace (from interviews, surveys, focus groups).
- Classroom application assignments completed by individuals or teams being trained, such as unit plans or case studies of student progress (evaluated with a common rubric, checklist, standards, or other criteria for quality).
- Content analysis of log/journal of each participant's growth in understanding and ability to apply program content.
- Performance of individuals or teams on training exercises, such as a "Three-Minute Paper," or games,

such as homemade versions of game shows (Jeopardy, Wheel of Fortune, etc.).

- Responses by participants to questions on program's content.

LEVEL 3: Data to evaluate full implementation/use of new learning
Evaluating how much learning from a professional development program has been transferred to the workplace is challenging and time-consuming. Ideally, data include observing teacher trainees, but the expense and obtrusiveness usually lead evaluators to seek other evidence.

An evaluation should also determine whether participants tailor what they learned to fit their context and whether they use the new knowledge regularly. Training homework such as classroom application projects can provide data to measure how participants adapt their new learning to their work setting.

Your guiding question: To what extent are participants using what they learned in this program?

Data that might be useful in evaluating participant implementation or use of what they learned are:

- Regular workplace observations to track participants' use of the new learning.
- Reports from principals and supervisors on participants' correct application of the new learning in the classroom/workplace (written reports, interviews, surveys).
- Content analysis of participants' logs/journals of working through implementation problems.
- Participants' self-reported descriptions of their implementation experiences in support groups.
- Student work (samples, portfolios, or individual case studies).
- Reports from students (interviews, focus groups, surveys).
- Reports from parents (interviews, focus groups, surveys).

- Classroom application projects participants undertake to apply what they learned (evaluated using a common rubric or rating scale for all trainees to determine quality).

LEVEL 4: Data to evaluate results

Specify at the start what you'll use to indicate the impact of a major initiative. Evaluating final results may be costly unless you use some data already collected. For example, if you use changes in the quality of student portfolios to indicate use of new learning, the portfolios can be used to demonstrate impact. Also plan to collect baseline data to compare with end results, too.

The professional learning experience is only one factor in any improvement. Changes in leadership, curriculum adoptions, schedule alterations, and staff changes also affect results. Collect data on those factors and determine the relative influence of each on the final results.

Your guiding question: Did the changes in staff behavior after the training impact positively on the organization, including improved results, such as higher student achievement?

Data that might be useful in documenting results are:

- School records of student progress (homework, grades, attendance, detentions, enrollment in particular programs).
- Student test scores (normed or criterion-referenced tests).
- Student work (samples, portfolios, or individual case studies).
- Reports from students (interviews, focus groups, surveys).
- Reports from parents (interviews, focus groups, surveys).
- Artifacts, such as minutes of team or faculty meetings (to document changes in school culture).

- Records of the percentage of new teachers successfully completing two years teaching (in a beginning teachers' induction program).

If we need a refrain to help us gauge which data to collect to evaluate professional development in a results-based world, it should be: "All professional development programs need to be evaluated for their impact, but not necessarily all at the same time, nor all with the same intensity, and not necessarily with the same kind of data."

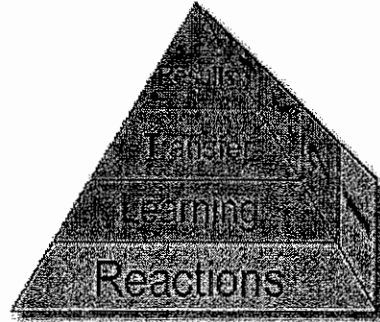
About the Author

Robby Champion is president of Champion Training & Consulting. You can contact her at Champion Ranch at Trumbell Canyon, Mora, NM 87732, (505) 387-2016, fax (505) 387-5581, e-mail: Robbychampion@aol.com.

[Click for NSDC Home Page](#)

Kirkpatrick's Four Levels of Evaluation

ASSESSING TRAINING EFFECTIVENESS often entails using the four-level model developed by Donald Kirkpatrick (1994). According to this model, evaluation should always begin with level one, and then, as time and budget allows, should move sequentially through levels two, three, and four. Information from each prior level serves as a base for the next level's evaluation. Thus, each successive level represents a more precise measure of the effectiveness of the training program, but at the same time requires a more rigorous and time-consuming analysis.

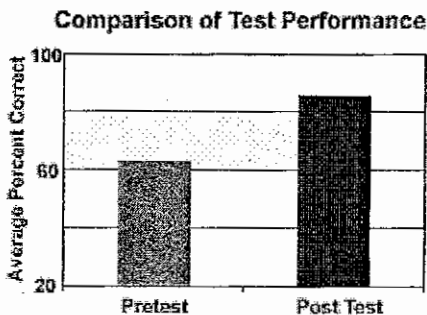


In Kirkpatrick's four-level model, each successive evaluation level is built on information provided by the lower level.

Level 1 Evaluation - Reactions

Just as the word implies, evaluation at this level measures how participants in a training program react to it. It attempts to answer questions regarding the participants' perceptions - Did they like it? Was the material relevant to their work? According to Kirkpatrick, every program should at least be evaluated at this level to provide for the improvement of a training program. In addition, the participants' reactions have important consequences for learning (level two). Although a positive reaction does not guarantee learning, a negative reaction almost certainly reduces its possibility.

Level 2 Evaluation - Learning



To assess the amount of learning that has occurred due to a training program, level two evaluations often consist of tests conducted before training (pretest) and after training (post test).

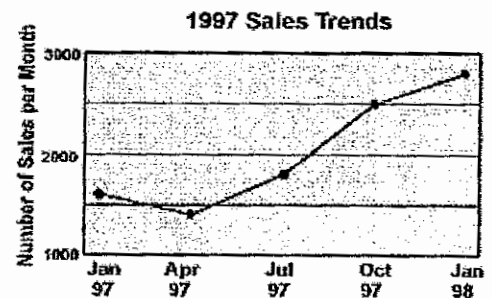
Assessing at this level moves the evaluation beyond learner satisfaction and attempts to assess the extent students have advanced in skills, knowledge, or attitude. Measurement at this level is more difficult and laborious than level one. Methods range from formal to informal testing to team assessment and self-assessment. If possible, participants take the test or assessment before the training (pretest) and after training (post test) to determine the amount of learning that has occurred.

Level 3 Evaluation - Transfer

This level measures the transfer that has occurred in learners' behavior due to the training program. Evaluating at this level attempts to answer the question - Are the newly acquired skills, knowledge, or attitude being used in the everyday environment of the learner? For many trainers this level represents the truest assessment of a program's effectiveness. However, measuring at this level is difficult as it is often impossible to predict when the change in behavior will occur, and thus requires important decisions in terms of when to evaluate, how often to evaluate, and how to evaluate.

Level 4 Evaluation- Results

Frequently thought of as the bottom line, this level measures the success of the program in terms that managers and executives can understand -increased production, improved quality, decreased costs, reduced frequency of accidents, increased sales, and even higher profits or return on investment. From a business and organizational perspective, this is the overall reason for a training program, yet level four results are not typically addressed. Determining results in financial terms is difficult to measure, and is hard to link directly with training.



Level four evaluation attempts to assess training in terms of business results. In this case, sales transactions improved steadily after training for sales staff occurred in April 1997.

Kirkpatrick, D.L. (1994). *Evaluating Training Programs: The Four Levels*. San Francisco, CA: Berrett-Koehler.

Evaluation Tip 1: Organizing Your Thinking About Evaluation

One way to think about evaluation is to use a model that provides concrete definitions of what can be learned from them. Kirkpatrick's four-level framework is one example that can be used. This framework consists of four levels that progress in difficulty from 1 (the easiest to conduct) to 4 (the hardest). When choosing the levels to include in your assessment, start by identifying the questions your evaluation needs to address.

KIRKPATRICK'S FOUR-LEVEL EVALUATION SCHEME

Level	Measurement focus	Questions addressed
1 - Reaction	Trainees' perceptions	What did trainees think of this training?
2 - Learning	Knowledge/skills gained	Was there an increase in knowledge or skill level?
3 - Behavior	Worksite implementation	Is new knowledge/skill being used on the job?
4 - Results	Impact on organization	What effect did the training have on the organization?

Examples:

Level 1: One way to assess trainee reactions and attitudes is to use a questionnaire.

- Questions can gather opinions about training methods, the instructor, the environment in which training took place, or other aspects of the training process.
- Pencil-and-paper surveys are convenient to use for trainees and the evaluator.

Level 2: Written or performance tests can assess change in knowledge/skills.

- The best way to measure changes in knowledge or skills is to test trainees before and after training.
- Note that even if a positive change is found, it is possible the trainees gained the new knowledge or skill from a source other than the training.
- If it is not possible to test trainees before training, their performance can be tested after training and they can be asked whether or not their understanding or skill came from the training session.

Level 3: Post-training testing or observations can assess use of skills on the job.

- This level must be completed outside of the classroom after trainees have had an opportunity to use what they have learned.

- This level is more difficult because it requires trainers or some other evaluator to follow-up months after training.

Level 4: Quantifiable measures are often used when assessing organizational impact.

- Some examples of measures that can be used are numbers regarding sales, injuries, or productivity.
- It can be difficult to determine the extent to which other factors besides training (i.e., economics of region) may have contributed to changes in organizational performance.

Key Points to Remember:

Different aspects of training can be evaluated.

- Level 1: Trainees' perceptions
- Level 2: Knowledge/skills gained
- Level 3: Worksite implementation
- Level 4: Impact on organization

Remember to direct your evaluation to specific questions.

- How did the trainees react to the training?
- Do people seem to have increased their skills?
- Are new skills being used?
- Has the organization been impacted?

Reference: Kirkpatrick, D. 2001. The four-level evaluation process. Ch. 12 in *What Smart Trainers Know: The Secrets of Success from the World's Foremost Experts*, L.L. Ukens, ed. San Francisco: Jossey-Bass/Pfeiffer, pp. 122-132.