Evaluation in the Implementation Phase

Evaluation in the Implementation Phase takes many forms. The Kirkpatrick Model is a popular summative evaluation model used to evaluate training programs. It is most appropriately used in the implementation phase after the product is stable (i.e., fully developed) to determine if learning has occurred and what impact the learning has inspired. However, other evaluation activities might also include an implementation fidelity study, negative case analysis, and exploring unintended consequences that result from the products use or programs implementation.

Evaluation in the implementation phase is primarily summative. The goal of summative evaluation is often seen only as a judgment of effectiveness, but this is not the case. Indeed, instructional products usually are created to meet specific learning objectives, and it may be important to determine whether those objectives have been met. However, we still need to know how well the final product turned out in terms of efficiency and user satisfaction. Likewise, not all products are designed with a specific learning objective in mind—these can also be evaluated summatively. In practice, an external evaluator will often be asked to conduct a summative evaluation after the product (e.g., an educational initiative, program, or policy) has been implemented for some time. The summative evaluation may become part of a maintenance evaluation. In this case, the evaluation can be formative, asking questions like: Is this educational program (or product) still needed? Should we continue to support it? Does it need to be updated? Is it still being implemented as intended? Do learners like the product? And, Is it effective? To answer each of these questions would require a variety of evaluation methods. These questions deal with long-term effectiveness, impact, accessibility, and satisfaction issues.

Kirkpatrick model

The Kirkpatrick Model is a popular summative evaluation model used to evaluate training programs. While it was designed to evaluate training, the strategies utilized in this model can be adapted to evaluate other instructional products. This model is most appropriately used in the implementation phase after the product is stable (i.e., fully developed). It requires a longitudinal time commitment and focuses on satisfaction, effectiveness, and impact. However, evaluation activities could be added to deal with accessibility (e.g., a negative case evaluation) and additional usability issues (e.g., implementation fidelity and efficiency). There are four levels or phases in the Kirkpatrick model.

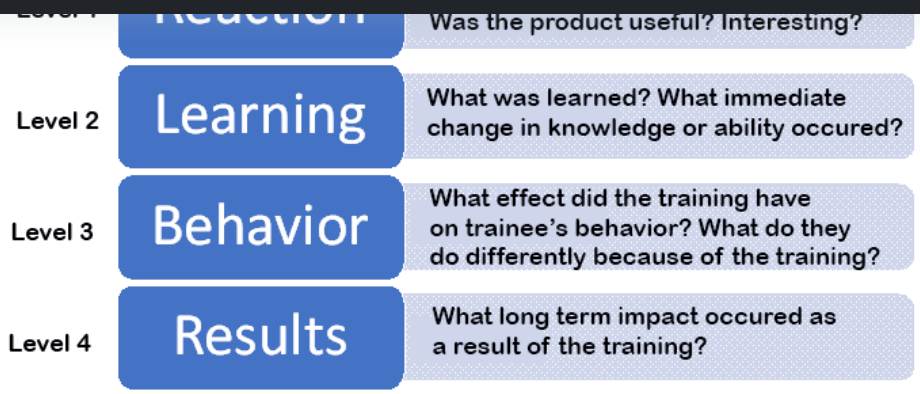

Figure 1: Phases in the Kirkpatrick Model.

Level 1: Reaction - the degree to which participants find the training favorable, engaging, and relevant to their jobs.

> Evaluation activities at this level focus on user satisfaction. Surveys, interviews, and observations might be used to ascertain users' reactions to the training (or product). The evaluator attempts to determine how individuals felt about the experience—the participants' perceptions. At this point, an evaluator might ask participants to self-report how much they learned? Was it beneficial? Did they enjoy the experience? What they plan to do differently because of the training (i.e., motivation)? The evaluator may also attempt to obtain some formative evaluation information regarding how the experience might be improved? This is also where an evaluator might conduct an implementation fidelity evaluation.
>
> An **implementation fidelity evaluation** judges the degree to which a program was implemented as intended. The evaluator looks at consistency and whether the person providing the training changes or adapts the training in any way. If changes were made, the evaluator asks why the changes were made then attempts to determine whether the changes were needed and appropriate (i.e., beneficial). The evaluation might also look at whether the training can be consistently implemented as intended (i.e., a usability issue).

Level 2: Learning - the degree to which participants acquire the intended knowledge, skills, attitude, confidence, and commitment based on their participation in the training

> At this level, the evaluation focuses on immediate learning gains. Here the evaluator should not rely on self-report data, which is notoriously inaccurate, but should assess (i.e., test) how well students achieved the intended learning objectives. Evaluation at this level addresses the criteria of effectiveness, actual, not just perceived. In addition to measuring student achievement, this is an excellent place to introduce a negative case analysis.
>
> A **negative case analysis** asks the effectiveness question in reverse. Most effectiveness evaluations are success case assessments—a negative case analysis looks at failures. The evaluator will identify those who benefited from the training, but more importantly, those who did not. Rarely will a product work well for all learners! The purpose of a negative case evaluation is to identify cases where individuals failed to achieve the intended learning outcomes then ascertain why this was the case.

Level 3: Behavior – the degree to which participants apply what they learned during training when they are back on the job.

> Here the evaluation assesses mid-range outcomes. The evaluator conducts follow-up observations and interviews to determine what effect the training had on changing behavior. Evaluation at this level also addresses the effectiveness criteria but focus on results other than acquired learning and ability. You will recall that training is needed (and created) to meet an identified performance gap. The purpose of the training is to

made a difference – did the training solve the performance problem? An additional performance gap analysis might be needed if the training did not achieve the desired effect. You will need to determine why these trained (more knowledgeable and capable) individuals still are not performing as expected. This is similar to a negative case analysis in that you want to determine the reason for the failure.

Level 4: Results - the degree to which targeted outcomes occur as a result of the training and the support and accountability package.

Evaluation at this level focuses on long-term impact. This is perhaps one of the most challenging things to do and often requires a long-term commitment to the evaluation. Measuring impact goes beyond gauging changes in the performance of an individual. It addresses the "so what" question. Now that employees are more capable and are doing the job as intended, what impact does this have beyond doing a task correctly? What benefits were achieved or goals met that resulted from doing the job well? Is the world a better place? Did the company see increased profits? Do customers have greater satisfaction? Do you have more customers? Do they recommend this product to others?

Measuring Effectiveness

Models outline what needs to be done but not often how. For some evaluation purposes the methods are clear. For example, measuring satisfaction requires the evaluator capture the perceptions of the participants. This inevitably means self-report data collection instruments like surveys and interviews. Measuring usability and efficiency also require user feedback. Observations, interviews, and focus groups can be utilized to obtain the information needed to accomplish usability evaluations. Measuring effectiveness is more challenging. It usually means using quantitative methods.

Objectives Oriented Evaluation. This type of evaluation judges the effectiveness of an educational product by testing a learner's ability after they have used the product. This works best when expected learning objectives are defined in great detail. It also requires valid assessment instruments (i.e., tests) that are designed to measure the learning outcomes. There are three common types of objectives: cognitive, performance, and affective objectives.

Cognitive objectives. These objectives measure thinking skills. They deal with lower-level thinking skills like knowing facts and understanding concepts. As well as high-level cognitive abilities like analyzing, critical thinking, synthesizing ideas, and evaluating. Tests that measure cognition require learners to recall, explain, compare, justify and produce logical arguments (see Figure 2). In schools, cognitive assessments are common and sometimes standardized tests exist to measure specific objectives. The evaluator must either find or create assessment instruments that align with the intended learning objectives. The data from these assessments are used to judge the effectiveness of the educational product.

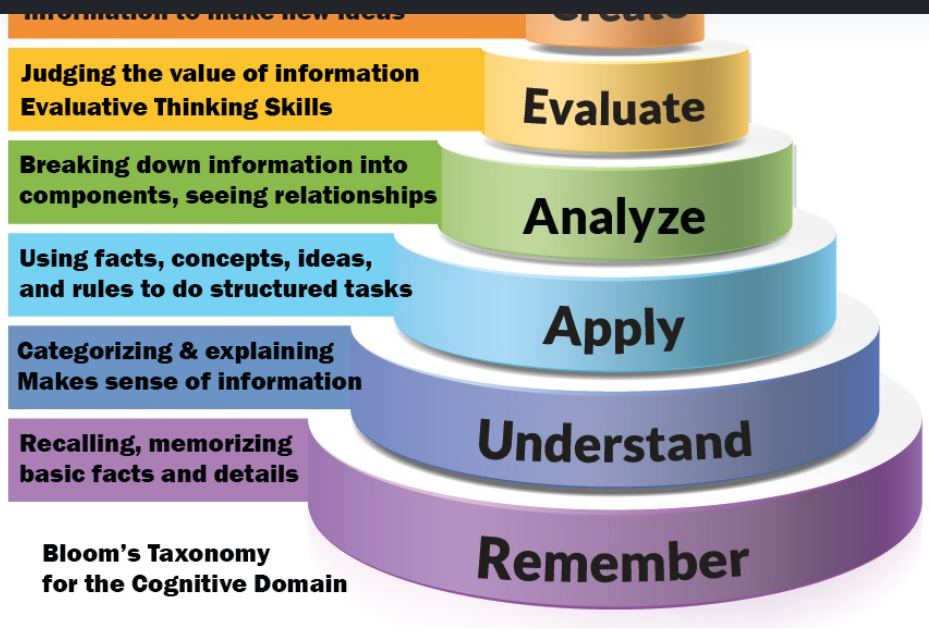

Figure 2: Bloom's Taxonomy for the Cognitive Domain. (adapted from source)

Performance objectives. These abilities must be measured using performance assessments. Tests that measure performance require learners to demonstrate their ability to perform a task or skill, not just know how to do it (i.e., the steps required). Behaviors or abilities that require performance assessments might include reading, speaking, singing, writing, cooking, or doing some other clearly defined task (i.e., a job). These assessments are subjectively scored and require the use of a rubric. The rubric (or scoring guide) outlines essential aspects of the skill and the criteria for judging competence. This is best done with expert reviewers or judges. For example, the ability to speak a foreign language should be tested with an oral proficiency interview. Those administering the test should be experts in the language and trained in the assessment's administration protocols. If the goal is to test a person's oral communication skills, it would not be acceptable to have a student pass a vocabulary test, or a reading test, then declare them a capable speaker. They must be able to speak fluently, have an adequate vocabulary, and respond appropriately (i.e., intelligently) to prompts.

Affective objectives. Affect refers to the personal feeling—one's beliefs, opinions, attitudes, dispositions, and emotions. When evaluating educational products we usually attempt to measure satisfaction (an emotion). However, at times the objective of a particular educational program is to help students develop specific dispositions or attitudes. A course curriculum may have the goal of developing students' character or generating a specific perspective. For example, educators prefer a student develop an internal locus of control (i.e., a belief that their efforts make a difference). Having this attitude is believed to result in better effort on the part of the student and as a result, more learning. However, with most affective characteristics, you cannot simply ask someone to tell you what they feel directly. You must measure the degree to which they hold a specific attitude or opinion using a scale. Some scales used to measure specific affective characteristics already exist, others would need to be created. This can be challenging (see scale development).

Experimental Methods. Measuring the degree to which objectives were achieved at the end of instruction (or after using a product) can be helpful, but it may not be an adequate indicator of a product's effectiveness. Sometimes it is enough to know that individuals have the desired abilities regardless of whether the educational product they used was effective. However, evaluation research usually needs to obtain evidence that the product facilitated (i.e., caused or contributed to) the expected learning. Experimental research designs are used to verify effectiveness. The problem with a post-test-only evaluation of learning is that we don't know whether the product facilitated the learning, whether the student already had the capabilities the instruction targeted, or some other factor caused the learning to occur. To do

Pre-Post experimental designs. Basically, this type of efficacy testing establishes a baseline assessment of ability or knowledge using a pre-test assessment. The results of the pre-test are compared to post-test assessment results. If the product facilitated learning, you would expect to see a meaningful increase in achievement or ability. It is also best practice to verify that the product was actually used or implemented as intended. For example, you may ask someone to use a product, but they may choose not to use the product or use it incorrectly. If you assume the product was used when it was not, the results will be flawed.

Control-Treatment experimental designs. You may also use pre- and post-tests when comparing treatment and control groups; however, to determine effectiveness in this way, a treatment group uses the educational product — the control group does not. You then compare results. If the treatment group results were meaningfully better than those of the control group, you can infer the product was effective.

Ethically you may not be able to withhold treatment. In these cases, a regression discontinuity model may be utilized, where the control group is tested without receiving the treatment; this result is compared to the post-test results of the treatment group. Later the control group receives the treatment and is again tested. If the treatment was effective, you would expect the control group to obtain similar achievement levels to that of the treatment group once they started using the educational product. With this kind of testing, the evaluator attempts to ensure the control and treatment groups are similar in their makeup (e.g., age, gender, developmental readiness, interests, ability).

Example of Inappropriate Effectiveness Measures

Evaluators often are asked to measure the effectiveness of a program or initiative. The client may want this information for a variety of reasons, however, often it is required to support requests for funding. Measuring effectiveness can be a challenge when the goal or objective of the initiative is affective in nature. For example, the sponsors of a character-building program wished to have an effectiveness evaluation completed. The specific aims of the program were to develop an attitude of acceptance of others, respect, kindness, as well as positive feelings of self-worth in participants. Unfortunately, they did not have and were unwilling to create measurement instruments that assessed the specific attitudes and perspectives the program was designed to promote. Instead, they proposed using student achievement data from the state's standardized test as an indirect measure of character. They assumed that if students did well in school, this would be an indicator of good character. Obviously, this would be inappropriate. How well students perform academically has little bearing on their personal feeling about themselves or their attitude toward others. How well students perform in school is not an adequate measure of the program's effectiveness because these data are not directly related to the program's learning objectives.

Measuring Impact

Impact and effectiveness are often used as synonyms – but they are not the same. Effectiveness is an indicator that the product worked. The impact of an initiative (or product's use) describes what happened because the product was effective beyond the fact that specified learning outcomes were achieved. Impact evaluation also includes identifying the unintended consequences of implementing a program or using a product (i.e., the side effects). For example, a training program may increase athletes' ability to jump higher. The product may be effective, but does that result in the team winning more games? And, because athletes are jumping higher, do more injuries happen? Impact addresses the

To determine the impact of an educational product, you need to identify the broader success indicators beyond whether the product functioned as intended and students achieved the stated learning objectives. You need to capture baseline data for those variables and compare them with data obtained after the educational product accomplished its purposes. Impact does not often occur immediately after a product, policy, or program is implemented; it usually requires time for desired long-term goals to be realized. Because of this, impact evaluation is not often done. Unfortunately, many educational initiatives are implemented and found to be somewhat effective, but the long-term impact is not determined. In addition, many evaluators fail to explore the unintended consequences of using a product or implementing an educational policy or program. Impact evaluation often requires using qualitative or mixed methods to properly understand a product, initiative, or policy's full impact. Focusing only on specific success indicators is a type of quasi-evaluation. The evaluator must be open to exploring the broader picture.

Impact vs. Effectiveness

A lawsuit was filed against a fragrance company for false advertising. The issue was one of impact vs. effectiveness. The product was intended to make people smell better – and it worked as promised. However, the advertisement for the product implied that the result (i.e., impact) of smelling better would be increased interest from the opposite sex (i.e., the success indicator) – this did not happen as suggested, nor as the participants hoped (something to do with personality and perhaps good looks). Even when a product is effective (i.e, it works), the intended, hoped-for impact may not ensue. For example, an educational program may be effective – it facilitates an increase in students' knowledge and ability; but, for various reasons, the impact may be small – students may not obtain employment as hoped.

- Evaluation in the implementation phase is mostly summative.
- Summative evaluation in this phase deals with effectivness and impact, but also might include satisfaction and negative case evaluations.
- The Kirkpatrick model is commonly use in this phase to evaluate training but can be adapted to evaluate other educational products.
- The Kirkpatrick model evaluates user satisfaction (reaction), effectiveness (changes in learning and behavior), and impact (results).
- Evaluating effectiveness is done by measuring learning outcomes.
- Outcomes might be cognitive, performance, or affective in nature. Each requires a different form of assessment.
- Experimental research designs can be used to validate effectiveness.
- Implementation fidelity studies are often used as part of an effectiveness study but also as a measure of usability.
- Summative evaluations should also consider unintended consequences of using a product or implementing a program or policy.
- Impact and effectiveness are not the same things. An educational product is effective if it accomplishes its intended purpose. The impact of a product goes beyond a product's effectiveness. It addresses the value and benefit of a product in a more general sense.

Discussion Questions

1. Consider a particular training program you have attended (or an educational product you have used). Explain briefly how an evaluation might address each level of the Kirkpatrick model? When would the evaluation activities occur and how would you gather information?
2. Think of an educational product. Explain how you might determine the product's effectiveness and its impact. Describe how you might obtain information about the unintended consequences of using the product, implementing the program or policy?

Access it online or download it at
https://edtechbooks.org/eval_and_design/summative_evaluation.