Evaluation and Design

Randall S. Davies

Evaluation Basics 1

Evaluation within a Design Context 3

Evaluation for Instructional Designers 11

Evaluation Planning 21

Research Principles Primer 31

Evaluation Approaches for Designers 39

Evaluation in the Analysis phase 45

Evaluation in the Design Phase 51

Evaluation in the Development Phase 59

Evaluation in the Implementation Phase 69

Personnel Evaluation 77

Data Collection Primer 81

Reporting Evaluation Findings 87

EdTech Books

URL: https://edtechbooks.org/eval_and_design

Davies, R. S. (2022). *Evaluation and Design*. EdTech Books. https://edtechbooks.org/eval_and_design

Randall S. Davies

Brigham Young University

Dr. Randall S. Davies is an associate professor of Instructional Psychology & Technology at Brigham Young University (BYU), where he teaches courses in educational inquiry, measurement, and evaluation. He previously worked at Indiana University South Bend as an Assistant Professor of Educational Research and at IU Bloomington in the Center for Evaluation and Education Policy. He received his Ph.D. in Instructional Psychology & Technology from BYU as well as a B.Ed. in secondary education and a B.Sc. in mathematics & computing science from the University of Alberta.

Evaluation within a Design Context

In the design process, evaluation questions are asked with the intent of providing information regarding the merits of taking a specific action. Decision-oriented questions describe an evaluation's purpose and justify the need for an evaluation. Evaluation questions may ask: Which option would be best? Which features of this product are essential, and which are unimportant?

Why Evaluate? Help me decide.

Many evaluation approaches have been developed. Those categorized as *decision-orientated or utilization-focused approaches* tend to work well in the field of instructional design. The instructional design process involves making a series of design decisions; evaluations are made to inform your design decisions and thus improve your designs.

The things we evaluate are called evaluands (or evaluees when evaluating people). In educational settings, we might evaluate people, programs, initiatives, policies, products, equipment, services, concepts, theories, or organizations. However, when we evaluate programs, services, and organizations, we evaluate the people who administer the program, provide the service, or run the organization. These people are what Stufflebeam (2014) calls – ghosts in the system. Regardless, the underlining purpose for conducting an evaluation is most often because we need to decide something about the evaluand (i.e., the thing we are evaluating). Rarely is an evaluation conducted just to obtain information. The information we collect is intended to facilitate our ability to make informed decisions.

The decisions that need to be made can be varied. In formal evaluations, the reason for conducting an evaluation is explained in a purpose statement; this includes the decisions that need to be made and often the central evaluation questions that will inform the decision.

The motivation behind a decision-orientation question is a need to know what *should* be done. Decision-oriented questions describe an evaluation's purpose and justify the need for an evaluation. On the other hand, evaluation-oriented questions ask for an informed opinion or judgment. The answers to evaluation-oriented questions form the rationale for the recommendations provided. This usually includes considering the potential benefits and disadvantages of doing one thing instead of another.

Decision-oriented questions are usually somewhat dichotomous; they help us choose one course of action over another. Rarely are these questions "right vs. wrong" types of determinations. They often involve deciding between "good, better, and best" options. We are trying to decide how to proceed, which may mean deciding not to do anything at all.

On the other hand, evaluation questions are rarely dichotomous. They often describe the "degree to which" the evaluand meets a particular set of criteria and can be judged of value, merit, or worth. The need for a judgment may be implied or stated but will always be part of the question. Evaluation-oriented questions inform decisions by asking: what is the best, most valuable, most important, safest, satisfying, beneficial, rewarding, desirable, practical, helpful, ethical, moral, or smartest thing to do?

The answers to evaluation questions can inform a variety of decisions. Should we continue to develop/fund a product or initiative? Does the product need to be modified or completely redesigned? Should the product be discontinued

unimportant? Which designs are viable? Is this initiative working as intended? Are users satisfied with this product? To what degree can the product be used, or the program be implemented as intended? To what degree was this training effective?

We assign merit based on specific criteria, making evaluation a very subjective process. People have different opinions because they value different things. Still, answering evaluation questions provides us with information that we need to make well-reasoned recommendations that are then used to inform decisions.

Design questions are posed because we need to decide on an action to be taken.

Evaluation questions are asked with the intent of providing information regarding the merits of taking a specific action.

When conducting an evaluation, you will note that not all the questions we ask will require a decision to be made or a determination of merit and worth. Information queries are asked simply to get information, and descriptive research provides information about relationships, potential causes, and general conditions. Experimental research makes comparisons to establish statistically significant differences (i.e., differences that likely didn't happen due to chance). This information can be valuable to evaluators. In general, research-oriented questions are posed to better understand an issue or topic. Research is conducted to draw generalizable conclusions that add to the body of knowledge in a particular field; they do not, however, attempt to make evaluative judgments. Still, research and information queries can be beneficial and serve distinct purposes for evaluators, designers, and developers. Evaluation questions, information queries, and research questions can all be used to provide the information we need to make informed design decisions.

Decision-oriented questions ask what should be done.

Evaluation-oriented questions ask what is best and always involve value-based judgments.

Research-oriented questions and information inquiries can also provide valuable information needed to evaluate.

A Basic Decision-Making Process.

Deciding involves choosing between a finite set of alternatives. Making an informed decision requires an analysis of the benefits and disadvantages of choosing one option over another. Making an informed decision involves making evaluative inquiries and collecting information.

A variety of decision-making models exist. To make an informed decision, you probably want to consider using a structured decision-making approach. To do this, you need to:

3. Make a list of possible alternatives
4. Identify the information you will need to make an informed decision
5. Identify the criteria you will use to evaluate possible solutions/alternatives
6. Determine how important each criterion is
7. Evaluate each alternative (judge the merit, worth, and value)
8. Determine the best solution

Sometimes the decision process can be completed quickly (minutes or hours). However, for some decisions, a more comprehensive evaluation is required to make an informed decision. Depending on the scope of the evaluation, the information-gathering activities required could take several weeks or longer.

Common Mistakes and Problems.

Sometimes we make good choices and other times not so much. Several common mistakes (biases) can affect the quality of a decision. Even knowing that they exist and that you may be vulnerable to them can help you make better decisions. We are all biased, but we can train ourselves to be less biased.

Many cognitive biases can affect decision-making. Here are a few mistakes (biases) that can cause poor decision-making.

Confirmation Bias.

Confirmation bias means paying attention to evidence that confirms our personal beliefs while ignoring anything that doesn't. Confirmation bias causes us to resist novel (different) ideas and solutions. This can cause us to intentionally or unintentionally interpret or twist information to fit our preconceived notions. Because of confirmation bias, two individuals with different beliefs will often draw different conclusions from the same information. It can also cause us to seek out information that supports our existing views or preferred solutions.

Confirmation bias often happens unintentionally so it is something we need to be aware of when making decisions. To overcome a confirmation bias, seek out people and information that challenge your opinions even if you are confident that the evidence supports what you want to do. Attempt to fully understand opposing viewpoints and the evidence that supports a different solution. Consider alternative interpretations of the data you have. You should also consider the potential unintended consequences of each decision.

Availability Heuristic.

The availability heuristic leads us to make decisions based on the first viable option or solution that comes to mind. This bias is reactionary and often emotional. Someone says something, and you act on your first inclination. If that inclination was to punch them in the nose, you might regret not taking time to consider other options. Impulse buying is another example of this bias. You identify a need (perceived or actual) and immediately go online and purchase the first solution (product) your search engine displays.

To overcome this bias, you need to be deliberate. Force yourself to consider alternative solutions. Delay deciding anything until the initial emotion of the situation has dissipated. Doing this will reduce the chances of experiencing buyer's (or decider's) remorse. Make sure you have all the information you need and that the evidence truly supports your making that decision. We tend to make better decisions when we pause, second-guess ourselves, and take time to obtain sufficient information to make an informed decision.

Survivorship Bias

The survivorship bias causes us to make decisions based on previous successes (or failures) without fully considering the context and circumstances that led to that success or failure. This bias causes us to resist change; it impedes us

change) something that works. Survivorship bias is often the result of fear – fear that change will be harmful or that we will be unable to make the changes that are needed. Remember, however, that fear of change may be warranted. Not all change is beneficial or needed. This bias restricts one's ability to make changes even when the evidence and circumstance warrant making a change.

Survivorship bias is mitigated by training yourself to be more introspective. Consider why you may be resisting a decision that requires change. Before deciding anything, consider the factors and conditions that resulted in previous successes or failures. Ask yourself whether the conditions that fostered those successes or failures still exist. Consider whether future success is likely given current trends and conditions. Force yourself to consider whether improvements could and should be made based on your evidence. Explore the possible benefits of making a change and the consequences of not doing so.

Anchoring bias and Priming

Anchoring bias causes us to fixate or rely heavily on the initial information we receive about a topic. When we reflect on that topic later, we interpret additional (possibly updated) information from the reference point of that anchor instead of seeing the new information objectively. This bias can happen regardless of whether the initial information is even related to the new information. When people are exposed to an idea or piece of information prior to making a decision, that idea or information influences their judgment and can skew one's judgment. Anchoring happens in these cases because the individual has been primed cognitively to consider the initial information as a baseline or reference.

What is black and white and read all over? This simple riddle is based on an anchoring bias. The individual is primed to think about colors. Then when asked to consider something that isn't a color (read vs. red), the individual has difficulty thinking about anything not associated with color. This riddle tends to work whether it is presented verbally or in writing.

Being aware of this bias and how vulnerable people are to this bias is important for evaluators and decision-makers. The best way to combat an anchoring bias is to slow down your decision-making process. Insist on taking time to consider the information carefully before deciding.

Halo Effects and Logical Associations Errors

First impressions are powerful. People have a hard time overcoming a first impression (positive or negative). We call it a halo effect when a first impression affects future decisions. A halo effect exists when your first impression (or experience) is negative and you still tend to think negatively about that person or product even when presented with contradictory information later.

Likewise, a logical association error can cause us to make poor decisions. If we like one thing about a person or product, we illogically assume we will like other characteristics of that person or product. Most product sales are not based on an objective evaluation of the product but rather on the charisma and likability of the salesman. Making a decision based on unrelated information is an example of a logical association error.

A healthy dose of skepticism can help counterbalance these biases. You should also consider evaluation criteria separately before making a holistic assessment. A person can excel in one area and fail in another. It is also possible for products to be beneficial for some and not for others. For this reason, it is crucial to conduct a negative case analysis when evaluating the effectiveness of a product or program.

Planning Fallacy

The planning fallacy describes our tendency to underestimate or overestimate costs and risks when we like or dislike a proposed solution. Providing or accepting incorrect information skews our assessment of the viability of taking a specific action. For example, we often underestimate the amount of time it will take to complete a task when deciding when to start an undesirable task; this is especially true if we must forgo a pleasant activity. A planning fallacy can

The best way to overcome this problem is to get accurate information. Get the information from reliable primary sources when possible. You will also need to be mindful about how you and others feel about a specific solution so you can gauge the accuracy of the information you are receiving.

Action Bias

Action bias describes a default tendency to favor action over inaction. We often feel the need to do something even when there's no compelling evidence to suggest that acting would lead to a better outcome. For example, sometimes updating a product produces no substantial benefit to users. The product functioned well enough as it was. But the product's producers feel compelled to include additional features, to do something. There may be an economic benefit to producing an "updated" product, but the action may not have been warranted based on functionality deficiencies and consumer demand (i.e., need).

The solution to this bias is to conduct a proper needs assessment and carefully consider the benefits of acting compared to the consequences of not acting.

Analysis Paralysis

Analysis paralysis is not a cognitive bias but rather a condition that stops us from making any decision at all. Paralysis occurs when Individuals or groups cannot make a decision because of the overwhelming amount of data available or the tendency to overanalyze the data. Working with large data sets can be a problem when we are unable to separate important data points from trivial information. This also can be a problem when the evaluator has a limited ability to organize and analyze the available data. For example, you may have thousands of open-ended survey responses to code and interpret with no one capable of or willing to conduct the analysis.

Overanalyzing is a condition that occurs when people get stuck reevaluating data for fear of making a poor decision. It also occurs when individuals go over the same data repeatedly, only to find additional questions rather than answers. One last reason for not being able to decide is because the decision-maker wants to eliminate all uncertainty, which is usually impossible.

Solving this problem can be a challenge because you need to permit yourself to make a mistake. Working with a representative sample of the available data can be a viable solution when working with big data sets. Setting parameters and limiting the time allocated to data analysis and discussion can also help.

Ambiguity Effect

The ambiguity effect is a cognitive bias that describes people's tendency to avoid ambiguity. This is a problem when a decision is avoided because the decision-makers feel they don't have all the information they need to decide. What can happen in these situations is our intolerance for uncertainty and ambiguity inclines us to select an option for which the probability of achieving a favorable outcome is known. In these cases, the decision-maker is not paralyzed; they simply tend to choose options where the probability of success is known over options where the chance of success is unclear.

This is a challenging bias to overcome. Some people are risk intolerant, while others are risk-takers. Neither of these personality characteristics is better than the other. Risk-takers often fail more often than they succeed, while individuals who avoid risk tend not to fail but may miss out on the rewards of taking a risk. To overcome this bias, one must make sure the decision is reasoned and not unduly influenced by their tolerance for risk.

Base Rate Fallacy

A base rate fallacy is the tendency to place more importance on the exception than the norm. For example, 90% of people surveyed may feel a specific aspect of your course's design was tedious and irrelevant; however, the decision-

The values of the decision-maker complicate this bias. For example, having an intuitionist-pluralist perspective leads a person to value individual impact, while having a utilitarian viewpoint favors solutions that promote the greater good. A base rate fallacy can be a logic vs. emotion challenge that is influenced by one's values. In addition, this fallacy can be based on the decision-maker's ability to understand statistics and whether they trust the statistics presented to them. When they do not understand or trust the statistics, they may lean towards something they can understand, personal experiences and individual stories.

Bandwagon Effect

The Bandwagon effect refers to the tendency of people to adopt certain behaviors or beliefs because lots of other people act or believe that way. This tendency can affect decision-making as well. For example, around the turn of the century (2000-2010), Educational Data Mining and Learning Analytics became a popular topic of discussion among those in the field of education. Many organizations decided they needed to use data analytics and data mining even though they had little idea what it was, how to do it, or in what way it might benefit them. It was a popular trend, and they needed to be seen as a leading-edge organization. It worked out for some and not so much for others.

Deciding to act on innovative ideas and practices is not necessarily bad. However, considerable time and effort can be spent implementing a potentially beneficial solution to little or no effect. In terms of decision-making, jumping on the bandwagon should be discouraged if the decision is reactionary or impulsive. Important decisions should be made carefully. They should be informed by evidence obtained through a careful evaluative thought process.

Chapter Summary

- In the field of instructional design, the reason for conducting an evaluation is often to make design decisions.
- Most design decisions are not “right vs. wrong” choices but determinations regarding “good, better, and best” options.
- Design-oriented questions ask "what should be done".
- Evaluation-oriented questions ask for informed opinions about the merit, worth, and value of potential options or solutions. What is best, most important, most desirable?
- Design and evaluation-oriented questions are closely related in the design process.
- Research-oriented questions are often descriptive in nature. The answer to research questions can provide useful information when conducting an evaluation.
- Design decisions are framed by the designer’s goals and values.
- Poor decisions are often made because of lack of information or cognitive biases.

1. Consider a specific decision you need to make. What information would help you make an informed decision? Provide examples of how a cognitive bias might affect the quality of your decision.
2. Some critics of the decision-orientated approaches to evaluation argue that the focus of an evaluation should consider the views, values, and interests of stakeholders other than the key decision-makers. Discuss the pros and cons of this point of view.

Evaluation for Instructional Designers

> *"Everything we evaluate is designed. Every evaluation we conduct is designed. Every report, graph, or figure we present is designed. In our profession, design and evaluation are woven together to support the same purpose—making the world a better place."*
>
> President John Gargani
> American Evaluation Association
> Announcing the Annual Conference theme for 2016

Evaluation is a transdisciplinary field. While teachers and nurses will receive specific training and certifications to work in schools and hospitals, trained evaluators work in a variety of workplaces and businesses. Evaluation is like statistics in this sense. People benefit from using statistics in a variety of different occupations. In fact, people effectively use statistics in their jobs all the time, even when they have limited training and a rudimentary understanding of the statistics they use. As a transdisciplinary art, evaluation is practiced in a variety of contexts. It could be argued that nowhere is evaluation more prevalent than in the field of instructional design. And while designers often conduct evaluations without receiving extensive evaluation training, training and practice will improve their ability to evaluate well; and as a result, it will improve their design and development efforts.

What is an Instructional Product?

As evidenced by the instructional design models developed in the late 1900s (e.g., ADDIE, Dick & Carey), the focus of instructional design was just that, the development of instruction or training. The instructional product was instruction. The modality of the instruction was typically limited to in-person classroom instruction (both academic and corporate). The process included the development of instructional objectives, tests to measure the learning outcomes, and resources (primarily textbooks, learning activities, and videos) the designer believed would achieve the specific instructional goals of that course. The designer would structure the course using a pedagogy they felt would facilitate the intended learning. This method of creating instruction still happens; however, as technology advanced and the internet became more widely available, the notion of what constitutes an educational product has expanded. In addition to instruction, instructional products include educational technologies, learning apps, and educational services in the form of collaborative learning tools, resource repositories, how-to guides, self-improvement and skill development apps, educational games, discussion boards, communication tools, and crowdsourcing apps. The primary modality for delivering instruction has also changed. In addition to classroom instruction, instructors provide training using e-Learning and online instruction, both synchronous and asynchronous, in blended and informal learning environments. In addition, some instructional products have educational purposes related to the facilitation and support of learning in general; these products are not tied to a specific course, and learners use these resources for numerous purposes and in a variety of ways.

It might be helpful to differentiate instructional products (those directly used for training and classroom instruction) from the more generic term of educational products (any product used in an educational setting), but they all have a similar end goal, to facilitate and support learning.

classroom setting. However, more recently, designers have been creating instructional products for non-traditional learners seeking educational opportunities outside of the classroom and any formal educational context. Many eLearning tools are created as supplementary learning resources and knowledge creation services for corporate training or personal enrichment. Several of the more contemporary instructional design approaches (e.g., rapid prototyping and design-based research) have adapted earlier instructional design models to accommodate this expanded view of what an instructional product might be. They still all utilize similar product development stages as all instructional products need to be designed, developed, tested, and maintained, which inevitably requires evaluation.

An instructional product might include any educational resource that facilitates or supports learning regardless of the setting or context.

Instructional Design Models

Many instructional design models have been proposed, but all tend to be an adaptation of the ADDIE model. ADDIE stands for **A**nalyze, **D**esign, **D**evelop, **I**mplement, and **E**valuate.

Figure 1: Original phases of the ADDIE Instructional Design Model.

From the acronym for this model, you may erroneously assume that evaluation only occurs after the designer has implemented the product. This was never the intent, and in practice, evaluations of various types are conducted throughout the project, as depicted in Figure 2.

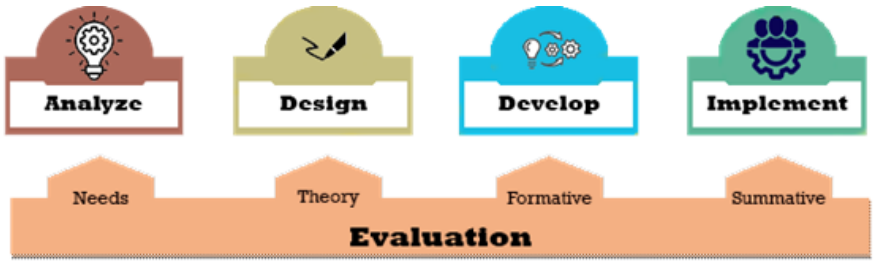


Figure 2: Evaluation Integration within the original ADDIE Instructional Design Model.

Scholars have created many innovative adaptations of the ADDIE model, including the Armed Forces' own modifications to their original training development framework. The PADDIEM version of ADDIE includes a planning phase to augment the analysis phase and a maintenance phase which expands the original purpose of the implementation phase. And while this and other design models each make subtle improvements to ADDIE, they all incorporate an analysis (concept planning), a design (theoretical planning), a development (creation), and an implementation (distribution and testing) phase. A few are presented here as examples of where evaluation occurs in the process.

Dick and Carey's ISD model was one of the early efforts to formalize the instructional design process. This model focused on lesson planning for classrooms and formal training situations. It was intended to help a designer figure out what to teach and how to teach it. It relies heavily on what has recently become known as "backward design." It starts by creating learning objectives and developing assessment instruments (tests) to measure whether students achieved the expected learning outcomes. The findings from the formative evaluation step informed revisions in the instruction. The summative evaluation took the form of an objectives-oriented evaluation, which focused primarily on whether students' test scores were deemed adequate. Achieving adequate test scores was seen as an indicator that the instruction was good and often was the only criteria used to judge the quality of the instructional product.

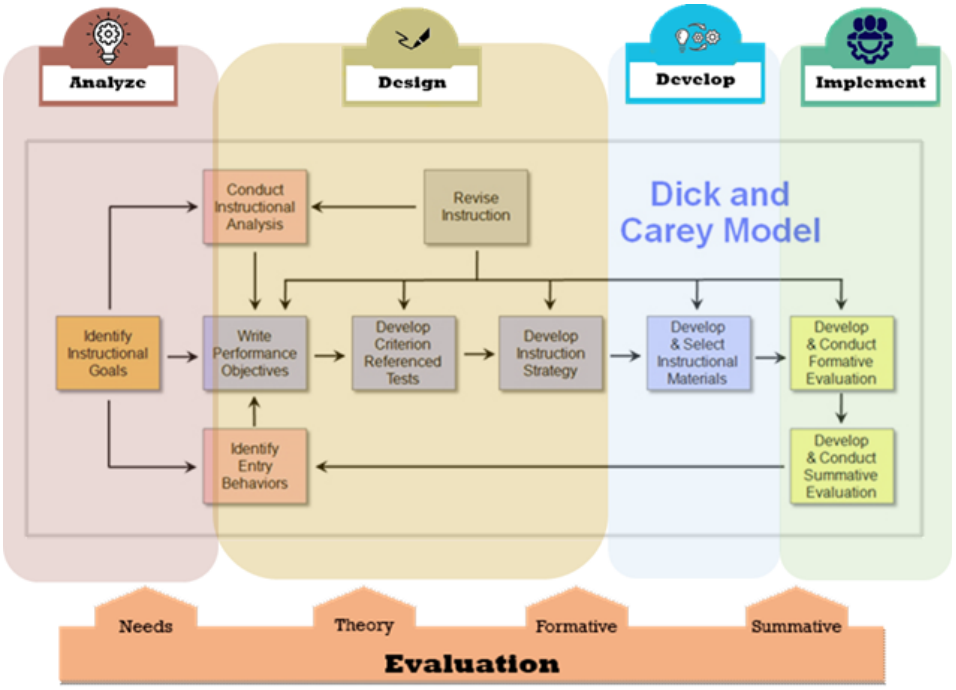


Figure 3: Evaluation Integration within Dick and Carey's Instructional Design Model.

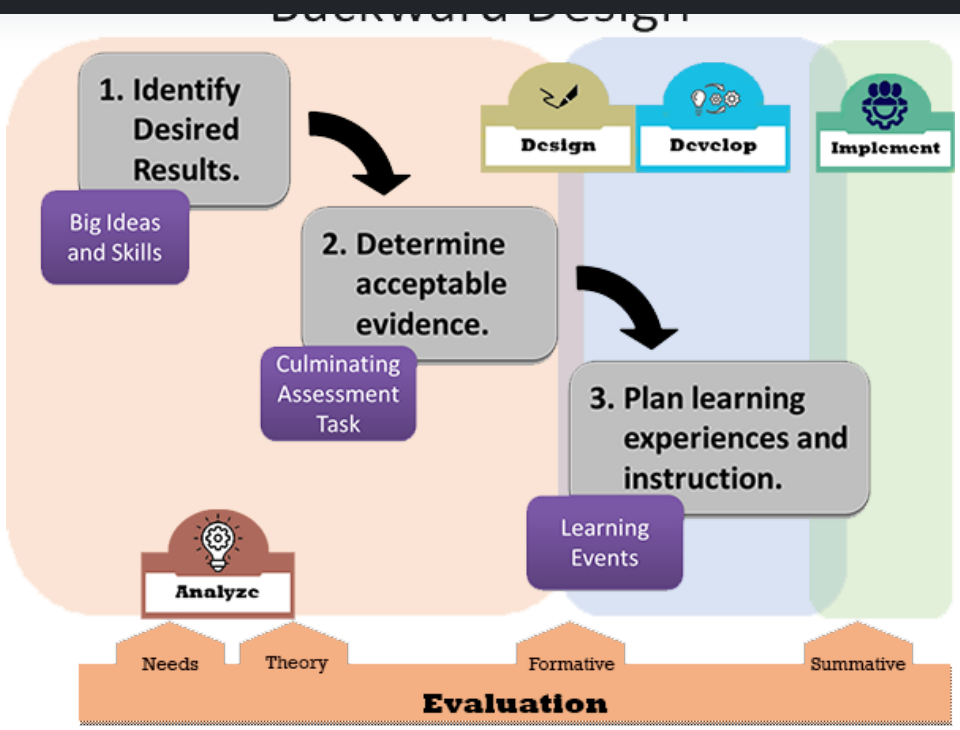


Figure 3a: Evaluation Integration within Backward Design Approach to Design.

Rapid Prototyping and ADDIE

Rapid prototyping was first used in the manufacturing industry. Instructional designers and others adopted rapid prototyping as a quick and cost-effective way to build and test a working version of their product. The innovation that rapid prototyping offers the design process is a quick iterative design and development cycle. The principle supporting this is similar to that used in action research, a trial and error method. The evaluation aspect of this model includes a quick formative review process that informs needed improvements and is repeated until the product meets specifications. Rapid prototyping activities are:

1. Define instructional goals and requirements,
2. Formulate a feasible solution.
3. Start building the product
4. Test it on users and others (evaluate)
5. Refine your design
6. Repeat the process until the product works as required

While this approach is practical, it still follows the same phases of the ADDIE model – just more quickly. In this model, the needs assessment is often limited, and a summative assessment may not occur. This model focuses heavily on the design and development phases. What this model tends to lack is a systematic evaluation of the theory and principles that support the design, which is not uncommon in other models as well.

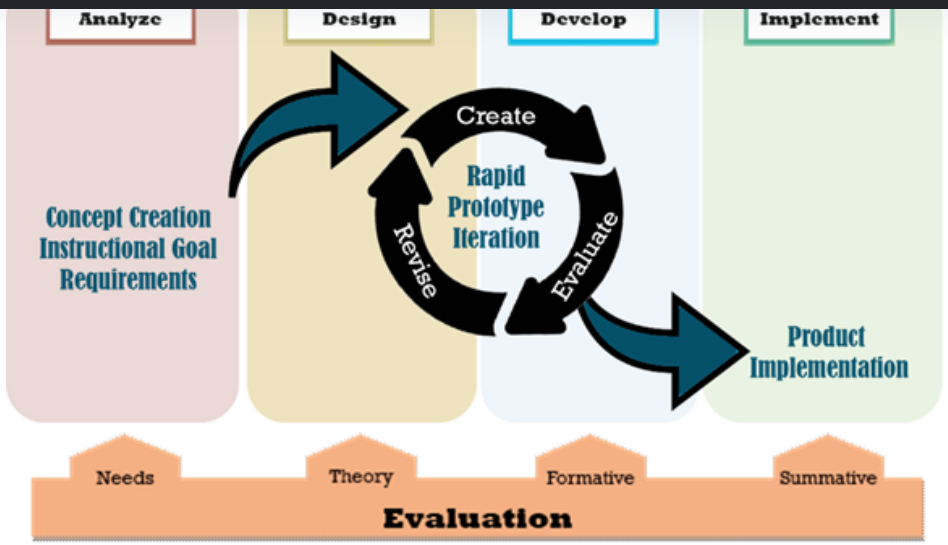


Figure 4: Evaluation Integration within a Rapid Prototyping Instructional Design Model.

Design Based Research (DBR) and ADDIE

McKenny and Reeves (2012) outlined three core processes of DBR: (a) analysis and exploration, (b) design and construction, and (c) evaluation and reflection. A hallmark of the DBR approach is its iterative nature, but you will note that the DBR approach represents another adaptation of the ADDIE model. This approach to design is similar to rapid prototyping but a bit more systematic. Each design iteration is formative in that the designer refines and reworks the product based on understandings obtained in the evaluation and reflection phase of each iteration. How a designer conducts each cycle will depend on the evaluation finding of the previous iteration, and a designer may perform different analyses and use different evaluation methods to complete a cycle. While the core processes identified by McKenny and Reeves do not explicitly state this, we can assume that, in addition to the analysis & exploration that occurs during development, a needs analysis would occur before designers initiate the development process. In addition, we can reasonably assume that a summative evaluation of the final product would occur.

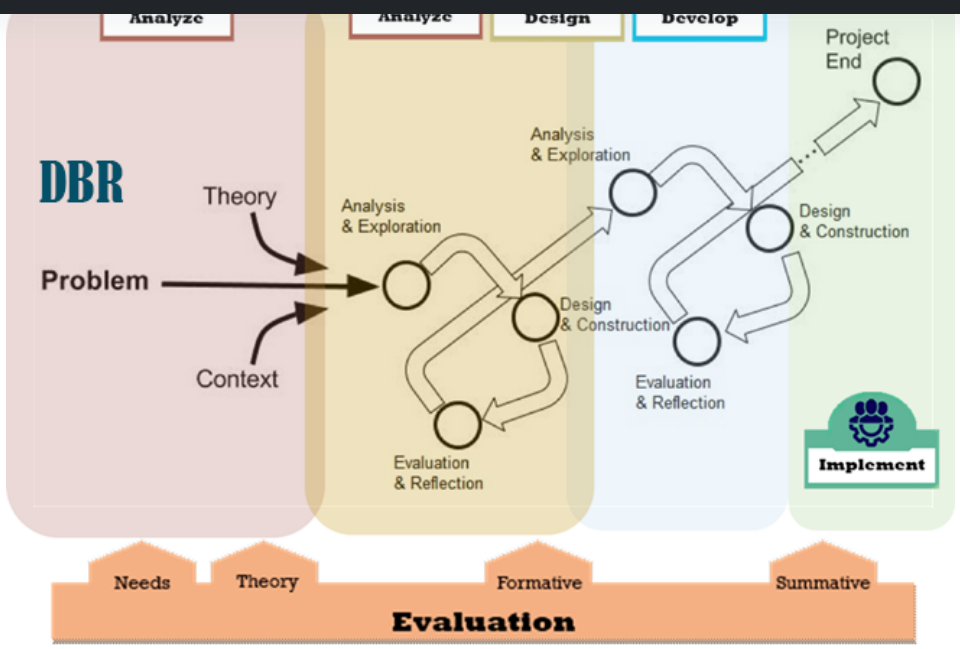


Figure 5: Evaluation Integration within a Design Based Research Model.

The Role of Evaluation in the Design Process

Evaluation is an integral part of the design and development process. Evaluation makes our designs better and helps improve the products we produce. We use evaluation throughout this process. Evaluation is an activity carried out before, during, and after a product has been designed and developed. The following graphic illustrates the various roles evaluation can play within specific stages of the instructional design process.

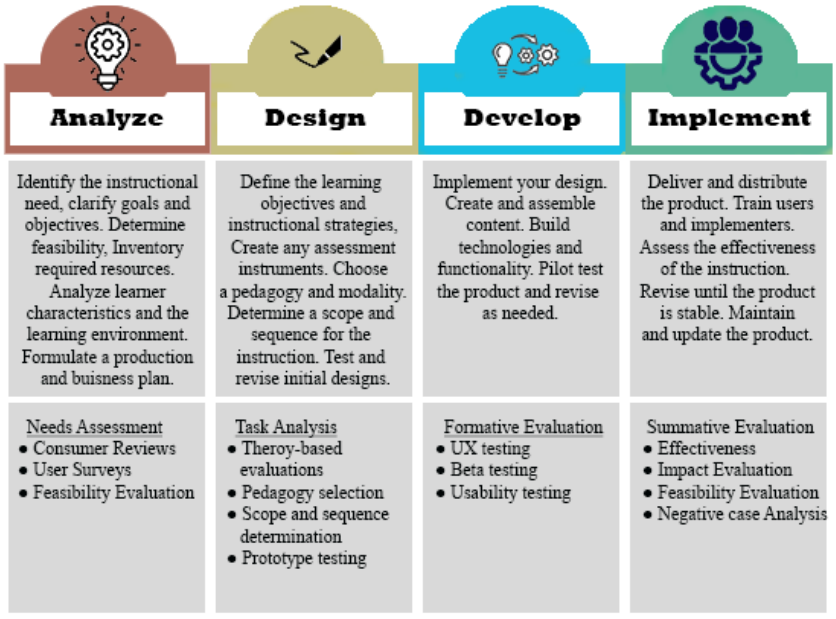


Figure 6: Role of Evaluation in the Design Process.

Evaluation by Design Phase

Designers use different types of evaluation at each phase of the design process; this is because they need to answer different design questions throughout the process. When referring to "an evaluation type," you will note that we refer to an evaluation with a specific focus or purpose. All evaluations follow a fairly standard structure; the purpose, methods, and scale of the evaluation are the things that tend to change.

While the previous chart seems to imply that we designate specific types of evaluation to a specific phase of the process, this is not the case in practice. While a specific type of evaluation may be particularly appropriate for a specific stage, smaller, more focused versions of a specific evaluation may need to be conducted in other phases of the process. For example, a designer may need to conduct a theory-based evaluation in both the design and development phases to help them make decisions. Likewise, designers may conduct consumer reviews as part of a needs analysis evaluation in the analysis and implementation stages, albeit in a modified form and for slightly different purposes.

The following discussion of evaluation roles does not represent a mandate for where evaluation must occur; it simply explores possibilities. We will discuss details of various evaluation approaches and types of evaluation in the next chapters.

Evaluation in the Analysis Phase

The analysis phase of the design process is mainly conceptual. In this phase, the designer analyzes their learners (the target audience and their needs) and attempts to understand any learning requirements and context restrictions (goals and constraints). The main evaluation activity for this phase revolves around needs analysis. A vital component of a needs analysis requires that the evaluator identify any gap that might exist between what is and what we want (need) things to be. For example, the designer may identify a gap between what students know and what they need to know (or be able to do). They then might identify a gap between the quality, effectiveness, or functionality of existing instructional resources and what is needed to facilitate the learning students are expected to accomplish. The designer might conduct a consumer review as part of the needs analysis. A consumer review evaluation will involve surveying users and reviewing and comparing existing products. It may also involve a theory-based evaluation of the product. Results from gap analysis and consumer reviews inform the designer's decision of whether to create a new product or utilize existing solutions. After identifying a need, the designer might conduct another needs analysis to determine the resources needed to produce a new product and the viability of such a project.

You will recall that a planning phase was added to the ADDIE training development framework to meet a specific need that wasn't being met in the original model. Planning focuses primarily on determining the project goals (objectives), requirements, constraints, budget, and schedules (i.e., project management stuff). Planning of this type is needed once the decision-maker decides there is a need for a product to be created or revised—project management benefits from a different set of evaluation activities.

Often designers work for a corporation or an academic institution as part of a design team (e.g., teachers or corporate trainers). In these situations, the client may not expect the products they produce to be sold for profit; they create them to serve a purpose (meet an instructional need within the organization). However, a designer often creates an instructional product to be sold. In these cases, the planning phase may also require the developer to create a business plan to evaluate the viability and cost of product development and whether there is a market for the product.

Unfortunately, too often, very little time is allowed for the planning and analysis phase. At times, clients and designers make quick decisions without carefully considering the need for a product. Cognitive biases that affect decisions made here include action bias, availability heuristic, planning fallacy, survivor bias, and the bandwagon effect. We may perceive a need simply because our personality compels us to act, or we see others developing products and feel compelled to do likewise. We may identify a genuine need but underestimate the cost and risks (viability) of developing

decisions.

Evaluation in the Design Phase

The design stage focuses on the design of the learning experience and the resources needed to support the experience. When designing an educational product, this phase requires a designer to consider the functionality of the product and how the product's design will accomplish its purpose and goals.The purpose of the design phase was initially conceived as a task analysis of the training a designer was hired to develop. A task analysis requires the designer to identify essential components of the learning and problems users experience when learning. Designers then make several decisions regarding the product's design (see Gibbon's layers). A designer must choose which content (information, exercises, activities, features) to include. Designers must also judge the best ways to present the content (i.e., the message) and how a student will interact with the product (modality). Evaluation activities in this phase often involve theory-based evaluation of the pedagogical ideas and principles that might best facilitate the learning and ways an instructional product can mitigate challenges students experience. Theory-based evaluations involve a review of research, and for existing products, an evaluation designed to judge the degree to which a product adequately applies pedagogical theory and principles. Prototype testing is also conducted in this phase to evaluate the viability of a design.

Evaluation in the design phase is essential because if the overall product fails, it is most likely due to a flaw in the design. Designs often fail because the designer neglects to consider existing research and theory related to the product. Even when theory is considered, the teaching and learning process is complicated. People have diverse needs, abilities, and challenges. They also have agency. Rarely will a single instructional design work for all learners. As a result, there is no certainty that all students participating in a learning experience will accomplish the expected learning objective. Likewise, experts often disagree on the best ways to teach. Designers need to judge for themselves which designs are best.

Evaluation in the Development Phase

The purpose of the development stage is straightforward. In this phase, developers implement the designer's vision for the instructional product – they create and build the learning assets outlined in the design phase. This might include the creation of assessments (tests and quizzes), assignments, practice exercises, lesson plans, instructor guides, textbooks, and learning aids. Developers may need to create graphics, videos, animations, simulations, computer programs, apps, and other technologies. They will also need to test and refine each of these assets based on formative feedback from experts, implementers, and the intended end-users.

As noted, the evaluations conducted in this phase are formative. The purpose of a formative evaluation is to identify problems. Formative evaluation can involve usability and user experience (UX) testing, both of which identify issues learners and providers might experience when using a product. It also utilizes beta testing to see whether products can be used and intended or as the designer envisioned (commonly called usability testing). The evaluator might use durability, usability, efficacy, safety, or satisfaction as criteria for their judgments. The methods used in these evaluations might include observations, interviews, surveys, and personal experience (trying it out for yourself).

Evaluation in the Implementation and Maintenance Phases

The implementation phase begins once the product is stable and ready to be used by consumers (e.g., instructors and students). Being stable does not mean the product is perfect – it just means it is functional. The product will likely need to be revised and improved through this and the maintenance phase based on additional testing.

Evaluation activities in this phase can be extensive if the developers decide to employ them. Effectiveness evaluation judges the degree to which learners can use the product to accomplish the intended learning outcomes.
Impact evaluation considers what long-term and sustained changes have occurred in the behaviors and abilities of learners — does the learning last, and does it make a difference? Implementation fidelity evaluations judge whether

wild). Testing in the development phase may suggest that users like everything about the product and indicate they would use the product. However, during implementation testing, you may find consumers only use some of the product features (they find some features beneficial but not others). Continued UX testing can also occur during this phase. Testing in this phase may also involve negative case analysis. Rarely will a product work well for all learners. A negative case analysis tells us who uses the product and who does not; it tells us which learners benefit from using the product and which do not.

The Navy added the maintenance phase to their ADDIE design model in recognition of the fact that products age. The maintenance phase is a commitment to continuous improvement of the product through its life cycle and requires ongoing product evaluations similar to those conducted in the development and implementation phase.

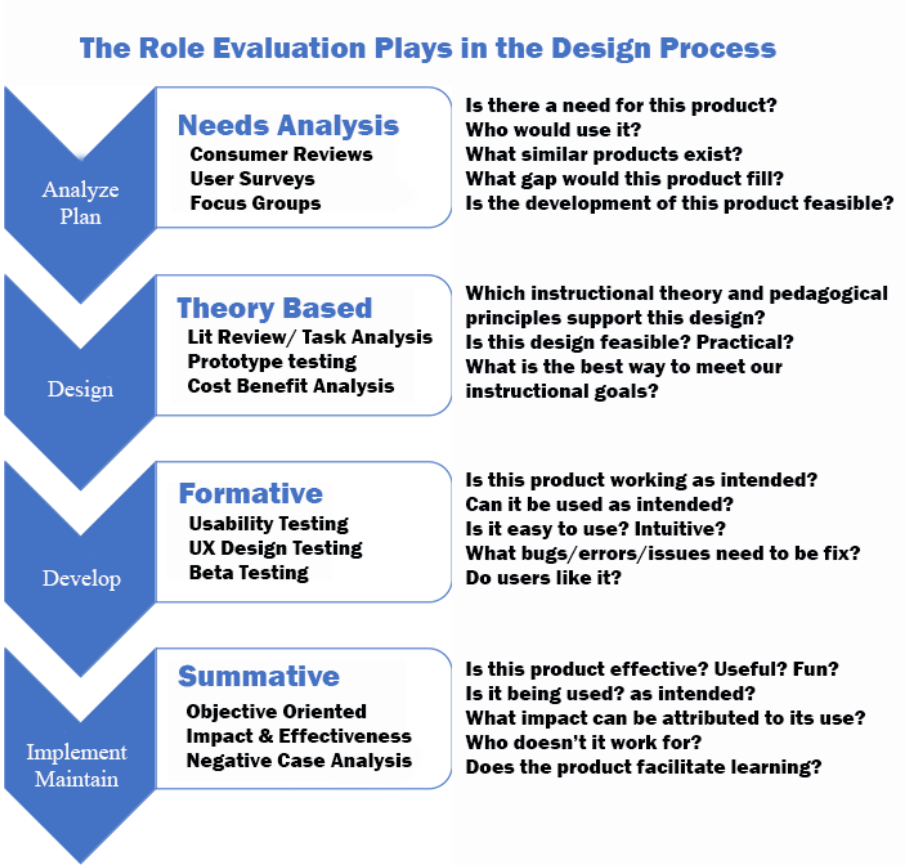


Figure 7: Role of Evaluation and Potential Guiding Questions

- An instructional product might include any educational resource that facilitates or supports learning regardless of the setting or context.
- Evaluation occurs throughout the design process.
- Most design models are adaptations of the ADDIE model.
- Four design phases occur in all design models: Analyze, Design, Develop and Implement.
- Specific types of evaluation are used in each phase of the design process.
- Evaluation is essential to improving the design decisions we make.

Discussion Questions

1. Consider an educational product that you use. What do you like about it? How does it compare to other similar products? Describe something the product lacks that would be nice to have. Describe something missing in the product that users might consider essential. Give reasons why the designer may have decided not to include the missing feature in the product's design.
2. Think about an instructional product people use. Describe the type of person (a persona) who tends to use this product. Think of a label you might use to describe the type of person who uses the product. Suggest reasons why some groups of consumers might use the product and not others?
3. Think about a learning activity instructors use when teaching their class. Why would a teacher believe it's a good learning activity? What pedagogical theory supports its use? Is the activity always effective? If not, why not?

References

Mckenney, S., & Reeves, T.C. (2012) Conducting Educational Design Research. New York, NY: Routledge.

Evaluation Planning

The Basic Evaluation Process.

Evaluations are often needed to make informed decisions. The basic process for any evaluation requires that you:

1. Describe the evaluand (what is being evaluated),
2. Determine the purpose for the evaluation (what do you want to know),
3. Establish criteria for judging the evaluand and the relative importance of each criterion,
4. Determine the best sources and methods for obtaining requisite information, then
5. Collect data, analyze results, and make recommendations

Sounds simple enough. However, the more closely you look at each aspect of the process the more you will appreciate the complexity of planning and carrying out a good evaluation. There are several things to consider.

Evaluator Titles and Jobs

While it is true that everyone conducts informal evaluations all the time, it is also true that those who conduct formal evaluations do not always identify themselves as evaluators.

Most jobs require evaluation skills and abilities; not all job descriptions specifically articulate this as a requirement. People conducting evaluations as a part of their job go by various names and titles. Some of these include – auditor, assessor, analyst, judge, compliance officer, arbitrator, counselor, consultant, director, specialist, manager, supervisor, and advisor. Specific jobs for evaluators in educational settings include – evaluator, teacher, the office of teaching effectiveness and innovation, instructional designer, training and support coordinator, evaluation and research assistant/associate/specialist, director of evaluation and assessment, curriculum specialist/director, program manager/coordinator, assessment/data analyst, and institutional researcher.

Internal and External Evaluators

We should also make a distinction between internal and external evaluators. Internal evaluators, those working within an organization as evaluators, work as full- or part-time employees. When evaluators work as consultants, completing as-needed contract work, they are referred to as external evaluators. There are benefits and disadvantages to both.

When an organization is committed to evaluation and the benefits of continuous improvement, and the organization is large enough to warrant hiring full-time evaluators, it can be cost-effective and efficient to have internal evaluators on staff. Internal evaluators can develop an in-depth understanding of the organization, its purpose, goals, politics, structure, and personnel. They have access to the information and informants they need to do their job and can work on several interrelated evaluation projects (of various sizes and scopes) within the organization. This can be extremely beneficial to an organization. However, internal evaluators can become biased or jaded. They may begin to advocate for specific solutions for political reasons or base findings on unimportant criteria. If internal evaluators do not have some degree of autonomy or do not develop and maintain solid evaluative thinking abilities, professional ethics, appropriate soft skills, and healthy relationships with those they serve, their evaluation efforts can be ineffective and impeded by others in the organization. Above all, evaluators need to be trusted.

evaluation. It can be cost-effective and prudent to hire an external evaluator to provide professional services on an as-needed basis. Sometimes, for political reasons, an evaluation needs to be perceived as unbiased and objective. An external evaluator's reputation as a competent ethical evaluator can provide the organization with results other stakeholders and the general public can trust. There are, however, challenges an external evaluator may face. An external evaluator needs to gain an understanding of the organization's goals and structure. They (or the evaluation team) need to develop relationships, get access to information and informants, understand how an individual project fits into the overall picture and the reasons the evaluation is being commissioned (i.e., needed).

Working with Clients

You will recall that one of the main differences between research and evaluation has to do with the role the principal investigator plays in terms of who is responsible for making final decisions regarding the inquiry's design (i.e., purpose, questions, and methods). Researchers typically work for themselves (i.e., they are in charge) while evaluators work for clients. There are exceptions, for example when a researcher works on an institutional research team or when an evaluator is conducting evaluation research. Still, evaluators typically work for a client as a service provider, on an evaluation team, or as an independent consultant. We use the term client as a broad description referring to the person commissioning an evaluation project. The client may be a supervisor or manager who assigns the project to an internal evaluator or a team of evaluators. The client may also be an employer hiring an external evaluator to conduct a specific evaluation. In either case, the role of the evaluator can be characterized by the involvement they have in the decision-making process and their responsibilities in implementing the evaluation. The relationship between the evaluator's involvement in the project and their role is depicted in Figure 1.

Service Providers. When the evaluator is hired as an external consultant or is the director of a department, and they have an extensive say in determining the purpose, goals, and questions for an evaluation, as well as, the methods utilized to complete the evaluation, they might best be described as a service provider. In these cases, the evaluator (or their team) is typically responsible for the implementation of the evaluation. They will then report back to the client once the evaluation is completed.

Evaluation Partner. An evaluator may work as a partner with a client. This is what Patton (2010) calls developmental evaluation. The evaluator may be working as an internal or external evaluator, but when the evaluator's role is one of partner, the control and the implementation of the evaluation responsibilities are shared between the client and evaluator.

Evaluation Consultants. An evaluator may also be hired solely as a consultant. In these situations, the evaluator lends their expertise to recommend the best way for the client to complete the evaluation. The evaluator may supervise and train individuals as they complete various evaluation activities. The purpose of an evaluator serving in this role is to build capacity and put in place needed evaluation processes so the organization can conduct its own evaluations without the evaluator in the future. Stufflebean (2014) warns however that this type of evaluation may become a pseudo-evaluation if the evaluator is expected to simply sign off on the evaluation as if the evaluation was completed by the evaluator and not the client.

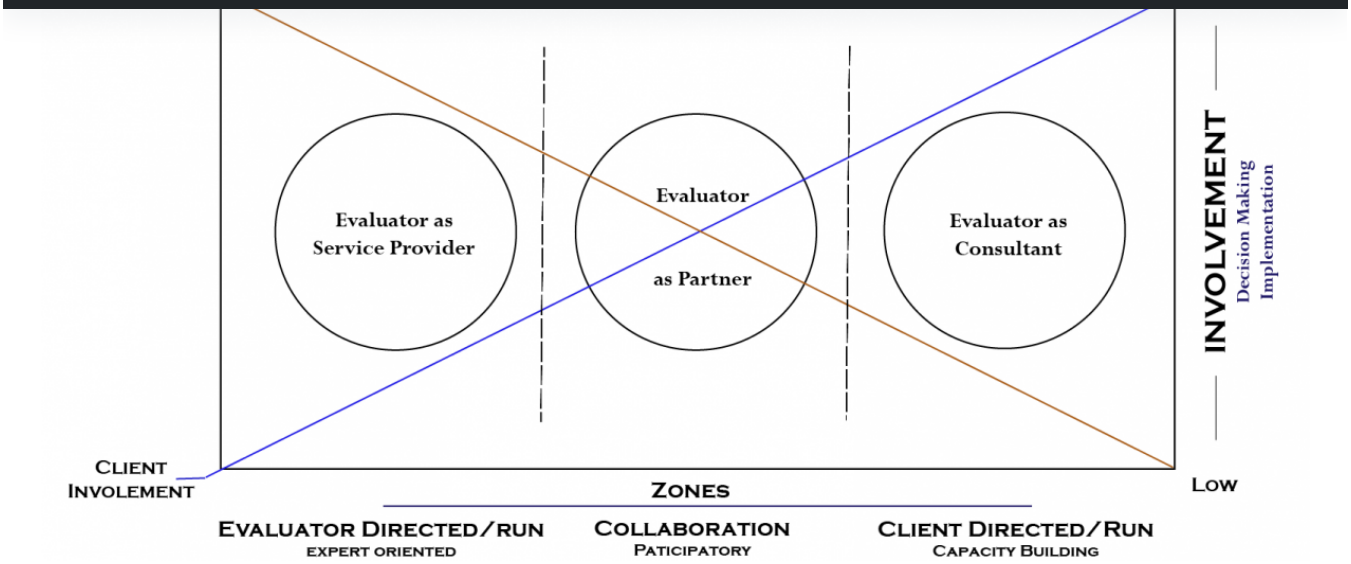


Figure 1: Evaluator Role by Involvement.

> **Evaluations should not be cast into a single mold [scientific].**
>
> **For any evaluation, many good designs can be proposed, but no perfect ones.**
>
> Cronbach (1982)

Making Plans and Proposals

The main difference between a plan and a proposal is the amount of detail required. Depending on who the evaluator is working for, a proposal need not include a detailed evaluation plan. Suppose you are working as an internal evaluator. In that case, your proposal needs to be clear, concise, and detailed enough that your employer/supervisor (i.e., the client) would sign off on your vision for completing the evaluation. However, if the evaluator is competing for an evaluation through a requestion for proposals (RFP) process, the plan may need to be quite detailed as the proposal will form the basis for an evaluation contract. In cases like this, you must persuade the client that your proposal is the best one submitted, and the prospective client should hire you to complete the evaluation. Once a proposal has been accepted, a more detailed plan may be needed to manage the project properly. Project management plans will include specific details regarding the project tasks, data collection, and personnel assignments.

Pre-Planning Activities

Before producing a proposal or plan, it is essential to clarify the client's request. It can take some time to fully understand and clarify the context, conditions, and potential value of conducting an evaluation. You will need to consider several things before deciding to proceed with an evaluation. To properly conduct an evaluation, you will need to fully understand the situation. Asking specific questions can facilitate this process.

- Who is requesting/funding the evaluation?
- What do they want to know? Why do they want to know?
- Who are the decision-makers? Other important stakeholders?
- Are there political considerations that may affect the evaluation?

Often the evaluation context can be complicated. Many educational initiatives and programs are sponsored (funded) by one entity, managed by someone else (often the client), implemented by several individuals (concerned stakeholders), and intended to benefit a diverse group of end-users. Likewise, the reasons for commissioning an evaluation can be varied. Politics is often involved. Sometimes the client has a hidden agenda; they may want the evaluation to justify or support a preferred decision or course of action. Various stakeholders may have different reasons for wanting the evaluation and may intend to use the evaluation results to inform a variety of decisions. To avoid problems, the evaluator must understand the context.

Understand the Evaluand.

Work with the client to fully understand what is to be evaluated. You might ask:

- What is its purpose?
- How does it work? How is it supposed to be used?
- Who are the intended users? Why is it needed? What are the intended outcomes?
- If the evaluand is a program, who is responsible for implementing the program?
- What is the evaluands current stage of production?

Evaluators need to gain a thorough understanding of the things they will be evaluating. Not only will they be required to describe the evaluand, but they will not be able to make valid recommendations without understanding the product and how the designer intended it to be used. Some evaluators use a logic model to accomplish this task. When evaluating an educational product, knowing where the product is currently in the design and development process can also inform the kinds of evaluation needed.

Constraints and Requirements.

- What is the timeframe for completing the evaluation?
- How much funding has been allocated for this project?
- Who will serve as your contact person? With whom will you be working?
- What resources will be available (personnel, equipment, supplies)?
- Will you have access to the information and informants you need?
- Are there any specific requirements and expectations?

It is likely that any evaluation you attempt can be conducted successfully to some extent. We say "likely" and "to some extent" because we do not live in a perfect world; we often lack sufficient resources, and sometimes we lack the desire or ability to complete requisite evaluation activities.

Consider the following common evaluation constraints and issues:

Desire. This can be a constraint when working for a client or with an evaluation research team. You may be willing and interested in completing a rigorous systematic evaluation, but the client (who has commissioned the evaluation), for legitimate reasons, may only be interested in paying for answers to a few specific evaluation questions.

Feasibility. In addition to desire, feasibility can be a constraint. For example, the preferred data collection method may not be possible. When this happens, compromise is required, and, in some cases, the activity may need to be abandoned.

cannot competently complete a specific evaluation task, they will need to recruit (hire) someone to help or be satisfied with the possibility of getting a less than perfect outcome.

Time. This is a prominent issue in evaluation practice. Time is usually of the essence; however, a specific window of opportunity often constrains our ability to capture data. For example, you can only get student achievement data once learners have finished taking their course and the final exam. This means an evaluator may have to wait several months to get the required data needed to complete an evaluation. In addition, some data collection practices take time to complete and are more costly to accomplish (e.g., interviews and observations). An impact evaluation can take months, if not years, to track participants longitudinally. If an evaluation needs to be completed within a specific time frame, evaluators may need to prioritize what can and will be done within that time limit.

Money. This is a constraint because clients (employers) have limited funds. Some evaluators are willing to work for free, but others wish to be compensated for their services (soft costs). In addition, there are always some hard costs associated with conducting an evaluation. These include the cost of travel, facilities, equipment, and supplies. These costs are sometimes forgotten when the evaluator is an internal service provider, but they exist nonetheless.

Access. When you plan evaluations, you might proceed under the assumption that you will have unlimited access to the data sources needed to complete the evaluation properly; this would be a mistake. In practice, evaluators must deal with many restrictions associated with data access. Ethical constraints prohibit evaluators from coercing or forcing individuals to provide information. Not everyone is willing to be a study participant, or you may need to get permission to collect data from potential informants (e.g., vulnerable populations). Sometimes the data you wish to use exists, but those controlling access to the data are unwilling to give it to you because of the regulations they must follow or the cost of providing that information.

Additionally, there are times when the data you need does not exist and cannot be easily obtained. You may lack access to key informants, or it is difficult to obtain accurate measures of the data you need because you do not have access to existing instruments or no suitable measurement instruments exist. At other times, those providing the information may not provide accurate information. For example, early elementary-aged children may not have the meta-cognitive capacity to understand and communicate their motives, likes and dislikes, goals and aspirations. Other participants may be unwilling to answer honestly because they are suspicious of an evaluator's motives; they may also provide skewed information for personal reasons. For example, they may (unintentionally or otherwise) downplay flaws in a product they value. Gaining access to and obtaining valid, relevant data can be one of the most significant limitations of any evaluation.

Not all evaluations can or should be conducted

Evaluability Assessment.

Part of any evaluability assessment involves understanding the evaluand's current stage of development and circumstances surrounding the product's use. The other part involves determining if an evaluation is warranted. Not all evaluations are worth doing. There are several situations where careful consideration should be taken to decide whether conducting a formal evaluation is sensible.

studied extensively, and its value has already been established. This can also happen when it is evident that research concerning the pedagogical theories and design principles underpinning the product does not support the product's development or its implementation. It may be hard to justify re-evaluating a product in such cases. There are exceptions as the results of many research studies cannot be replicated consistently. An evaluation of well-studied products may produce new information when circumstances and conditions change substantially over time. For example, a considerable amount of evaluation research was conducted to determine the potential efficacy and effectiveness of online learning. Much of this research concluded that online learning was less effective than in-person learning due to inadequate pedagogies and the lack of technology needed to successfully implement online learning. However, conditions and circumstances have changed considerably from when this research was initially conducted—improvements in online technologies and increased access to technologies by learners in many locations have made online learning a viable and effective alternative to in-person learning. In cases like this, changing circumstances may make conducting new evaluations reasonable.

2. **Trivial results are expected**. The value of conducting an evaluation might be suspect if the purpose and scope of the evaluation are insufficient or limited in some significant way. For example, if a proposed effectiveness study to determine whether a product or program facilitated student achievement was needed, but the evidence to be collected was only to consider parents' perceptions of their student's learning. Results, in this case, represent a trivial facet of effectiveness and would likely be misleading. Without a more robust and systematic data collection process, an evaluation of this type might not be beneficial.
3. **The product is not needed**. An evaluation may not be necessary if other similar popular products exist and there is no interest in the product being proposed as the evaluand. When few users adopt a product, it may be difficult to conduct some kinds of evaluation (e.g, effectiveness and impact). Not only that, depending on how similar the product is to other already tested products, results for a usability evaluation may be somewhat inconsequential. A marketing evaluation might benefit a designer to know why people choose other products, but an extensive evaluation of the product may be futile given the lack of interest in the product and its similarity to other existing products.
4. **The product cannot be used.** In some situations, you may find that users do not use a product not because they don't want to but because they cannot. For example, in many places throughout the world, learners do not have adequate access to the internet. It would be unwise to evaluate a product that depends on internet access in these areas.
5. **Results are unlikely to be used.** Many educational programs have a limited lifespan. For various reasons, stakeholders may decide to discontinue an initiative regardless of any results an evaluation might provide. Sometimes effective programs are discontinued because the person championing the program leaves, the cost of implementation becomes too great, or the funding for the program runs out. When there are no plans to continue an initiative for one of these reasons, conducting an evaluation may be unimportant. In addition, you may not be surprised to hear that some evaluations are commissioned for political reasons. At times a client commissions an evaluation to avoid making a decision or to make it look like something is being done while all along having no intention of using the evaluation results.
6. **There is nothing to evaluate**. In some cases, certain kinds of evaluation cannot be conducted because the product is only a concept or is still in development. If the product is likely to undergo substantial changes during the design and development phases, effectiveness evaluation and impact testing may need to be postponed until the product is stable enough to be evaluated.

There are certainly many reasons why an evaluation might not be warranted; those presented here are but a few. However, once you determine that an evaluation is viable and needed, you can now start planning the evaluation and developing a proposal.

Description of the evaluand.

An evaluation proposal (or plan) often starts with a description of (or introduction to) the evaluand, the client, end-users, and important stakeholders (i.e., those directly associated with the implementation of the product or with a vested interest in the product's use). Information gathered from the pre-planning phase will be indispensable in accomplishing this task.

For evaluation in an educational setting, the evaluator needs to describe the purpose and function of the instructional product, how the designer intended it to be used and who the intended users are. It may include an explanation of the theory and principles supporting the product's design and any contextual aspects relevant to the product's use. If the product has not already been developed and implemented, the introduction section of the proposal should explain the product's current status in the design and development process. Often, evaluators are asked to serve as external evaluators for a funded development project. In these cases, the evaluator may propose various evaluation activities appropriate for each stage of the product's development. For example, it is typical for a developmental evaluation to focus on formative evaluation activities during the project's initial months (or years), then switch to summative evaluation activities once the product's development has been completed and the product has been implemented in its final form. This may include efforts to establish the effectiveness and impact of the product. It may also include a negative-case evaluation to determine which users benefit from the product and which do not.

Write a purpose statement.

A purpose statement describes the reason (or need) for the evaluation. This statement will explain the decisions a client needs to make and how evaluation results will be used to inform these decisions.

Sometimes the purpose for an evaluation is provided for you by the client. Other times, all the client knows is that they need or want an evaluation of a specific product. Some clients may be asking for the evaluator's expertise to help them determine a valid and viable purpose for the evaluation. Other times the evaluator needs to suggest utilizing various types of evaluations to answer questions they may not have considered. For example, it is not uncommon for clients or interested stakeholders to ask for a summative evaluation of the effectiveness and impact of an educational product. This may sound reasonable, but if the evaluand has not yet been developed or has yet to be implemented, these types of evaluation may not be possible. Likewise, it is often the case that an evaluator needs to suggest alternative or additional reasons and purposes for evaluating a product. For example, while an objectives-oriented evaluation of the product's effectiveness might be beneficial, an implementation evaluation may be needed as well. An implementation evaluation provides evidence that an educational product can be and is being used as intended or that a program is being implemented properly. Without this, clients, decision-makers, and other interested stakeholders will often misinterpret and possibly misuse the results of an effectiveness evaluation. Misunderstanding results can adversely affect the decisions people make.

When deciding on a purpose for an evaluation, you need to consider all the possibilities then establish an appropriate yet feasible reason for conducting the evaluation. The purpose you eventually decide on will be influenced by context, situation, and existing constraints. However, initially, you should ignore constraints and brainstorm ideas as if anything were possible (divergent thinking). Think about everything you would like to know about the product you plan to evaluate. Consider what you would need to do to answer these questions. After identifying all the possibilities, determine which are essential (need to know objectives) and which are less critical (nice to know objectives). Consider constraints to narrow the evaluation purposes and scope (convergent thinking). The purpose you ultimately decide on will often represent a compromise between what you would like to do and what is possible. The purpose of any evaluation should represent an important reason or need.

List Evaluation Questions and Specify Criteria

An evaluation proposal should list specific evaluation questions needed to inform the decisions proposed in the purpose statement. Your proposal should also include a description of the criteria you will use to judge the merit and

might ask how successfully a product facilitates students' achievement of specific expected learning objectives. A secondary question (i.e., an important but not essential question) might ask in what ways the product might be improved to increase its utility. The effectiveness criterion is primary, and the usability criterion is secondary.

Types of standards/criteria

You will recall that a standard is a generally accepted set of criteria used to make a judgment. In contrast, a criterion may represent a personal value held by individuals but not generally agreed on as essential or even necessary. Criteria can be classified as rules and requirements, a scoring structure, or specific principles and attributes individuals value. Criteria commonly used to judge educational products include utility, usability, feasibility (viability), availability, cost, effectiveness, efficiency, impact, satisfaction, preference, desirability, relevance, coherence, social acceptability, safety, and sustainability, to name a few. Evaluators may use some or all of these criteria to judge the merit and worth of an educational product or program. However, some of these will be more important than others to decision-makers. For example, decision-makers may value effectiveness and efficiency over user preference and product appeal. Indeed, a product that works is better than one that does not, even if it looks good. Still, individuals may not use a product they find unappealing even if it is more effective than other more attractive products. So while it may be unimportant to decision-makers that users like the look of a product, it still could be an important criterion for evaluating a product – or not.

Proposing Data Collection Methods

Once evaluation questions have been established, an evaluator must determine what data is needed to answer the evaluation questions. They must then propose sensible data collection procedures that could be used to capture the required data. Proposed methods will include identifying the sources (e.g., the people, existing data, documents) from which the data will be retrieved. It will also describe the procedures required to obtain the data and how the evaluator will analyze the data. Precise detail need not be included in a proposal. Still, enough specificity should be provided to convince the client that the data collection efforts will likely produce enough relevant information to answer each of the evaluation questions. For example, it would be insufficient for a proposal to only say that interviews and surveys will be used to collect data. The proposed methods should include some details regarding who will be interviewed and surveyed, what instruments will be used, how data collection instruments will be developed, and how the data obtained will be analyzed and used to answer the evaluation questions.

Determining a Budget

Developing a budget can be tricky, especially when the evaluator is an external evaluator responding to a request for proposals. Budgeting too little will mean you may be doing the evaluation for free, or worse, you may have to pay out of pocket to complete the evaluation you contracted to do. Asking too much may diminish your chances of obtaining work now or in the future. Understanding actual costs and how much you are willing to charge for your services is essential to the budgeting process. However, budgeting issues can be alleviated to some extent by asking clients upfront what they have budgeted for the evaluation. Knowing beforehand any budget restrictions can allow the evaluator to plan activities and limit the scope of the evaluation accordingly. Giving a client the option to include or exclude certain costly activities they are unwilling or unable to pay for can also help.

As mentioned previously, there are soft costs and hard costs that need to be considered. Soft costs represent the evaluator's time, and hard costs represent the cost of travel, facilities, equipment, and supplies. These costs are sometimes forgotten when the evaluator is an internal service provider, but they exist nonetheless. When working as an internal evaluator, budgeting usually entails specifying the personnel required and estimating the time needed to accomplish each task.

Often external evaluators have to submit a proposal and hope for the best. However, the evaluator can sometimes negotiate terms and requirements with the client beforehand or after submitting a draft proposal. Working collaboratively with the client on the proposal's details can be a very productive way to develop an evaluation plan. Once a proposal is approved, a more detailed management plan will likely be needed.

Chapter Summary

- Evaluations are conducted within the context of a specific situation. The purpose and methods may vary as time, available resources, and politics constrain what can be done.
- Evaluation plans and proposals are similar, but a proposal can be less detailed. Once a proposal is accepted, a more detailed management plan will be needed.
- Plans are essential but often need to be adapted to account for changing circumstances and conditions.
- Pre-planning activities are necessary to properly understand the context, conditions, and potential value for conducting an evaluation.
- An evaluator needs to carefully consider whether an evaluation can or should be done.
- A proposal needs to include a description of the evaluand, a purpose statement, proposed methods, a timeline, and a budget.
- Establishing an evaluation's purpose requires divergent and convergent thinking to determine an achievable yet meaningful objective for the evaluation.
- A purpose statement describes the reason an evaluation was commissioned. It explains the decisions that need to be made and how evaluation results will be used to inform these decisions.
- Proposed methods should be viable and likely to produce sufficient data to answer the evaluation questions.
- Timelines and budgets are needed to ensure the planned activities can be accomplished within the timeframe specified for the evaluation and within any existing funding constraints.
- Pseudo-evaluations should be avoided as they are conducted to promote a specific predetermined solution. These include politically-inspired and advocacy-based evaluations.
- Quasi-evaluations provide good information, but the value of the findings is limited in some way. Evaluation classified as quasi-evaluations could be improved by expanding the scope of the evaluation and the criteria used to determine merit and worth.

1. Provide an example of an evaluation you would likely choose to decline. Give reasons why you would be hesitant to take on such an evaluation.
2. Suppose a client asks you to help them with an evaluation project. Describe the benefits and disadvantages of working as a service provider, evaluation partner, or consultant. Which would you prefer and why?
3. Think of a specific educational product you have used or would like to use. For that specific product, describe any questions that need answering before you could develop an evaluation proposal for that product. Provide possible answers to these questions.
4. Imagine a fictitious client asks you to help them evaluate a product they have developed. Using a divergent thinking process, list various types of evaluation the client might employ and the benefit of each in terms of decision-oriented questions that could be answered. Using a convergent thinking process, consider potential constraints that would limit the viability of each of these evaluation options.

References

Cronbach, L. J., & Shapiro, K. (1982). *Designing evaluations of educational and social programs*. Jossey-Bass,.

Patton, M. Q. (2010). *Developmental evaluation: Applying complexity concepts to enhance innovation and use*. Guilford press.

Stufflebeam, D. L., & Coryn, C. L. (2014). *Evaluation theory, models, and applications* (Vol. 50). John Wiley & Sons.

Research Principles Primer

The following chapter presents research principles and ideas relevant to evaluation. You will not be surprised to find that people do not always share a common understanding of words and ideas. A shared understanding of terms is needed as many concepts are understood and used differently in various fields and disciplines. A brief discussion of a few crucial concepts is provided here as a foundation for understanding the evaluation principles, approaches, and methods presented later.

Inquiry Methods & Paradigms

Research Paradigms are deeply held beliefs and assumptions that guide the way we do research. The reason people do research is to learn some truth about the world (i.e., make sense of reality). However, individual researchers often make their inquiries in very different ways. This is because they have different beliefs about the nature of reality and truth (ontology). They also hold different opinions about how best to gain understanding and obtain the truths they seek (epistemology). A person's ontological and epistemological perspectives form the research paradigms they hold. This affects how they explore the world, the information they seek (methodology), and what evidence they will trust when making decisions.

Ontology has to do with the nature of reality and truth. Three common ontological viewpoints include the belief that:

- there is one reality (absolute truth, one correct interpretation),
- there are multiple realities (individual truths, perceptions), or
- truth is relative (context-dependent, debatable, and interpreted).

Many theorists view realism (absolute truth) and relativism (contextual with multiple realities) as opposing extreme positions on an ontological continuum. They often consider the idea of multiple realities to be a subset of constructivism since both are personal interpretations of reality that will likely change after reflection and dialogue with others.

Relativists believe that a person's interpretation of reality is context-dependent. Individuals can perceive reality a certain way given a specific set of circumstances and have a completely different understanding of what is true in a different situation. Although different, the person's perception of reality would be accurate (true for that individual) in both situations. Therefore, truth is not absolute. In addition, two individuals can have very different interpretations of the same situation. These conflicting versions of the reality (what is true) are accurate for each person—they represent what each individual believes (how they feel), and that should be respected as their truth.

Realists accept that two conflicting views of reality can exist, but they consider multiple realities like they would competing hypotheses—an individual can believe two different things to be true, but the evidence will support one and not the other; both cannot be accurate; although, both may be inaccurate. For a realist, the goal of research is to obtain a correct and complete understanding of reality that thoroughly explains and can be used to predict outcomes and behaviors.

In practice, realism (i.e., the belief that there is absolute truth) is prevalent in the hard sciences (e.g., physical sciences like biology, physics, chemistry). Relativism is more prominent in the soft sciences (e.g., social and political sciences). The physical sciences tend to have identifiable rules that govern things in the real world and can be used to predict

identify and apply. Social contexts are complicated because people are complex and frequently unpredictable. In education, for example, we are often unable to control for confounding variables (e.g., agency), making the results difficult to generalize and often impossible to replicate. Social theories are often developed conceptually, under ideal conditions or controlled environments. In practice, classroom conditions are not always controllable, and students are unpredictable. A teacher may find that what worked today will not work tomorrow. Often the best researchers can do is identify promising best practices and hope for the best.

Epistemology deals with how knowledge is acquired and how we can come to know what we know. It describes how we find truth and understanding—how we make sense of the world we live in and our interactions with others. There are a few generally accepted ways an individual can come to know or believe something:

- **Empiricism** (Physical Senses),
- **Rationalism** (Cognition, Reason, Logic, Analysis),
- **Intuition** (Emotions, Feelings, Conscience), and
- **Authoritarianism** (experts sharing/communicating truths).

You will note that these ways of knowing are not exclusively or independently applied. We often empirically experience something through our physical senses and then use authoritarianism, reason, and intuition to build understanding. (see Piaget's schemas, accommodation and assimilation).

Figure 1: Methods Paradigm Foundations (source Kimmons, 2022)

Common Research Paradigms

There are three widely recognized research paradigms: positivism, pragmatism (or post-positivism), and constructivism. Other paradigms have been identified, but these tend to align with one of the three main paradigms and differ only in how they situate themselves within a specific research field. For example, a critical theory paradigm is a form of constructivism that situates its research in social justice issues and seeks to address political, social, and economic issues.

exists and that reality can be understood objectively. They argue that perceived truth is not the actual truth regardless of how strongly a person believes their perspective is accurate. They rely primarily on empiricism as an epistemology but tend not to trust subjective interpretations of an observation preferring to use measurement to verify observations rationally using statistics. Fundamental positivism rejects the notion that intuition and personal values should be used to determine truth. They stick to what we can observe and measure, not what we think and feel. From a positivist viewpoint, the purpose of research is to establish generalizable cause and effect relationships. A few common ways positivists gather, process and analyze information include:

- **Statistical Empiricism** – this approach to knowing relies on mathematical verification of what we observe; it requires researchers to quantify observations so they can be analyzed statistically.
- **Scientific method** – the scientific method is based on rationalism and is used to test a theory or hypothesis in a structured, logical way. The scientific method is a deductive approach. Results of experimentation that use this process either support a proposed theory (understanding of reality) or not. It relies on replication.
- **Authoritarianism and Scholarship** – Learning from experts is acceptable in the positivist paradigm if the expert's knowledge was obtained using appropriate methods (e.g., see for example what works clearinghouse). Scholars gain an enhanced understanding of reality by reviewing research that, in their perspective, was of high quality.

Many of the methods used by positivists are quantitative.

- **Experimentation** – deductive hypothesis testing is a standard positivist method for building knowledge. There are many experimental forms, but the gold standard is a random controlled trial. From a positivist paradigm, it is best to verify results mathematically (i.e., statistical empiricism); thus, most experiments would include a measure of some dependent variable and measures of independent variables thought to influence the dependent variable.
- **Correlational studies** – establishing the existence of a correlation (relationship) between two variables is one of the first steps to determining causation.
- **Structured data collection**. – surveys, interviews, and observations that ask all participants identical questions. These bounded sets of questions are designed to produce accurate and reliable measures of a specific variable (e.g., behavior or attitude).
- **Tests** – assessment instruments designed to capture objective measures of knowledge and ability.

The Constructivist Paradigm. Unlike positivists, constructivists believe that reality (truth) cannot be determined statistically. Constructivism, or Interpretivism, is based on the idea that individuals must build (i.e., construct) knowledge personally and that knowledge is based on a personal interpretation of one's lived experience. This paradigm originates from a relativist ontology. It advances the idea that our understanding of reality (truth) does not exist as an absolute but is a personal perspective and is defined contextually by each individual independently. It involves an inductive process. However, fundamental constructivism rejects the idea that people can be totally objective. A person's effort to process information and gain understanding will always be subject to bias, personal values, goals, and preferences. While constructivists accept empirical evidence, they would not feel compelled to verify truths statistically. Instead, they would describe it qualitatively. They accept intuition and rationalism as valid ways of knowing and do not worry that individuals arrive at different conclusions. Multiple versions of reality (personal beliefs and feelings) are accurate for that individual and should be respected as their truth. The best way to establish a collective truth is through dialogue, negotiation, and persuasion. The goal of research isn't necessarily to predict outcomes and behavior but to understand them. Methods constructivist use often seem similar to positivist, but they are applied differently. Some common ways constructionists gather, process, and analyze information include:

and extant data.
- **Authoritarianism** – similar to positivism, this way of knowing involves learning from experts—those who have insights and understanding others do not. However, expert sources might include academic researchers, instructors, experienced practitioners, regular people, or God (information revealed by a superior being).

Many of the methods used by constructivists are qualitative.

- **Naturalist Inquiry** – is accomplished when a researcher uses a specific perspective or cultural context to observe, describe, and interpret the experiences and actions of people.
- **Hermeneutic Phenomenology** – an interpretation of people's lived experiences using the first-person point of view.
- **Case studies**– an in-depth study of a person, group, or event designed to explore every aspect of the subject's life in order to understand their behavior.
- **Grounded theory**–is an inductive process that gathers, synthesizes, analyzes, and conceptualizes qualitative observational data to develop theories (explanations of how the world works).
- **Ethnography**– involves a description of the customs of individual peoples and cultures. This often requires the researcher to completely immerse themselves in the culture and everyday life of the people the researcher is studying.

The Pragmatist Paradigm. Pragmatism came about primarily as a result of the *paradigm wars* (or Quant-Qual wars) of the 1980s (Guba & Lincoln, 2005). Pragmatism, also referred to by some as post-positivism, began among scholars who argued that a mono-paradigmatic orientation of research was not adequate. Understanding (truth) could not be established entirely using statistics and random controlled trials, nor could it be altogether revealed using subjective descriptions of an individual's perceptions. The pragmatist paradigm promotes a practical pluralistic approach to research that uses the most appropriate methods for studying the phenomenon. This gave rise to a belief that advocates for mixed methods and an appreciation for the value-laden complexity of reality (axiology). Pragmatists use a variety of research methods depending on the purpose of the research and the questions being asked. However, their perspective of what constitutes quality research is not based on using a particular method but on the appropriateness, credibility, and trustworthiness of the method's application. In addition to any of the other previously mentioned methods used to gather, process, and analyze information, pragmatists might also use:

- **Action Research** – Action research is not a method but rather an approach. It differs from other research in that there is less concern for the universality of findings and more concerned about whether the findings can be used to solve a particular problem or answer a specific question. When a problem is identified, an action research protocol would consider various solutions and test the most promising to see if it solved the problem at hand.

Table 1: Common Paradigms

Paradigm	Typical Stance
Positivism	Absolute truth exists. Reality is objective, knowable, and generalizable. Truth is discovered by experimenting and testing theories (i.e., deductive, scientific, and statistically verifiable
Pragmatism	Some absolute truths exist (e.g., in the physical sciences); others are individually or collectively perceived (e.g., social sciences). Not all explanations of reality can be understood perfectly or mathematically. The best way of knowing will depend on the question being asked (methodological pluralism, mixed methods).
Constructionism	Truth is context-dependent and varies (personal perceptions of truths, multiple realities). Collective reality (truth) is ever-changing, socially constructed, and represents a negotiate

Quantitative and Qualitative data

Research is often described as either quantitative or qualitative. When people say this, they are describing the type of data being collected.

Quantitative research collects information that can be represented as a numerical value and analyzed statistically. Quantitative data (i.e., numbers) describes some attribute of an object or person (i.e., amount or magnitude). These numbers (i.e., measures) are used to communicate and compare qualities a researcher is interested in studying. For example, we could say someone is tall; this is a qualitative description of a personal attribute. We quantify this attribute by measuring the person's height (i.e., assign a precise numerical value that represents a person's height). This can then be used to communicate the precise height of an individual or compare the height of one person with another.

Many things can be quantified, but not all information can be easily reduced to a number. Information that cannot be quantified is referred to as qualitative data. Qualatative data is non-numerical information. It cannot be analyzed statistically—it is described, categorized, and interpreted. Examples of qualitative data include personal accounts, photographs, documents, and video recordings. Qualitative information may also represent a person's emotional state, reasoning, or beliefs. A particular attribute's frequency and magnitude (e.g., strength) can be quantified but not the various ideas they represent. For example, a researcher may ask people to explain why they behaved a certain way. A given reason cannot be quantified (e.g., I was angry, afraid, interested, determined, concerned, or felt responsible). These are nominal data points (i.e., labels without order or incremental value). Assigning a number to a nominal data point would have no meaning. However, the frequency of a reason being given can be calculated, and the magnitude (i.e., strength) of a particular feeling represented by a given reason can be quantified (e.g., how angry were you?).

Objective and Subjective interpretations

Positivists tend to value quantitative data because it is objective and can be statistically analyzed. While constructivists often collect qualitative data which needs to be interpreted subjectively (i.e., using values-based criteria). Researchers with a positivist paradigm may collect qualitative data but attempt to quantify it for analysis.

Many people mistakenly assume that objective analysis is better than subjective analysis. However, both are valid and often necessary ways of interpreting data. To be objective simply means that people tend to agree on the meaning of the data. A number has a specific meaning because people agree on that meaning. Using numbers to communicate can be more effective in conveying a precise meaning. For example, you may find that "*several*" people indicated they held a specific belief, but saying it in this way (i.e., qualitatively) does not communicate that finding very well. However, to say 43% of participants held a specific opinion (e.g., agreed or strongly agreed) would be a more precise (i.e., objective) way to communicate this finding. Still, people do not always agree on the interpretation of research findings. For example, a p-value is a number that represents the probability (i.e., likelihood) that an observed phenomenon happened by chance. A p-value is a quantitative piece of information. It is objective because people tend to agree on its meaning. However, a researcher must interpret a p-value to determine its statistical and practical significance, which requires a subjective analysis. Statistical significance requires we set an alpha value, usually .05 but not always; it depends on the context of the research. Even when a statistically significant result is determined (i.e., the p-value is less than the alpha value), the researcher must subjectively consider the finding's practical significance. This may be facilitated by calculating an effect size, but the practical significance of a finding is most often a subjective determination. For example, an evaluator might calculate the change in attendance rates attributed to a specific intervention. The resulting difference may be statistically significant (i.e., not likely to have happened by chance). The effect size may be adequate (i.e., accounts for a fair amount of the variance, and the difference in standard deviation units is acceptable). Still, those running the program may believe the magnitude of the change has no practical value (e.g., a statistically significant increase in attendance may represent an insignificant number of students).

subjective interpretations (i.e., evaluative judgments) are often necessary.

Inductive and Deductive Reasoning

Inductive reasoning, or **induction**, involves making an inference based on observation. Induction is often used to develop theory. Deductive reasoning, or **deduction**, involves making inferences based on widely-accepted facts or theories (i.e., premises). Deduction is often used to test theories. Both these forms of reasoning involve logic, reflection, and interpretation. Both are used in evaluation and are aspects of evaluative thinking. Evaluators use an inductive process to describe situations and make recommendations. They might use a deductive process to judge the effectiveness and impact of an initiative.

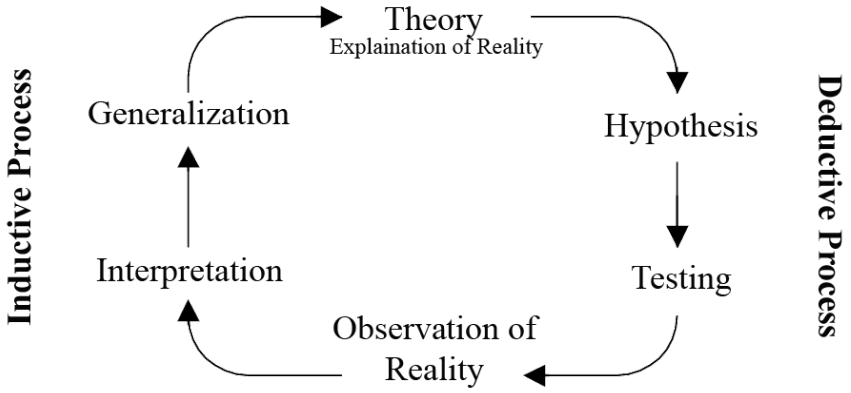


Figure 2: Theory Development using Inductive and Deductive Logic

Axiology

Axiology is the study of values and value judgments. Axiology is concerned with issues of ethics, morality, and aesthetics (personal preference and satisfaction), as well as other forms of value. Axiology is vital for evaluators because our values affect our judgments of what has merit, worth, or is valuable. The value (merit or worth) assigned by individuals to an object will differ depending on their values (morals, preferences, interests, goals, ethics).

The distinction between intrinsic and extrinsic value is central to axiology. Things with intrinsic value are valuable in and of themselves (e.g., they are life-sustaining or essential to our well-being, see Mazlow). Objects and people have extrinsic value because they are useful or desirable to someone for some reason. For evaluators, understanding our own values helps us to make good evaluations; understanding what others value helps us understand the decisions they make.

Personal benefit and the greater good. One particular axiological issue evaluators must consider is the tension between valuing what is best for the greater good and what is best for an individual (i.e., personal interest and benefit). For example, results from a negative-case analysis might find that most individuals participating in a program achieve the initiative's intended goals (i.e., they benefit from participating in the program). However, this may not be the case for all participants. This becomes a problem when the only way to increase the benefit for the few would mean diminishing the benefit for the many.

Evaluators are encouraged to “strive to contribute to the common good and advancement of an equitable and just society” (AEA guiding principles, 2018). However, it would be incorrect to believe that there is one overriding value or belief that describes society's best interest (Davies, 2021; Patton, 1985; Schwandt, 2002). Thus, the issue remains

Chapter Summary

- Research paradigms are deeply held beliefs and assumptions that guide the way we do research and evaluation.
- A person's ontological and epistemological leanings will influence their research paradigm.
- There are three widely accepted research paradigms: positivism, pragmatism, and constructivism.
- An individual's research paradigm impacts the types of data they value and the methods they use.
- All data begins as qualitative information; some data can be quantified (i.e., described using numbers).
- We consider data to be objective when people generally agree on its meaning.
- Quantitative data is described as objective because it deals with numeric data.
- Qualitative data is described as subjective because it needs to be interpreted.
- Both objective and subjective interpretations are valid and needed in research and evaluation.
- Inductive reasoning involves making inferences based on observation and is often used to develop theory.
- Deductive reasoning involves making inferences based on widely-accepted facts or premises and is often used to test theories.
- Axiology is the study of values and value judgments, a fundamental underpinning of evaluation.
- One important issue evaluators must consider is whether they should focus their criteria for judging merit and worth on individual benefit or the greater good.

Discussion Questions

1. Describe your own research paradigm. Include your preference for quantitative (objective) or qualitative (subjective) data. Explain with examples how this might affect the way you evaluate.
2. Hedonistic theories attempt to explain what motivates us to behave in the way that we do. A basic premise of hedonism suggests that people will do things that increase pleasure and decrease pain. Using a deductive process, describe a way you might test this theory. Using an inductive process, suggest a revision to this theory by giving an example from your experience that might explain (describe) how values and a person's belief about what is good for us might contradict this theory.

References

American Evaluation Association. (2018). AEA-American Evaluation Association: Guiding Principles for Evaluators. [online] Eval. org. https://edtechbooks.org/-pSNL

Davies, R. (2021). Establishing and Developing Professional Evaluator Dispositions. Canadian Journal of Program Evaluation, 35 (3).

Guba, E. G., & Lincoln, Y. S. (2005). Paradigmatic Controversies, Contradictions, and Emerging Confluences. In N. K. Denzin & Y. S. Lincoln (Eds.), *The Sage handbook of qualitative research* (pp. 191–215). Sage Publications Ltd.

Patton, M. Q. (1985). Cross-cultural nongeneralizations. In M. Q. Patton (Ed.), Culture and evaluation: New directions for program evaluation (vol. 25). Jossey-Bass.

An alternative resource for research basics can be found at Educational Research.

Evaluation Approaches for Designers

When describing how people should conduct an evaluation, you will find that most suggestions can be classified as general principles or approaches, with a few being classified as models. General principles describe things evaluators need to do so the results of an evaluation are deemed credible. For example, clearly defining the purpose for an evaluation and the criteria you will use to judge merit and worth is a general principle all evaluators should follow. The evaluation principles someone might provide will not tell you what the purpose should be nor which criteria are most important—just that you need to state the purpose and criteria clearly. Likewise, suggestions for how to conduct an evaluation might be classified as an approach (not a model) because they do not prescribe specific methods—they suggest best practices given a potential evaluation purpose or need.

For example, a model is an example of how a thing should look—those using a model attempt to replicate the model precisely. Evaluation models are prescriptive; they will provide detailed steps an evaluator or researcher must take if they claim to be using a specific model or design (e.g., random controlled trials). However, an evaluator will rarely conduct an evaluation the same way twice. Evaluators may use the steps of a recommended model as guidelines but will not attempt to replicate the procedure. The purpose, goals, context, and constraints of an evaluation will require the evaluator to adapt and revise any proposed model. The goal is not to replicate but to credibly adapt and approximate the model's proposed design.

Whatever you prefer to call them, the next few chapters present descriptions of various approaches and models you might use when conducting an evaluation at various phases of the design and development process.

Pseudo and Quasi-evaluations

Before presenting any of the commonly used evaluation approaches, you should be aware of two situations that often affect the quality of the evaluation results we obtain and thus the decisions we make. Stufflebeam & Coryn, (2014) refers to two types of evaluations we should either avoid or take steps to improve: Pseudo-evaluation and Quasi-evaluation. Any of the approaches described in this chapter have the potential to become a pseudo- or quasi-evaluation. An evaluation may seem well designed but might be compromised in some way.

Pseudo-evaluation

Pseudo-evaluations are flawed mainly because the evaluation is conducted in such a way as to confirm a predetermined outcome. Some pseudo-evaluations are founded on ill-intent; others are inadvertently compromised by stakeholders or limited due to unavoidable constraints that restrict our ability to conduct a proper evaluation. Either way, these should be avoided.

Some examples of the ways an evaluation might become a pseudo-evaluations include:

for example, a success-case evaluation and only collected positive reviews to make the product look better than it is.

- **Politically Mandated Evaluations** – Often, an evaluation may be commissioned for political purposes. There may be a legitimate reason for doing the evaluation, and the evaluators may have every intention of conducting a credible evaluation. Still, if the evaluators are denied access to essential information or only use information that leads to a specific recommendation, the evaluation becomes a pseudo-evaluation. A politically mandated evaluation may also be classified as a pseudo-evaluation when those commissioning the evaluation wish to avoid making a decision, or they wish to make it look like something is being done while all along having no intention of using the evaluation results.
- **Advocacy-based and Pandering Evaluations** – Advocacy-based evaluations often become a form of pandering—working toward a predetermined outcome. An evaluation becomes a pseudo-evaluation when the evaluator caters to the client's desire for a result that supports a specific recommendation. So, while advocacy for an important cause may be admirable, advocacy-based evaluation often fits the definition of a pseudo evaluation because the evaluator, by definition, is promoting a particular viewpoint and pushing for a predetermined set of recommendations.
- **Empowerment Evaluation** – Building evaluation capacity can be a legitimate objective for an evaluator. An evaluator serving as a consultant may wish to help the company build evaluation capacity and put in place needed processes so the organization can conduct its own evaluations without the evaluator in the future. Stufflebean and Coryn (2014) warns, however, that this type of evaluation may become a pseudo-evaluation if the evaluator is expected to simply sign off on the evaluation as if the evaluation was completed by the evaluator and not the client.

Quasi-evaluations

In contrast to pseudo-evaluation, quasi-evaluations have less value because they are incomplete or limited in some way by the scope of the evaluation's purpose, the types and sources of data collected, or the criteria used to determine the merit and worth of the evaluand. For every evaluation, an evaluator could ask several different questions and use a variety of criteria. Quasi-evaluations can be beneficial to evaluators, but they often do not provide a complete picture and thus could be improved.

Some examples of the ways an evaluation might become a quasi-evaluations include:

- **Limited Data** – Some quasi-evaluations are limited because the evaluator has limited or no access to the information they need to conduct a proper evaluation. Ethical issues may constrain the evaluator; the data may not exist or cannot be obtained directly. When compromises are made, the results may be limited. For instance, an evaluator may not have access to key informants (e.g., young children); so, as a compromise, they ask someone associated with the key informants (e.g., a child's parent). Getting information from an indirect source may provide some information but likely is not as good as a direct source.
- **Narrow Set of Evaluation Questions or Criteria** – Experimental studies and objectives-oriented evaluations are often described as quasi-evaluation because they tend to answer a limited set of questions (usually only one) and base the evaluation on a single or a limited number of criteria. For example, an assessment (test) is often used to obtain information regarding the degree to which students have accomplished the learning objective for a course. If this were the only data collected and the criteria for judging the quality of the course was based solely on student achievement, the evaluation likely could be improved. There are many things an evaluator might consider when judging the quality of a course or the instructor (the ghost in the system); student grades are but one of those things.
- **Personal Values** – Employing an expert to conduct an evaluation (i.e., connoisseurship) can be an excellent way to produce useful results. We value expert opinion because these individuals have experience and understanding others may not. However, when the values and criteria a connoisseur uses to judge merit and worth do not align with those of the client (i.e., what the client thinks is important), the evaluation may have limited value.

The CIPP Model

There are many evaluation approaches and models. Most align well with a specific phase of the ADDIE model. However, Stufflebeam's (2003) **Context, Input, Process and Product**, CIPP model is a comprehensive approach to program evaluation that spans all facets of the design and development process. We present it here, and other evaluation approaches in later chapters dedicated to specific phases of the design and development process.

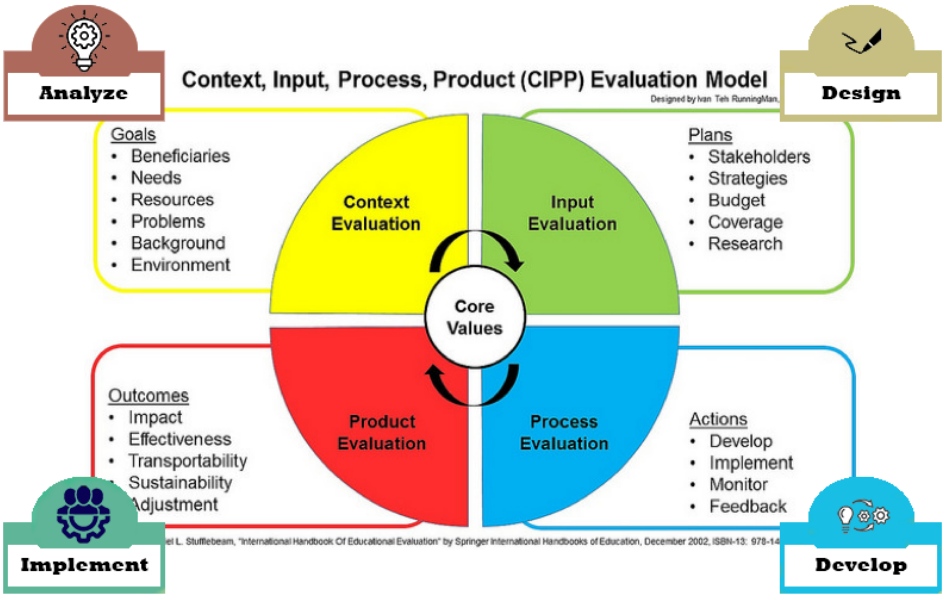
The CIPP framework is a decision-oriented approach to evaluation. It aims to provide an analytic and rational basis for program decision-making at various stages of a program's life cycle (i.e., conceptualization, planning, development, implementation, and maintenance). The CIPP model attempts to make evaluation directly relevant to the needs of decision-makers during the phases and activities of a program's development.

You cannot apply the CIPP model all at once. Each facet of the model must be applied separately depending on the program's current stage of development. Each of the four components of the model aligns well with one of the four phases of the ADDIE model (analysis, design, development, and implementation).

The CIPP model recommends asking formative questions at the beginning of the program's development, then transitions to a summative evaluation of the program once it has been implemented. Guiding questions for each phase include:

- **Context**: What needs to be done? (analysis)
- **Input**: How should it be done? (design)
- **Process**: Is it being done? (development)
- **Product**: Is it succeeding? (implementation, maintenance)

The CIPP model is more of a framework than a prescription for completing an evaluation. Detailed descriptions of the CIPP framework can be obtained from various sources. Additional ways to accomplish each component of the CIPP model are provided in subsequent chapters.



Chapter Summary

- Suggestions for how to conduct an evaluation can be classified as general principles, approaches, or models.
- General principles describe things evaluators need to do so the results of an evaluation are deemed credible.
- Models are prescriptive and provide specific steps that must be followed.
- An approach may approximate a model but the goal is not to replicate the design.
- The purpose, goals, context, and constraints of an evaluation will require the evaluator to adapt and revise any proposed model.
- Pseudo-evaluations should be avoided as they are conducted to promote a specific predetermined solution. These include politically-inspired and advocacy-based evaluations.
- Quasi-evaluations provide good information, but the value of the findings is limited in some way. Evaluation classified as quasi-evaluations could be improved by expanding the scope of the evaluation and the criteria used to determine merit and worth.
- The CIPP model is a comprehensive framework that spans each phase of the ADDIE model for developing instruction (i.e., analysis, design, development, and implementation).
- The CIPP conceptualization phase aligns with the analysis phase.
- The CIPP input phase aligns with the design phase.
- The CIPP process phase aligns with the development phase.
- The CIPP product phase aligns with the implementation phase.

Discussion Questions

1. Think of an evaluation you might consider completing. Provide an example this evaluation might become a pseudo- or quasi-evaluation. What steps should be taken to avoid this?

References

Stufflebeam, D. L., & Coryn, C. L. (2014). *Evaluation theory, models, and applications* (Vol. 50). John Wiley & Sons.

Stufflebeam, D. (2003). The CIPP model of evaluation. In T. Kellaghan, D. Stufflebeam & L. Wingate (Eds.), Springer international handbooks of education: International handbook of educational evaluation.

Evaluation in the Analysis phase

Evaluation in the analysis phase often involves a gap analysis to identify and describe the difference between what is and what we want it to be. The needs analysis uses the information from the gap analysis to identify the cause of the problem and evaluate potential solutions. Before deciding to design a new solution you also need to determine the nature of the problem and whether solutions already exist.

Some typical evaluation approaches and designs used by instructional designers in the analysis phase are summarized here.

Gap Analysis (also called a Needs Analysis)

A search of this topic will likely produce different descriptions of what a needs analysis (or needs assessment) might entail. Some sources refer to a needs assessment as an evaluation whose purpose is to determine the degree to which an objective or goal was met. We refer to these types of evaluations as objectives-oriented or effectiveness evaluations.

What many call a needs analysis might more appropriately be referred to as a gap analysis. The terms *need* and *gap* are often used synonymously, but technically, an identified gap necessitates a solution; the gap defines the need. Said another way, we need a solution to fill a gap. These two activities (i.e., the gap and needs analysis) are usually performed together as one process. A gap analysis is used to identify and describe the difference (i.e., gap) between what is and what we want it to be. The needs analysis uses the information from the gap analysis to identify the cause of the problem and evaluate potential solutions. Thus, a gap analysis identifies problems (i.e., gaps), and a needs analysis recommends a potential (needed) solution.

You may detect various kinds of gaps. The gap might be a performance gap, a knowledge gap, a competency or ability gap. Many of these gaps lead to a need for training or practice, and other gaps are reduced by providing an educational product or tool. Sometimes the evaluator recognizes the gap cannot be filled. Before you can confidently recommend a solution, you will need to identify the cause of the gap. You will also need to consider how important it is to fill the gap and the likelihood that any solutions you propose will have the desired effect.

A gap can represent a perceived, present, unrecognized, or potential need. The gap may refer to the difference between intended goals (or objectives) and the actual (or desired) outcome (i.e., perceived need). It may represent the difference between a desired state and the current situation (i.e., a present need). Or, it may refer to something users don't have but could use (need) to accomplish some task (i.e., something a student or provider needs to accomplish a task). In addition, a gap may represent something an individual may need in the future (i.e., a potential need) even though the individual may not appreciate or be aware of the need (i.e., unrecognized need).

Those conducting a needs analysis are often susceptible to an *availability heuristic* cognitive bias, where decisions are based on the first viable option or solution that comes to mind. For example, it is not uncommon for people to quickly conclude that training must be developed or technology needs to be purchased before identifying the actual cause of the problem and whether that solution would be appropriate.

The basic steps required to accomplish a generic needs analysis are:

Step 2: Identify the problem (need or gap).

Step 3: Describe the various users experiencing the problem.

Step 4: Identify the cause or reason for the problem.

Step 5: Assess how important it is that the problem be solved.

Step 6: Recommend solutions.

To a hammer, all problems are a nail. However, in education,

not all learning gaps can be solved by creating more (or better) training.

There are several ways a Needs Analysis might be applied

Performance gap analysis

In their 1997 book, Mager and Pipe present a model for analyzing performance problems. Various depictions of the models exist (flowchart 1, flowchart 2, flowchart 3).

In Mager and Pipe's model, once the gap (problem) has been identified, the evaluator must identify the cause of the problem, then select an appropriate solution for the performance problem. This is done by asking a series of questions. Each question is intended to provide information so the evaluator can determine the actual cause of the problem and propose a targeted solution.

It is not uncommon for instructional designers to produce training whenever it's asked for, even when training isn't going to solve the performance problem or produce the desired results. Training is only the best intervention when the performance issue is the result of a lack of knowledge and skill. Some issues that would not require training as a solution include:

Being out of practice	Skill is rarely used
Performance is punishing	Its boring, difficult, unpleasant
Performance is unrewarded	Poor compensation, lack of acknowledgment
Nonperformance is rewarding	Conflicting activities are more enticing, lack motivation
Obstacles exist prohibiting the performance	Policies, conditions, facilities, lack of resources, tools
Need to perform well isn't important	Intended outcomes are trivial, little impact

Training is needed when the performance deficit is caused by a lack of knowledge or skill. In these situations, a training needs assessment should be conducted. A training needs analysis identifies the knowledge, skill, and abilities that can be facilitated through instruction. When training is required, gaps may represent knowledge and skills an individual will need in the future or abilities a person needs now to accomplish a particular job or perform a specific task. While a training need may be identified in the analysis phase as part of a gap analysis, some form of task analysis will also be needed, which is often carried out in the design phase.

However, when training is needed, it is not uncommon for instructional designers to spend an exorbitant amount of time and resources developing training for basic knowledge and skills while devoting insufficient time creating instructions for important topics which may be difficult to teach and challenging to develop. A training needs and task analysis will not solve this problem. The designer must use the information provided by a training needs analysis to make design decisions that will facilitate learning in an appropriate, effective, and efficient manner.

The goal of a training needs assessment is to create learning objectives for the training so assessments, instructional materials, and learning activities can be designed and developed. The basic decision-oriented question that needs to be answered is—what do we need to teach? Some ways we can answer this question include:

- Consulting with experts
- Referring to establish curriculum content standards
- Identifying prerequisites and prior knowledge needs
- Listing special qualifications and personal attributes needed to accomplish a particular job or task
- Completing a task analysis

Consumer Reviews

A consumer review is often utilized once a gap is revealed and one or more products (i.e., solutions) exist that could potentially fulfill the identified need. In an educational setting, solutions may include a training program, policy, service, or product. The evaluator's role in a consumer review is to provide an expert opinion.

Results from a gap analysis and consumer reviews inform an evaluator's recommendation to create a new product or utilize an existing solution. The purpose of a consumer review is to provide information to the decision-makers about the various solutions so they can decide which, if any, should be employed. If none of the existing solutions are viable, an evaluator might recommend developing a new product or revising an existing product.

The evaluator's judgment of the product might be determined through measurement or personal experience. For example, the value of a product may be measured and compared against specific performance benchmarks. Or, the evaluation may be based on the evaluator's professional opinion, other users' experience and preference, or both. The evaluator must capture and report findings in a way that helps individuals make an informed decision regarding a solution's merit, worth, and value.

Once a significant problem or need has been identified, the basic steps required to accomplish a consumer review are:

Step 1: Identify existing products (potential solutions).

Step 2: Establish criteria for judging each solution.

Step 3: Test each product using the standards (criteria) you selected.

Step 5: Make a recommendation.

Chapter Summary

- A few evaluation activities that commonly occur in the analysis phase include: gap analysis, needs analysis, and consumer reviews.
- The goal of gap analysis is to identify problems and deficiencies.
- A needs analysis is used to identify the cause of a gap and recommend solutions.
- A consumer review can help evaluate existing educational products that could serve as a solution to an educational problem (a gap). If no viable solutions exist, a new educational product or solution may be needed.
- Not all educational problems require a training solution—only those involving a lack of knowledge, skill, or ability.
- A performance analysis is a type of needs analysis that can be used to identify the cause of a performance problem.
- A training needs analysis is used when the needs analysis suggests training is required. Its purpose is to create learning objectives for the training.

Discussion Questions

1. Consider an educational product you use or might be interested in using. Identify the purpose of the product and the problem (gap) it solves. Think about how the product works. Who might benefit from this product? Are there alternative (similar) products on the market? Briefly describe how the products differ. Why would someone choose to use one product over another?
2. Think of a task you or someone else is required to perform, something that people may not do well. Using Mager and Pipe's model for analyzing performance problems, work through the steps, identify the potential causes of poor performance and suggest specific ways you might solve the problem.

References

Mager, R. F., & Pipe, P. (1997). Analyzing Performance Problems: Or, You Really Oughta Wanna (3 ed.): Center for Effective Performance.

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

Evaluation in the Design Phase

Theory-based evaluation is often underutilized if utilized at all. Many designers fail to consider the theoretical underpinning of their design choices. Many educators may not consider the pedagogical theories that support a specific instructional approach. The following evaluation approaches are commonly used by instructional designers in the design phase.

While the following evaluation approaches are commonly used by instructional designers in the design phase, it is not uncommon for them to be used in other ways in other phases of the design and development process.

Theory-based evaluation

Theory-based evaluation can be used to a) describe the theory supporting a specific design (i.e., evaluating the appropriateness of the theory) and b) evaluate how well a specific theory was put into practice. The first approach is most suitable in the analysis and design phases, the second in the development and implementation phases as part of an effectiveness evaluation. However, a consumer review may attempt to do both, as you would need to understand the theory used in the design before assessing how effectively the theory was put into practice. For instructional designers, this could entail evaluating pedagogical theories and design principles.

Unfortunately, theory-based evaluation is often underutilized if utilized at all. Many designers fail to consider the theoretical underpinning of their design choices. Often designers cannot articulate the design theory they employ; they only know it seems to work. Likewise, educators may not consider the pedagogical theories that support a specific instructional approach. Sometimes failing to conduct a theory-based evaluation has little impact; other times, it can cause an instructional product to fail completely.

The first step in any theory-based evaluation is identifying the theory being used. Chen (1990) suggests that an evaluator can determine the theory supporting a product's design in one of two ways. They can work with stakeholders (i.e., designers and implementers) to discover the reasons for designing and assumptions stakeholders have about the product. Or, they can use their own knowledge of educational psychology and social science theory to describe the supporting theories and design principles used.

Donaldson (2007) describes a process that balances both tactics. A modified version of the basic steps he recommends to accomplish a theory-based evaluation are:

Step 1: Engage relevant stakeholders - Ask them to articulate the product's purpose (short and long-term objectives and goals) and why they think it will work. Ask them to describe how it is supposed to work.

Step 2: Develop a draft of the theory and present it to stakeholders for discussion, reaction, and input.

intended outcome might be achieved in this way.

Step 4: Communicate your findings to stakeholders and revise your description of the theory if necessary.

Step 5: Probe the theory for more specificity - Ask stakeholders what resources and conditions must exist for the product to work. Consider critical links between theory and practice. Make sure the theory's description includes an accurate account of how the product is intended to be used.

Step 6: Finalize the theory and report evaluation findings - make a judgment regarding the merit of using this theory to support the product's design. Is it likely this product can produce the intended outcomes? Can it be implemented as intended?

If the evaluation is testing theory-to-practice, you might add:

Step 7: Conduct an objective-oriented evaluation – Test the degree to which the learning objectives were accomplished. Attempt to verify a relationship (experimental or correlational) between the product's use and the desired outcomes.

Step 8: Conduct a usability study – Observe the product being used to determine whether the product can be used and is being used as intended. Determine whether the product's use is viable, sustainable, desirable, and efficient.

A good resource for educational theories in instructional designers can be found here.

Logic Models

Logic models are often used during the analysis and design phase of an educational program or initiative. An evaluator might use a logic model to describe the evaluand (program) and the theory supporting the initiatives. While an objective-oriented evaluation can provide information about the effectiveness of a program, it does not explain why it was effective. Logic models make an explicit, often visual, statement about how a program is supposed to work, why it is supposed to work, what is needed to make it work, and what change you might expect to see as a result.

A typical logic model might include:

Inputs: A description of the resources needed to make the program or product work as intended.

Activities: A description of the program's key components. This might include an explanation of how the program is supposed to work and a description of what those implementing the program will be doing and why they are supposed to do it (program theory or logic).

Outputs: Describe any tangible products or deliverables that result from the activities. What will participants be required to do?

Outcomes: Identify the intended benefit and expected results for participants (short-term outcomes and long-term impact).

Several public resources exist that describe the process of developing a logic model (see intro video, steps).

There are several ways to accomplish a task analysis; each approach has a slightly different focus based on the purpose for the analysis. A task analysis is often conducted as part of a training needs analysis. A training needs assessment aims to create learning objectives for the training so assessments, instructional materials, and learning activities can be designed and developed.

Common Steps for Implementing a task analysis:

1. Identify the target skill.
2. Identify prerequisite skills.
3. Breaking down the skill into its component parts (or steps).
4. Confirm that the task is thoroughly analyzed.
5. Determine how the skill should be taught.
6. Implement training and monitor effectiveness.
7. Revise the task analysis and instruction as needed.

Some additional resources that cover this topic can be found online (see basics, steps). A few different ways a task analysis might be conducted include:

Job Task Analysis (JTA)

Once you've determined that training is needed, a JTA aims to determine the knowledge and skills an individual requires to complete a specific job. Several discrete tasks may be required for one job. You use the results of this analysis to create learning objectives that will inform the design of any proposed training. For example, training might provide information if the work requires specialized knowledge. If the work requires specific skills, the training should include practice activities (simulated or authentic) that address the highest priority behaviors.

A JTA is appropriate when a job has clearly delineated work requirements and people are hired to do a specific task. A JTA would not work well if the job description was generic and the employee responsibilities varied. For example, suppose the only requirements for a job were to show up on time and follow directions. In that case, a JTA won't help define learning objectives, and a training course is likely not needed because on-the-job training would be more effective. However, if the job requires people to do a specific task and have specialized knowledge, then a JTA can help define the learning objectives for a course. However, remember that not all training solutions require the development of a course—sometimes a training aid would suffice.

The goal of a JTA is to collect specific information, for example:

- A list of the tasks employees must perform to satisfy specific aspects of their jobs.
- How often the task is performed (i.e., frequency).
- The difficulty of each task.
- The importance of each task.

Skills and knowledge objectives to target included those that are essential, complex, or required frequently. However, the existence of any of these factors may be sufficient to label an objective as high-priority. For example, a good candidate for a learning objective might include a simple task that must be done accurately every time regardless of how frequently it is required or how difficult it is to complete (e.g., hand washing or data entry). If the action has severe consequences when done wrong, it would be vital that training help individuals develop the required knowledge and skills.

You can obtain the information you need in a variety of ways:

- Conduct observations of people doing the job. (cognitive interviews)
- Verify your observation with current employees.

Critical Steps Analysis (CSA)

The term task analysis is often used to refer to a detailed description of the steps required to complete a specific task. We call this a critical steps analysis (or CSA) to differentiate it from other types of task analysis. A CSA breaks down a skill into discrete steps or behaviors needed to complete a task. This may or may not refer to a task associated with a specific job. The number of steps in a CSA will depend on the nature of the skill (i.e., its complexity). It will also depend on the individual learner's ability to understand the process (i.e., their age or cognitive capacity). It may be sufficient to define a task in very general terms, or listing detailed steps may be required. For example, most people understand how to brush their teeth. However, you may need to break this task down for the very young.

1. get a toothbrush
2. place toothpaste on the brush (not too much)
3. make sure the toothbrush makes contact with your teeth
4. clean all your teeth (brush for 2 to 3 minutes)
5. rinse your mouth and the toothbrush

The instruction provided may be verbal (on-the-job training), or it may involve the use of a training aid.

Retrieved from: https://www.alphadentalgroup.com.au/blog/oral-care/how-to-brush-teeth/

There are several ways you might break down a specific skill or behavior into smaller steps, including:

Observation. Watch several individuals and document the steps they take as they complete the task. Note steps people take that are not required, those that are essential, and steps that are difficult to do properly. While it is best to conduct this analysis with those who perform the tasks exceptionally well, information obtained from observing poor performers can be helpful. This information might help you decide when practice activities are needed or more detailed instructions is required.

Cognitive interviews (knowledge audits). Ask an expert in the target skill or behavior to explain what they are doing as they complete the task. Make sure to ask why they are doing each of the steps. Ask those demonstrating the skill to rate the difficulty and importance of what they are doing in terms of the likelihood that they will get a satisfactory result.

Curriculum Content Standards Analysis

Course development in a general education setting often does not have a specific job or career in mind. The curriculum outlines a broad set of competencies (e.g., literacy, numeracy, and social skills) an institution has determined students will need to succeed in an advanced educational setting or as lifelong learners. This general education philosophy is designed to provide students with transferrable skills that will prepare them to gain knowledge, acquire new skills, and broaden their perspectives. Its goal is to help individuals adapt to society's ever-changing needs and become productive workers in the economy.

Just as the term curriculum means different things to different people, so does curriculum analysis. Whereas curriculum development involves building the curriculum to present a comprehensive and coherent educational plan, curriculum analysis involves unpacking the curriculum. Like a task analysis, an instructional designer must evaluate how the parts fit together. The analyst must determine: What need the curriculum is responding to? Who is the curriculum designed for? What are the prerequisite knowledge and skill requirements? What content does it cover? What resources are needed to teach the curriculum?

For instructional designers, conducting a curriculum content analysis is only an evaluative process in that the instructional designer (e.g., a teacher) must develop instructional products that align with a chosen or, more likely, a mandated curriculum. Primarily as a result of the standards-based reform movement in education, curriculum developers constantly review and revise the curriculum. As a result, designers need not spend a lot of time deciding on the learning objectives for a course as they should be clearly outlined in curriculum standards.

Learner Analysis (or Target Audience Analysis)

This is a special type of task analysis. It does not consider what needs to be taught (i.e., learning objectives). You conduct learner analysis to understand the ability and needs of those whose knowledge, skill, or ability you hope to improve. This will help you determine the level of specificity required when listing critical steps. More specifically, you might collect general group information about their:

- Background knowledge
- Developmental readiness (ability)
- Comfort level with technology
- Preferences regarding delivery method (virtual, face-to-face, or self-study)

Overall, you use the information gathered during the learner analysis to make instructional design decisions, such as the level of scaffolding or guidance required, the technology support needed, or the most appropriate delivery methods. Another way to accomplish this task is to use personas (see personas).

Like the learner analysis, an instructional context analysis does not consider what needs to be taught (i.e., learning objectives). Instead, it considers the best way to provide the instruction and the constraints imposed on the instruction by the context. It is not as common as the other types of task analysis, but it is important nonetheless. It was originally used to evaluate business environments (internal and external) but has been adapted for education. To conduct this analysis, you ask yourself what context the learner will be in when they interact with the training and what modality (i.e., conditions) would be best. In some cases, the context may severely limit how instruction can be delivered and the instruction's likely success. For example, teaching online (possibly due to covid or distance) may be a problem for some topics and learning objectives. When students do not have access to the expensive specialized equipment they need, they cannot practice the skills targeted by the learning objectives. This will significantly diminish the effectiveness of the instruction—attempting to design instruction in less than ideal conditions will be challenging and probably ill-advised.

By analyzing the instructional context before designing any instruction, you are able to tailor your instruction with regard to any potential limitations. For example, online instruction may be designed to be accessed using a personal computer but may also be accessed using smartphones. Other issues may constrain the feasibility of designing instructions, including internet access or audio/visual needs. Identifying potential implementation problems can also draw attention to situations where alternative modes of instruction may be needed.

Chapter Summary

There are several evaluative activities that should be considered in the design phase. The consequences of overlooking these evaluations can be costly and often will impact the success of a designer's development and implementation efforts.

- **Theory-based evaluation** - describes the theory supporting a specific design or how well a specific theory is put into practice.
- **Logic models** - logic models describe an evaluand (e.g., a program or instructional product)—how it is supposed to work, the resources needed, the intended outcomes, and the theory supporting that initiative.
- **Task analysis** - A task analysis is conducted to clarify training needs (i.e., learning objectives). There are several ways to accomplish a task analysis.
    - **Job-task Analyses** help us identify the high-priority objective on which we should focus the ensuing training.
    - **Critical Steps Analysis** breaks down a task (i.e., the general description of a task) into the precise steps needed to complete that task.
    - **Curriculum Content Standards** are developed so designers can select appropriate learning objectives from an established set of general learning objectives. Like a task analysis, this analysis must determine the gap or need the instruction will address, the intended learners, and the prerequisite knowledge/skill requirements.
    - **Learner Analyses** describe the intended learners' attributes and characteristics. This helps us make decisions regarding delivery methods and the level of guidance and background information to provide.
    - **Instructional context analyses** identifies the best conditions for providing the training and the constraints imposed on the instruction by the context.

1. Consider an educational product you use. Describe the pedagogical and design theory that makes it effective.
2. Think of a task you regularly undertake (you being an expert at doing the task). List the critical steps a novice might need to understand to accomplish the task.
3. Think of some training you have completed. Describe how a learner analysis and instructional context analysis might inform the design of the instruction.

References

Chen, H. T., & Chen, H. T. (1990). *Theory-driven evaluations*. Sage.

Donaldson, S. I. (2007). *Program theory-driven evaluation science: Strategies and applications*. Routledge.

Mager, R. F., & Pipe, P. (1997). Analyzing Performance Problems: Or, You Really Oughta Wanna (3 ed.): Center for Effective Performance.

Stufflebeam, D. L., & Coryn, C. L. (2014). *Evaluation theory, models, and applications* (Vol. 50). John Wiley & Sons.

Evaluation in the Development Phase

Formative evaluation is prominent in both the design and development phases. It can be part of prototype testing in the design phase or a beta testing process in the development phase. In practice, designers continually evaluate a design's effectiveness, efficiency, and appeal throughout these stages. It is good practice to begin user testing early in thedesign and development process.

While instructional designers commonly conduct **formative evaluations** in the development phase, formative evaluations are also common in the other phases when creating instructional products. For example, in design-based research (DBR), formative evaluation is prominent in both the design and development phases but also can occur in the analysis phase (see figure 1). It can be part of prototype testing in the design phase or a beta testing process in the development phase. In practice, designers continually evaluate a design's effectiveness, efficiency, and appeal throughout these stages; it is good practice to begin user testing early in the design and development process.

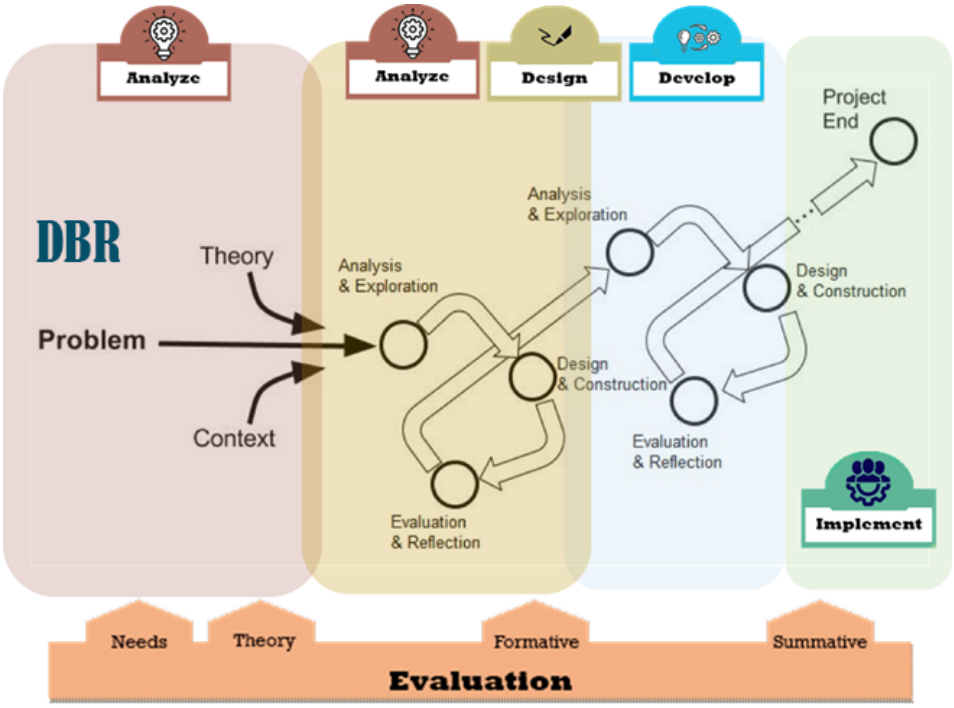


Figure 1: Evaluation Integration within a Design Based Research Model.

The evaluations carried out in the development phase are often short but numerous as hundreds of design decisions need to be made. The products we evaluate are typically beta versions; the final version may become something entirely different. Evaluation in this phase helps refine the product to the point that it is good enough to implement, even if it is not perfect. The implemented product needs to be an adequate solution to the instructional problem (i.e., gap or need), not a perfect solution. Although, even if a product works, it also needs to appeal to the user.

User testing involves getting information from actual users (view video). We are not testing the users; we are testing the product's design and how users interact with the product. We want to know what they need and want the product to do. This is why many call this usability testing. The concept of user testing is based on human-centered design principles and the idea that products are designed for people to use. Human-centered design requires product developers to empathize with the end-user, understand their needs, and build products they want and enjoy using. To do this, designers need user input and formative evaluation.

We use many labels to describe the evaluation activities performed in this stage of production; all are related and often represent distinctions without a lot of difference. For example, user experience (UX) testing and usability testing both fall into a broad category of User Testing.

UX testing vs Usability testing

Often people use the terms usability and UX testing interchangeably. User testing was the original term, followed by usability testing. UX design testing is the more recent term and is debatably more widely used.

Some definitions suggest usability is concerned only with functionality, ease-of-use, and learnability (i.e., how intuitive the product is to use). They define UX design (and testing) more broadly to include usability, but also additional aspects of the end-users experience associated with marketing, branding, findability, support, accessibility, and overall appeal (see Figure 2, adapted from).

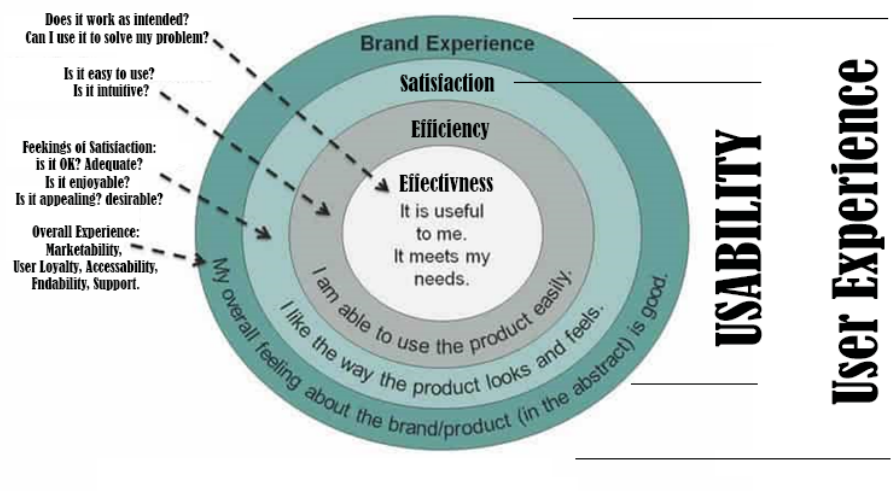


Figure 2: Usability and UX Design Testing.

However, the International Standards Organization (ISO) defines usability in terms of effectiveness, efficiency, and satisfaction (ISO 9241-11, see video). Some suggest that a product can be desirable and not be useful or usable—making UX design a subset of usability or perhaps just overlapping constructs (see Figure 3). The difference is framed as a contrast between science (i.e., usability) and art (i.e, user experience). Those purporting that usability and user experience are different describe usability as analytical, while user experience is subjective; They suggest usability focuses on users' goals, but user experience focuses on how it makes the user feel.

So the main distinction seems to be how you interpret the term satisfaction. Satisfaction meaning "good enough" (i.e., it's functional, I am pleased with how it works), or satisfaction meaning "desirable and appealing" (it works well, AND I love how it looks and how it makes me feel).

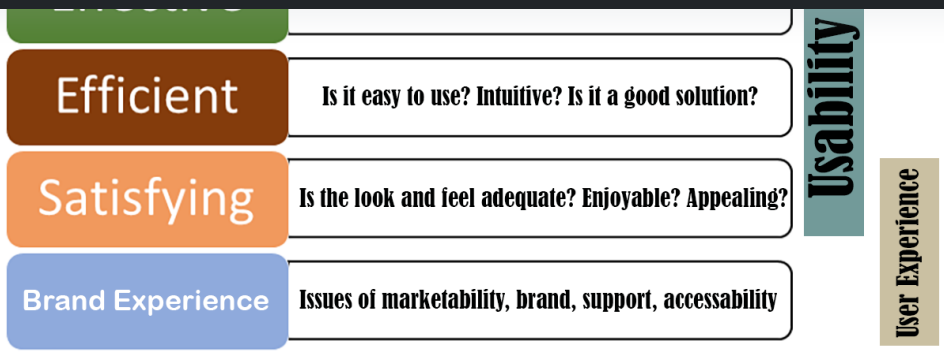


Figure 3: Usability and UX Design Testing.

You can decide for yourself the degree to which these terms are similar or different and what you want to call the evaluation activities you perform in this phase of a product's development. In terms of how formative evaluation benefits the design and development process, we need to consider several issues.

Evaluation Criteria and Purpose

The purpose for usability and UX testing vary, but the evaluator's goal usually is to:

- **Identify problems** in the design of the product or service,
- **Improve the functionality or quality** of the product to enhance the product's performance and increase user satisfaction,
- **Uncovering opportunities** to add features or deal with users' un-met needs,
- **Learn about the target user's** behavior and preferences, or
- **Determine how satisfied** users are with the product.

UX experience and usability testing both use three general factors or criteria to judge the product's value, merit, or worth.

1. **Effectiveness** – The primary criteria for determining effectiveness are utility and usefulness. Judging effectiveness requires that you answer questions like: Does the product work? Does it do what it was designed to do? Can I use it to solve my problem? Is it useful?

   To capture this information, you will need to observe how well users utilize the product to solve a specific problem or complete a task.

2. **Efficiency** – The primary criteria for determining efficiency are ease of use or usability. Can the product be used as intended? Is the design elegant? Intuitive? Fast?

   To capture this information, you will need to observe how users interact with the product.

3. **Satisfaction** – The essential criteria used to evaluate satisfaction are varied. Satisfaction is subjective and depends on one's values (i.e., what is most important to the individual). Basic satisfaction might be determined by the product's usefulness and utility; however, deeper levels of satisfaction might consider the product's safety, cost, support, presentation, and overall appeal. Evaluation efforts primarily focus on the users' experience. How do users feel about the product? Do they like using the product? Is it safe, cost-effective, and enjoyable?

   To get this information, you listen to what users say (e.g., interviews) or document how willingly they use the product (i.e., frequency of use and reuse).

Computer Interaction or HCI). However, UX and usability testing can be applied to any instructional product or service, not just technology or physical products. In addition, much of the information sources for this topic focus on guidelines for designing and developing technology-enabled resources rather than how these products are evaluated. Still, design guidelines and principles can be used as specific criteria by which products might be evaluated (for example, see rules1, rules2).

When Is it appropriate to conduct user testing?

As mentioned earlier, formative evaluation should be started as soon as possible. Gathering information from users can be part of a needs analysis, a consumer review, prototype testing in the design phase, effectiveness testing in the implementation phase, but it is essential during the development phase.

Test Subjects

UX stands for user experience; as such, UX testing cannot be done without users. Both usability and UX testing gather information from users to learn how *they* experience a product. However, some of the evaluation data obtained in a usability study can be acquired from experts (e.g., usability heuristics analysis).

While the designer and experts will need to make some evaluative judgments, formative evaluation of an instructional product needs to get data from those who will actually be using the product. This may include those hoping to benefit from the instructional product's use (i.e., the learner) and those providing or facilitating the expected learning (e.g., teachers, parents, instructors). Both groups are considered primary stakeholders as they will be directly involved with the product's delivery and use. Therefore, both should be asked to provide information about the product's utility, effectiveness, and appeal.

The Typical User

When testing a product, you need to recruit study participants that are representative of your target audience (see video). As your intended users will be diverse, so should the group of individuals you choose to test the product. And while it may be best to select novice users (i.e., those who have never used the product), you can also gain insights from proficient users as well (i.e., those who regularly use the product or have expert knowledge).

Personas and the Intended User

Personas are fictional characters that describe your intended user. Several publically available resources exist that explain the process of developing a persona (video, resource, resource1, resource2, resource3). You may need to develop several personas as there will likely be various groups of individuals who might benefit from using your product. Each persona represents a homogeneous group of potential users with similar characteristics, behaviors, needs, and goals. Creating personas helps the designer understand users' reasons for using a product and what they need the product to do. Identifying a persona can also help select an appropriate group of people to test the product.

Sample Size (and the Rule of 5)

With the exception of a consumer review of existing products, the goal of a user test is to improve a product's design, not just to document its weaknesses. In the development phase, when a product's design is revised based on user feedback, you will want to run additional tests of the product. In each iteration, your test group need not include large numbers of people. If you have a representative sample of key informants, each test iteration can use a small testing cohort (3-5 participants, see source. video). This is called qualitative sampling. Using a limited number of users, you can often identify the majority of issues you will need to address. However, in your initial testing iterations, you may only need a single user to uncover severe flaws in the design. If this happens, you may wish to suspend testing to fix these issues before resuming your analysis with additional testers. This will definitely be the case if the issue is a safety concern. However, you may need a larger group to conduct

Test Session Basics

Before you start testing, a few decisions you need to make include:

Moderated vs. Unmoderated - Moderated sessions allow for a back and forth discussion between the participant and facilitator. Facilitators can ask questions for clarification or dive into issues during or after the user completes tasks. The participant completes unmoderated usability sessions with no interaction from a facilitator. They are asked to explore using the product independently and report back.

As a general rule of thumb, moderated testing is more costly (i.e., facilities, time, and setup) but allows the facilitator to get detailed responses and understand the reasoning behind user behavior. Unmoderated testing is less expensive and is more authentic. However, unmoderated user sessions can provide superficial or incomplete feedback. The facilitator may need to conduct a detailed interview or have the user complete a survey once they have finished testing the product.

Remote vs. In-person – Remote testing is typically unmoderated and, as the name suggests, is done outside a structured laboratory setting in the participant's home or workplace. Remote unmoderated testing doesn't go as deep into a participant's reasoning, but it allows many people to be tested in different areas using fewer resources. In-person testing is usually done in a lab setting and is typically moderated. However, an unmoderated session can be conducted in a lab setting. The evaluator may record or observe the user interacting with the product in an unmoderated session, but they analyze body language, facial expression, behavior without interacting with the user.

Gorilla Testing – is testing in the wild. Instead of recruiting a specific targeted audience, participants are approached in public places and asked to perform a quick usability test. The sessions should last no more than 10 to 15 minutes and cover only a few tasks. It is best to do gorilla testing in the early stages of the product development—when you have a tangible design (wireframes or lo-fi prototypes) and what to know whether you're moving in the right direction. This method is beneficial for gathering quick feedback to validate assumptions, identify core usability issues, and gauge interest in the product.

Lab testing – The term laboratory may be misunderstood when describing a setting in which products are tested. Indeed, participants may be invited to a location where specialized apparatus or materials will be used (e.g., eye tracking equipment), but whenever you invite someone to test a product in an environment of your choosing, it might be considered a laboratory test. A lab setting is testing done in unique environments under specific conditions and supervised by a moderator. In contrast, field studies are defined as observations of users in their own environment as they perform their own tasks. Any time you test in a controlled setting, you run the risk of getting skewed results to some extent. Lab testing is essential; however, you will also need to test in a more authentic setting once the product is ready to implement.

When testing the design of a new asynchronous online course, designers conducted several remote unmoderated evaluations of the product with a diverse group of participants from the target population. Users testing the product were given access to the course and asked to work through the material and give their impressions. One aspect of the design included external links to supplemental information. Under laboratory conditions, those testing this feature of the course indicated they loved the opportunity to search and review these optional materials. Some of the reviewers reported spending hours working through the elective content. However, summative evaluation results conducted once the product was implemented revealed that students enrolled in the course never used this feature, not once. Students working in an uncontrolled authentic setting determined that accessing this information had no impact on their grades; as a result, they didn't. So while user testing under laboratory conditions confirmed the potential benefits of external links, testing in the classroom exposed this as an unrealized potential (i.e., a theory-to-practice issue). You cannot always control for all the confounding variables that affect actual use. (source Davies, 1999)

A few testing methods you might consider include:

Expert Evaluation (usability heuristics analysis) - Expert Evaluation (or heuristic evaluation) is different from a typical usability study in that those providing data are not typical users. Experts evaluate a product's interface against established criteria and judge its compliance with recognized usability principles (the heuristics). Heuristic analysis is a process where experts use rules of thumb to measure the usability of a product's design. Expert evaluation helps design teams enhance product usability early in the design and development process. Depending on the instructional product, different design principles will apply. Identifying appropriate heuristic principles can be the focus of a theory-based evaluation. (video, steps, example of website heuristics)

A/B testing - A/B testing (or A/B split testing) refers to an experimental process where people are shown two or more versions of something and asked to decide which is best. A refers to the 'control' or the original design. And B refers to the 'variation' or a new version of the design. An A/B split test takes half of your participants and presents them with version A and presents version B to the other half. You then collect data to see which works best. A/B testing is often used to optimize website performance or improve how users experience the product. (see primer, steps)

Card Sorting - Card sorting is a technique that involves asking users to organize information into logical groups. Users are given a series of labeled cards and asked to sort them into groups that they think are appropriate. It is used to figure out the best way to organize information. Often the designer is has a biased view of the organization based on their experience. Card sorting exercises can help designers figure out an organization scheme that best matches users' mental model of potential users rather than what the designer thinks is most logical. This can also be used to organize the scope and sequence of instructional content and is an excellent method for prioritizing content. Card sorting is great for optimizing a product's information architecture before building a prototype, lo-fi mockup, or wireframe. (see examples)

Cognitive Think-aloud Interviews – this technique goes by different names (e.g., context inquiries), but the basic technique asks test participants to perform a number of tasks while explaining what they are doing and why. This is an unmoderated testing approach where the evaluator tries to capture what users think as they perform the task without intervention. The evaluator does not interact with the user; they record the user's actions, their explanations, and note any problems. Several publically available resources exist that cover this topic (see Intro).

Cooperative evaluation is a moderated variant of a think-aloud interview. In addition to getting the user to think aloud, the evaluator can ask the user to elaborate or consider "What if ?" situations; likewise, the user is encouraged to provide

and how well the person conducts the interview.

Before you begin, you will also need to consider the following:

Creating Scenarios

A scenario is a very short story describing a user's need for specific information or a desire to complete a specific task. There are various types of scenarios you might create, depending on the purpose of your test. You can also ask users for their own scenarios then watch and listen as they accomplish the task. A scenario should represent a realistic and typical task the product was designed to accomplish. The facilitator should encourage users to interact with the interface on their own without guidance. Scenarios should not include any information about how to accomplish a task or give away the answer. Several publically available resources describe this process. Several publically available resources exist that cover this topic (see video explanation, resource1, resource2).

Moderator guidelines

An essential aspect of any moderated user test is the person facilitating the evaluation. An inexperienced moderator may inadvertently thwart the interview process. This can be done by failing to establish rapport, asking leading questions, failing to probe sufficiently, and neglecting to observe carefully. Usability testing can yield valuable insights, but user testing requires carefully crafted task scenarios and questions.

A few basic rules for interacting with evaluation participants include:

- Given the purpose of the test, determine the best way to conduct the test and how to interact with the participant.
- Respect the test participants' rights and time.
- Consider the test participants as experts but remain in charge.
- Focus on the goal of the evaluation. Use carefully crafted scenarios.
- Be professional but genuine and gracious. Be open, unbiased, not offended, surprised, or overly emotional.
- Listen, let the test participants do most of the talking!
- Don't give away information inadvertently, explain how to do a task, or ask leading questions.
- Seek to fully understand. Use probing questions effectively.

An excellent resource on this topic is provided by Molich et al. (2020) [alt link]. Several additional free resources that describe this process are available online. (see video explanation, common mistakes)

When deciding on which educational psychology textbook to use in a course, the instructor decided to ask several students to give their opinion. He provided them with three options and asked which would be best. This was an unmoderated remote evaluation of the textbooks using a simple A/B testing option. The student tended to agree on one textbook. When asked why, students indicated they liked the design and colors on the front of the book. Aesthetics are important—but the unmoderated format and lack of a carefully created guiding scenario resulted in a failed evaluation. The usability of the textbook should have been determined using a set of scenarios devised to evaluate the usefulness and efficiency of the design and not just the appeal. A more thorough evaluation might also have included an expert review of the content (i.e., correctness) and the design principles used.

Session overview

A typical usability test session should not last too long (less than an hour) and might include the following:

- **Introduction** - Make the participant comfortable, explain what will happen, and ask a few questions about the person to understand their relevant experience.
- **Present the scenario**(s) - Then watch and listen as they attempt to complete the task proposed in the scenario. Prompt only to gain understanding or encourage the user to explain what they are thinking or feeling. If relevant, ask participants for their own scenarios. What would they like to accomplish with the product?
- **Debriefing** - At the end, you can ask questions about the experience and follow up on any information provided about the product that needs further explanation. You might ask the user for suggestions or a critique of the product. If appropriate, ask the user how satisfied they are with the product's functionality, esthetics, appeal, and desirability.

Triangulation

One last thing to remember is to trust but verify. Not everything the user says will be accurate or reasonable, and opinions about how to proceed can be diverse. Use multiple sources and look at the problem from multiple points of view. Combine multiple types of data and obtain information using several methods. Recommendations should be reasonable, ethical, plausible, and for the most part, required. Remember, not all changes can or should be done (even if deemed necessary), and not all nonessential changes should be ignored if they improve the product and are reasonable.

- Formative evaluation is typically conducted in the design phase.
- User Testing is a fundamental aspect of formative evaluation.
- By User Testing, we mean having the intended end-users test the product's design to determine how users interact with the product.
- Both UX testing and Usability testing focus on human-centered design principles and the idea that products are designed for people to use.
- The ISO defines usability in terms of effectiveness, efficiency, and satisfaction.
- Formative evaluation should begin early in the design and development process.
- Typical users and subject matter experts should be used to evaluate the product.
- Personas can be developed to describe the typical intended users of a product.
- Formative evaluation test groups need not be large (Rule of 5).
- Qualitative sampling should be used to identify key informants.
- User testing can be moderated or unmoderated, remote or in-person, conducted in a laboratory setting or as a field study.
- Various types of testing can be employed, including expert evaluations (heuristic analysis), A/B testing, card sorting, and cognitive interview (context inquiries).
- The value of the information obtained from a user test depends on the task scenario used and how well the moderator conducts the interview.
- Triangulation is needed to verify data and fully understand issues.
- Recommendation for modifying a product should be reasonable, ethical, plausible, and for the most part, required.

Discussion Questions

1. Consider a product you would like to evaluate. Describe the best way to test the product's usability in terms of conducting a moderated vs. unmoderated, remote vs. in-person, and laboratory vs. field study. What would you recommend and why?
2. Consider an educational product you are familiar with. Describe a persona (a user group) that typically would use this product.

References

Davies, R. (1999). Evaluation Comparison of Online and Classroom Instruction. *Higher Colleges of Technology Journal*. *4*(1), 33-46.

Molich, R., Wilson, C., Barnum, C. M., Cooley, D., Krug, S., LaRoche, C., ... & Traynor, B. (2020). How Professionals Moderate Usability Tests. Journal of Usability Studies, 15(4).

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.

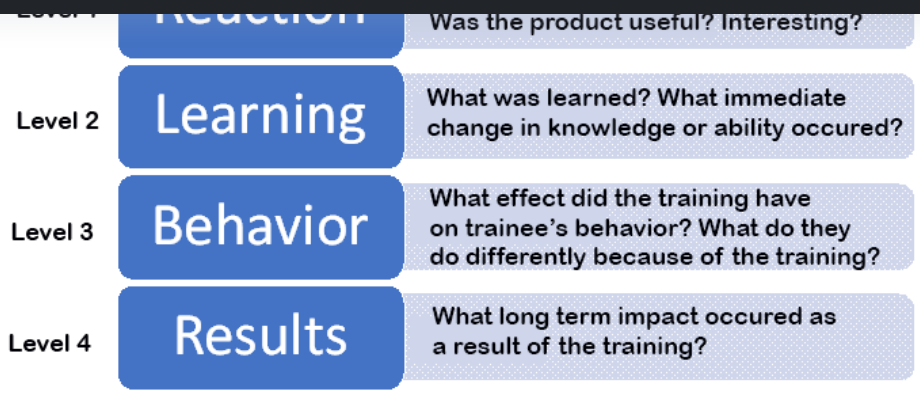


Figure 1: Phases in the Kirkpatrick Model.

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.

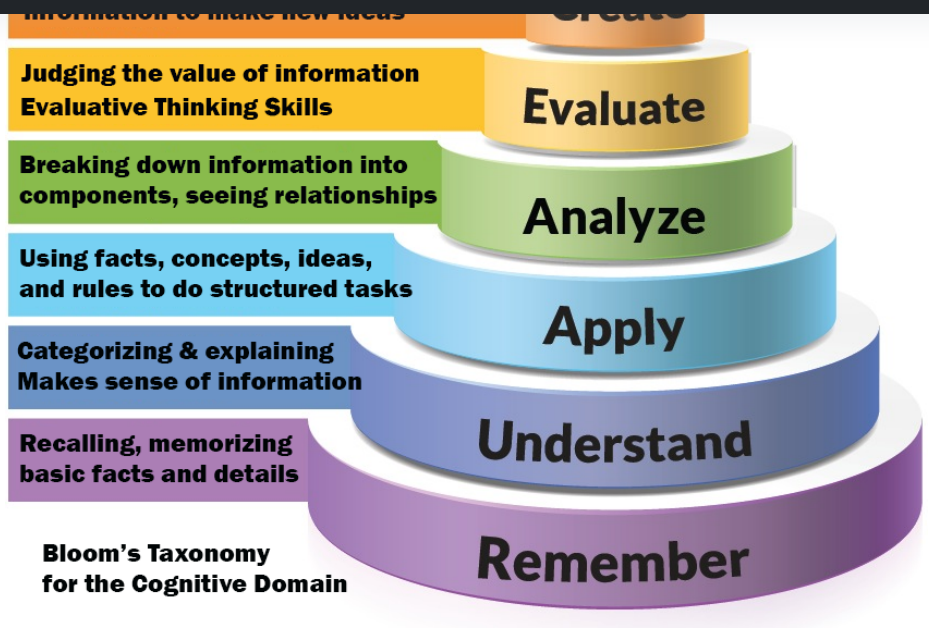


Figure 2: Bloom's Taxonomy for the Cognitive Domain. (adapted from source)

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.

Personnel Evaluation

A fundamental tenet of product evaluation is that you evaluate the design, NOT the user. However, there are times when this is not necessarily true. Stufflebeem & Coryn (2014) describe what they call *ghosts in the system*. For example, in a program evaluation, the designer creates an instructional product to be implemented or provided by someone. Likewise, a policy must be implemented by people and followed by individuals. In addition, educational products must be utilized by the intended users. However, most people don't like to be evaluated because they are being judged, but making judgments is a fundamental aspect of evaluation. So, we say we are evaluating the product but we are also evaluating the people implementing the program, using the product, or applying the policy (i.e., the ghosts in the system). The transition from judging the design of a product to judging the user often happens between the formative and summative evaluation phases. However, it can happen during a needs analysis when conducting a Performance Gap Analysis.

In addition, a fundamental assumption when evaluating a product is that the user intends to use the product and is capable of benefiting from the product. This may not always be the case. When conducting a formative evaluation test subjects often attempt to be helpful. However, once the product has been implemented, users are not always so accomodating. When conducting formative evaluates of a product, you might ask whether users *can* use the product as intended. However, when conducting summative evaluations, you might ask whether users are *willing* to use the product. The difference is subtle but important. The formative evaluation of a product judges its design (e.g., can it be used); however, once the product has been judged adequate and has been implemented, you often judge how willing the intended users are to utilize the product to accomplish specific learning objectives. You might also ask how capable individuals are when judging the effectiveness of a product. It is not always a question of whether the product is effective but rather how willing and interested learners are in accomplishing the expected learning.

When conducting a Performance Gap Analysis, you are basically performing a personnel evaluation. You assess how well someone performs a task and judge whether the performance is adequate. You make determinations about the cause of the problem and decide what action to take. Sometimes the solution is to create training or provide practice, but often the solution is to establish a policy (remove obstacles or arrange for consequences). The purpose may be framed as whether an instructional product is needed or what might be done to improve the effectiveness of a product (its implementation or application), but essentially, in a summative evaluation, you are evaluating people and what you can do to get them to accomplish the required tasks or implement the product as intended.

Understanding this difference is important when making recommendations. We note the deficiency in users when, for example, we fault user error in plane crashes or automobile accidents. The user may be driving too fast for the road conditions or driving impaired or distracted – it is not the product that needs improvement. Likewise, many perfectly viable educational products exist. Some might need to be improved, but often a more likely reason a product is labeled "ineffective" is that it is not used at all or as intended. When a product would facilitate learning if used correctly, it would be ill-advised to recommend an instructional product be abandoned or revised simply because a user is unwilling to take advantage of the opportunity.

For some time now, one criterion used to judge educational products and programs has been that the instruction must be interesting and engaging. As a result, students don't always feel inclined to do anything that is not interesting or fun. However, many educational tools (or topics) are essentially uninteresting—they aren't meant to be interesting; these tools are believed to be useful (or in the case of a topic, essential). Even if they are interesting at first (i.e., initially novel), educational products will eventually need to be used to accomplish a task, not to entertain. Students must be intrinsically motivated to learn; they cannot always be extrinsically enticed to participate. They can and need to accomplish difficult, uninteresting tasks. However, when conducting summative evaluations in authentic contexts, students often comment that an activity or product is uninteresting, too complicated, or difficult to understand. They have selected specific criteria that focus on certain aspects of satisfaction rather than effectiveness and utility. When receiving this type of feedback, the user might really be saying that they don't want to learn what is expected of them, or they prefer not to do anything that challenges their ability.

In a negative case evaluation, you often find that some individuals find a product facilitates their ability to accomplish a task or learn a topic. However, others do not. It may not be the fault of the product's design; it may be the result of an ineffective teacher, the learning environment, or it may be that the learner is unwilling, resistant, unprepared, or incapable of accomplishing the expected learning. Students with an external locus of control may blame external factors for their failure rather than accept responsibility for their learning and expend the effort needed to accomplish the required learning. In these situations, care needs to be taken as to what recommendations to suggest and how to convey sensitive recommendations.

You can't make me! A Corporate Example

Learning doesn't just happen in schools. Corporate training benefits from instructional design just as much as schools. Companies need to train employees to do specific tasks accurately and consistently. Unfortunately, a training program may be effective in that employees know what and how to do something; however, they may choose not to do what is expected for various reasons. For example, employees may become lazy or tire easily; they may not perform well in adverse situations (e.g., dealing with the unrealistic expectations of demanding customers), or they may not care.

After receiving training, you may assess an employee's performance and find that they can do what is expected. You may also later observe employees fail to perform the task consistently when they don't think they are being watched. In these cases, a performance task analysis is needed. It has nothing to do with whether the training needs to be improved but how to motivate individuals to do their job. Sometimes, implementing a policy, removing obstacles, or arranging for positive or negative consequences might solve the problem. Other times terminating the employee may be the best option. Each of these solutions has less to do with a formative evaluation of the instructional products used to train people and more to do with summative evaluations of the personnel.

- Product evaluation typically focuses on evaluating the product design, not judging the user. However, sometimes summative evaluations must evaluate users, and a performance gap analysis essentially evaluates personnel, not products.
- When a product is found to be ineffective, it is essential to identify the actual cause of the problem.
- Not all products need to be revised or improved.
- Not all products need to be interesting or fun. Satisfaction also must consider whether a product is effective and efficient. Choosing appropriate evaluation criteria is essential.
- Sometimes an evaluation finds the problem with ineffective products is an issue of user error, a lack of interest, or learner intent.
- An instructional product may be effective and individuals may still choose not to use what they have learned or behave (perform) as they have been trained

Discussion Questions

1. Think of a product that works perfectly but is not often used. Explain the reason it is not used. Consider ways you might entice potential users to use the product.
2. Describe an educational situation where an effective instructional product is used but students choose not to act on what they have learned. What recommendations would be appropriate? Which recommendations might not be appropriate?

References

Stufflebeam, D. L., & Coryn, C. L. (2014). *Evaluation theory, models, and applications* (Vol. 50). John Wiley & Sons.

Data Collection Primer

Evaluation models often explain what needs to be done but rarely how to do it. This chapter presents a primer on data collection methods. The precise way these activities are performed depends on the evaluation context and the information needed. However, the basic principles apply regardless of the context.

Data Quality

The first and most important principle related to data collection is quality. A common way to remember this principle is GIGO (Garbage In, Garbage Out). No matter how well you design your evaluations, if the information you collect is inaccurate or incomplete, the interpretations of your results and the recommendations you make will be flawed. One of the biggest mistakes researchers and evaluators make is assuming the data they collect or are provided is accurate without taking steps to confirm the data is accurate (i.e., trust but verify, triangulate).

Data collection can go wrong in many different ways, depending on the purpose for collecting data and the methods being utilized. When the information we collect is incorrect, we call this measurement error, bias, or noise. Sometimes the data errors result from inadequate data collection instruments; other times, bias occurs because the respondent is unable or unwilling to provide the information we wish to obtain. Errors can also occur when interpreting the data we obtain. Measurement error is described differently and is subject to different kinds of errors depending on the type of data you are collecting.

Self-report Data Collection Issues. The term *response bias* is commonly used to describe issues that occur when collecting self-report data (i.e., surveys and interviews). Response bias occurs in many ways. These errors occur because the respondent cannot or will not provide accurate data. The accuracy of the data we collect affects the integrity of the results and the credibility of our interpretations. Self-report data collection is most appropriate when evaluating user satisfaction, usability, and UX testing (i.e., learner reactions, perspective, and feeling). Several types of response bias associated with surveys and interviews include: (see source)

> **Recall Bias.** This is common in self-report situations when respondents are asked to provide information retrospectively. Human memory is imperfect. Some information is more likely to be remembered than others. A person's ability to recall events and feelings will depend on the metacognitive ability of the individual and the significance of the event to that particular person (i.e., a *vividness effect*). Recall often depends on the time interval between the event and when the individual is asked to recall their perceptions. A person may have forgotten the event altogether; they may remember incorrectly or revise their recollection (see prestige bias).
>
> **Social Desirability & Conformity Bias.** It can be hard for respondents to openly express non-conformity when asked to self-report their behavior, beliefs, and opinions; this is especially true when the respondent believes they may be ridiculed or despised. In such cases, respondents tend to provide a socially acceptable response (sometimes subconsciously) over their true feelings. For example, a respondent may tend to agree with a statement more strongly than how they truly feel when the item addresses something that is generally seen in society as desirable or expected.
>
> **Prestige Bias or Report Bias.** This bias is related to social desirability bias but it is based on an individual's personal desire to be seen in a positive light. This bias is based on personal feelings, not a general instinct for conformity. It involves selectively revealing or suppressing information. For example, respondents may round up

instead, the individual may honestly remember the facts inaccurately (i.e., they have revised details of the event in their mind). They believe what they remember as true, even though it is not. Respondents often tend to view or recall their own situation in a more favorable light than is actually the case—subconsciously protecting their self-image or inflating their ego. It is often good practice to assume that if a question has a potential prestige component, the responses are likely inflated to present the respondent more favorable. Exactly how much they are inflated will depend on the question, context, and respondents.

Acquiescence or Agreement Bias. This bias is like conformity bias. However, unlike conformity bias, in this case, the respondent will, in general, and inadvertently, agree with statements. With this bias, participants tend to select a positively worded response option (i.e., a *framing effect*) or disproportionately indicate a positive direction to their feelings. This bias will skew results towards the positive.

Item and Option Order Effect Bias. Order bias can be the result of both item order and response option order. The order in which survey items are presented can affect a respondent's answers due to a *priming effect*. People tend to contextualize their responses. Because of this, survey or interview questions that come just before a particular query may provide information that respondents will use as a context when formulating their subsequent answers. This is not always a negative. It is only important to know that if an alternate primer was presented, the responses might be significantly different.

Two common response biases associated with response option order are primacy and recency bias. *Primacy bias* is the tendency for respondents to pick one of the first viable options presented to them. This can happen when a respondent quickly reads through the survey and picks one of the first response options they agree with. *Recency bias* is the tendency to pick an answer option presented at the end of a list. This is especially problematic when a long list of options is presented; the choices respondents read last are more memorable, so they tend to select answers near the end of the list.

Mood Bias and Emotional Mind-Sets. One's mood or mindset will affect the way responses are provided. For example, if a participant is exceptionally happy or angry for some reason while taking a survey, their emotional state affects the general pattern of responses provided. Given time, the respondent's current extreme emotions may subside, which will modify the intensity of the responses provided. Emotional responses can be intense in either a positive or negative direction. You will also see this when the survey addresses an emotionally charged topic. Responses may tend to be on the extreme ends of the response scale, possibly because those who choose to complete the survey have strong opinions; however, mood bias becomes a problem when the respondent's current emotional state temporarily exaggerates opinions.

Central Tendency Bias. This bias refers to the tendency of some individuals to avoid responding in extreme ways. For example, some people may never indicate they strongly agree or are extremely dissatisfied (i.e., nothing is perfect, and nothing is entirely without merit). This is the opposite of a mood bias in that responses from those who have this bias will trend closer to the center of the response scale.

Demand Characteristic Bias. A demand characteristic is used to describe specific cues in research that may inadvertently influence a participant's response. A demand characteristic can manifest in a number of different ways if the researcher is not careful when designing and proceeding with a study. In social science research, demand characteristics can create bias when the subject becomes aware of the purpose of the study (i.e., a *Hawthorn effect*). This may potentially bias or invalidate the outcomes. When a respondent becomes aware of the reason or purpose of the study, they may intentionally provide answers they feel would influence the results. For example, if a respondent figured out that the results of a survey will be used to set policy, the individual may attempt to answer in a way that they feel would be beneficial to them.

Random Response Bias. Random response bias can occur when a respondent honestly does not know the answer to the question but answers anyway. This can happen when you ask a respondent to answer a question

another individual, prompting a random response bias.

Another way this bias can manifest is when an individual has an opinion but hasn't considered their true feelings carefully. Like a central tendency bias, these individuals also tend to choose options toward the middle of the response scale. At times, people with this bias will choose the exact middle point (on an odd-numbered response scale) simply because they don't want to think about the issue or don't really care. This bias can also manifest itself maliciously when an individual intentionally responds in a random fashion without actually reading the survey questions or carefully considering what is being asked.

Additional issues can affect the quality of data obtained through interviews. There are several ways to conduct an interview, including structured interviews, unstructured interviews, and focus groups. In addition to response bias issues, interviews are also susceptible to problems regarding how the interviewer conducts the interview (see moderated user testing). A few examples include:

Authority Bias. This happens when an evaluator values information from what they perceive as a reputable source—disbelieving or devaluing other sources of information.

Personality Issues. This is a concern when gathering data using a focus group. A focus group can be more efficient than individual interviews but only if they are conducted properly. The makeup of the group is also essential. A focus group can be a synergetic dialogue or a one-sided conversation. Understanding personality characteristics matters because your goal is to hear the voices of all participants and understand the issues fully from various perspectives. People who are shy, insecure, non-confrontational, or prone to authority bias are less likely to share their thought and feeling. In contrast, individuals who are outgoing, assured, opinionated, angry, or in a position of power over others in the group may take over a conversation or restrict the free flow of honest dialogue.

Measurement Issues. Data collection quality issues also occur when using tests that measure cognitive and affective learning outcomes. The terms *valid* and *reliable* describe the quality of data obtained in this manner. Validity in assessment refers to the degree to which a test measures what it was designed to measure, and reliability refers to the consistency of the results. A measurement instrument must adequately and consistently provide evidence (i.e., results) of whether the learning objectives associated with an instructional product have been achieved. An assessment instrument may provide a measure of student learning and ability, or a measure of an individual's attitude, personality, or beliefs (e.g., a scale). In either case, the measurement instrument must adequately align with and target all aspects of the intended learning objectives. This type of data collection is most appropriate when conducting an effectiveness evaluation. An assessment instrument may be included in the design of the instructional products. However, when there are no specific learning objectives provided or the assessment tool is missing or inadequate, the evaluator may need to create a data collection instrument.

Another issue associated with assessment is that of interpretation resulting and the incorrect use of statistical methods. Method errors happen when the evaluator uses an inappropriate statistical procedure or fails to verify that assumptions have been met. Method errors occur when the evaluator is unskilled in interpreting the result correctly. Measurement issues can also happen when the results of a particular analysis or the assessment results are difficult to understand.

Observations Issues. Observations can be a valuable tool when collecting evaluation data. This type of data collection is used when conducting effectiveness evaluations, usability tests, implementation fidelity studies, and negative case analyses. In an effectiveness evaluation, observations are needed when the intended learning objectives involve performance. To properly assess performance, a rubric is needed that outlines the criteria you will use to judge whether the performance meets a specific standard. In other types of evaluation (e.g., usability testing), observations are needed to understand the phenomenon better. Observations can be completed in a laboratory or an authentic setting. Observing individuals in an authentic setting has the advantage of avoiding certain kinds of bias; however, they have

may exhibit some degree of performance anxiety. These situations are the result of *observer influence*.

In addition to the issue of observer influence, observations are particularly prone to interpretation error resulting from observer biases. These errors happen when the evaluator is unskilled or unaware of how a specific bias might influence them. Interpretation error occurs when the evaluator fails to consider alternative explanations, neglects to conduct member checks and peer reviews, or has a specific cognitive bias. Observer bias can also occur when the evaluator is unduly influenced by their personal values, the desire to obtain a specific result, or their inability to understand what they observed. These are often unintentional but not always. A few examples include:

Confirmation Bias. Confirmation bias describes the tendency to look for evidence that supports one's prior beliefs, ignore contradictory evidence, and interpret ambiguous information in a way that confirms a desired position or finding.

Belief Bias. Similar to confirmation bias, belief bias occurs when an evaluator judges the strength of an argument or the importance of evidence based on how well the evidence supports their values, beliefs, or previous understanding.

Attention bias (or blindness). This happens when an observer pays attention to specific details while ignoring or failing to see other potentially significant evidence. This often happens in structured observations when the observer is tasked with collecting specific data. This can also be the result of priming. When an observer is told to expect something, they tend to see it.

Attribution Bias. This is an interpretation error. It happens when the observer notes a specific action, behavior, or event and misunderstands its cause. For example, when you see someone smile, you might attribute the expression to the individual feeling happy, amused, love, or satisfied. However, alternative explanations might include feelings of superiority, disdain, disregard, fear, nervousness, or submission. We might also misinterpret the target of the emotion. We may assume the person is smiling at us when they might be looking at or thinking about something completely different. Likewise, observations are not suitable for identifying another person's intentions.

Halo Effect. A halo effect happens when an evaluator forms a positive or negative impression based on previous knowledge or experience. In these situations, the empirical evidence is ignored or excused.

Reporting Bias. This is typically applied to a respondent who selectively reveals or suppresses information. However, it can also apply to observers when they under or over report specific observations. It may happen when a person is hesitant to report something sensitive or potentially controversial. It can also happen when they over-emphasize observations they think are interesting or unique.

Cultural Bias. This bias involves the tendency for evaluators to interpret observations and judge them based on the values and standards of their own culture.

Data Collection Instruments

A second general principle related to data collection is asking the right questions in the right way.

Too often, evaluators fail to validate their data collection instruments and protocols adequately. Instruments can be flawed in several ways. They may fail to ask the right questions or the question may be asked in a way respondents cannot understand or misinterpret what is being asked. The questions that are posed may be well designed, but the instrument (or protocol) neglects to ask other important questions (a form of attention bias). An interviewer may also fail to prompt a participant to fully explain their answers, thinking they understood fully what was being said (an attribution issue). Pilot testing the instrument with someone from the target population can alleviate issues. Conducting member checks and peer reviews can help when interpreting results. (see survey planning example)

Another important principle of data collection is getting information from direct sources.

Basically, you need to get information from those who have the data—preferable an objective source. For example, you might want to know how often students attended a class or how much time they spent studying. You plan to use this information to disaggregate results as part of an implementation fidelity study or an effectiveness study to verify the instructional product is being used. When planning your data collection you need to consider where you will get the information you need. In this example, you might ask a student's parent how often their child attends class, but they may not be able to provide factual information – only their perceptions or beliefs. The student may be asked to recall how often they study but may be unwilling to respond accurately or may have forgotten. You might obtain the data from a homework log where students record when they started and ended their studies. This would be good if the students faithfully and honestly recorded their study time. Suppose the learning activities are contained in a technology-enabled learning management system (LMS). In that case, you might also rely on data analytics obtained from the system that recorded when a student accessed the online activities. However, even in this situation, you may encounter measurement errors. For example, a student may have logged into the LMS and accessed a homework assignment but also watched television, listened to music, took bathroom breaks, texted friends, and perused the refrigerator for sustenance. When planning your data collection, you need to make sure the source of information will provide accurate data. One of the biggest challenges can be getting access to those who have the information you need.

Chapter Summary

- Data collection is crucial in the evaluation process.
- Unfortunately, we often mistakenly assume that the data we collect is accurate and complete.
- The first principle of data collection is data quality. It is essential that we take steps to ensure our data is accurate and unbiased. We likely will never eliminate all measurement errors, but we can take steps to alleviate the likelihood that they will occur.
- The second principle of data collection is associated with the data collection instruments and protocols we use to collect information. Asking the right questions, in the right way is essential to obtaining the data we need to answer our evaluation questions.
- The last principle requires we use direct sources. We must obtain the information we need from those who have direct knowledge or direct access to the data we need.

Discussion Questions

1. Consider the data you plan to obtain for one of your projects. Describe which threats are most likely to affect the quality of your data? What can you do to alleviate these issues?

Reporting Evaluation Findings

Evaluation reports are a form of storytelling. There is no one correct way to write a report (or tell the story), but depending on the purpose and intended audience, specific guidelines will help you present your evaluation findings accurately and effectively. Both the Logical and Emotional Arguments can help persuasively tell the story you need to communicate.

Wise men speak because they have something to say;

Fools because they have to say something. (Plato)

The following chapter presents ideas related to presenting evaluation findings. There is no one correct way to write a report, but depending on the purpose and intended audience, specific guidelines will help you present your evaluation finding accurately and effectively.

It is not enough for evaluators to conduct a rigorous and insightful investigation; they must also present their findings, interpretations, and recommendations in a logical and compelling manner. Unfortunately, reporting is often a step that evaluators fail to address adequately. Evaluators need to consider how the result will be presented for maximum impact and usefulness. They need to consider the best ways to present information to stakeholders so they will read the report and consider the recommendations being presented.

Basic structure of an evaluation report

All reports will follow a basic outline. This includes an:

- Abstract (sometimes optional)
- Introduction
- Purpose Statement and Evaluation Questions
- Background
- Methods
- Results and Discussion
- Conclusions and Recommendations

If you have written a formal proposal, several of these sections will have already been written and might only need to be updated for accuracy. For example, you may have proposed specific methods, but things may not have worked out as intended, or you now have specific details of how the data was captured (e.g., response rates and data sources). You can repurpose the revised sections of the proposal for the final report. The rest of the report (i.e., findings, discussion, conclusions, and recommendations) can only be written once you have collected and analyzed data, then interpreted the findings. In some cases, certain sections can be omitted or collapsed into one section or may need to be separated into two specific sections. You need to have a good reason for altering the basic structure of a report but remember that you will not necessarily write the report sections in the order you eventually present them. For example, if an abstract or executive summary is required, these are often written last.

Interim Reporting

obtained can be a learning opportunity. The evaluator can get the client's reaction, and stakeholders can provide insights and explanations that help the evaluator understand the data they have collected. This includes getting the client's interpretation of the results. Rather than acting as the expert, the evaluator and client might work collaboratively to develop conclusions and recommendations.

Report Formats

There are many ways a report can be formatted. The best way to format a report will depend on the purpose and the intended audience. A report may take the form of an executive summary, an in-person presentation, or a formal written report. Depending on the circumstances, you may need all three.

For academic reporting of an evaluation project or publication in academic journals, you may need to follow specific formatting guidelines (i.e., APA). There are good reasons for following these guidelines; most notably, researchers are often required to search through hundreds of studies and evaluation reports each year. Having these reports structured in a particular way makes their reading more efficient. Using appropriate heading and formatting helps the reader productively navigate the information. The tone need not be extremely formal; however, the writing should not be too casual (i.e., idiomatic or colloquial). Traditionally reports have been written in third person, past tense; however, many journals now allow first-person reporting in some situations.

Remember, good writing requires rewriting. It would be best to plan time to write your report so you have time to rework sections, complete copy editing tasks, and reconsider how best to present information.

Logical and Emotional Arguments

The data you collect provides answers to the evaluation questions you posed and supports the recommendations you make. However, reporting is a form of storytelling. An appealing story can present a logical or an emotional argument. Most academic evaluation reports are expected to present a logical argument, including a rational appraisal of the evidence. However, the most compelling reports also include an emotional appeal. The problem with presenting information logically is that it doesn't always persuade our brains to pay attention. Even though it's logical, it doesn't get remembered, retained, or recalled because it doesn't tap into an emotion. Still, if you base your recommendations solely on emotional arguments, you may have difficulty getting readers to take your evaluation results and recommendations seriously.

An effective evaluation report presents information that is relevant to the specific problem and evaluation questions proposed; then, because evaluations rely on value-based criteria, the answers we arrive at, and the recommendation we make need to be persuasively argued from multiple perspectives. Your evidence must be presented in a way that speaks to both our logic and our emotion.

Results Section

It is important to remember that you have nothing to say until you have collected and analyzed the data. Because evaluations often acquire a vast amount of quantitative data, you should organize these data into charts and tables before writing the results section (see examples). This way, you can decide how to present the findings, which should be included in the results section, and which might best be located in an appendix or a supplemental resource. Qualitative data is presented differently than quantitative data but can be coded and quantified to represent the strength of trends and patterns.

The quality of what you report will depend on the quality of your data and your ability to analyze these data. The purpose of the results section is to organize the evidence you obtained in a way that the reader can understand. You need to focus attention on significant aspects of the data. This will include interesting information as well as those results that do not support your hypothesis. Sometimes an evaluator may present the findings separately (i.e., the fact) and then have a separate discussion section. Other times, the results and discussion section can be combined.

At times an evaluation will uncover information that may be difficult for the client to accept. No one likes to be the bearer of bad news. There are, however, ways to lessen the potential backlash. An evaluator might:

- Prepare the client beforehand by reporting negative findings promptly. This can be done during regular communications with the client or when providing interim reports.
- Allow the client to review the data and make suggestions for interpreting negative findings.
- As the client begins to accept the findings, discuss ways to address the findings and present the results to others.
- Include the client and relevant stakeholders in making recommendations for the final report.
- Present positive results first, then discuss negative findings in the most gentle way possible.
- When practical, present results in person before providing the final written report. Some stakeholders are more likely to accept negative findings when they are presented orally in a relatively friendly manner.
- Be open to revising your recommendations should additional information warrant changes. However, removing or misrepresenting the evidence would be unethical if the data does not support such changes.

Discussion Section

In addition to focusing attention on significant aspects of the data, you should provide readers with an understanding of the issues and introduce new ways of seeing the information in the discussion section. You will need to promote an accurate, balanced, and fair portrayal of the results. This section should persuasively present the results so the logical argument you wish to make is clear. However, check your biases; look for alternative explanations and unintended consequences that occurred or may occur. All your interpretations need to be reasonable, and they need to be supported by the data; still, you need to present the result in context. Don't push a perspective that matches what you hoped to find if the data does not support that interpretation. Include and discuss all results needed to answer the evaluation questions. If you have multiple evaluation questions or purposes, you may want to present the result organized by the evaluation questions.

The American Evaluation Association's Guiding Principles (GP) provide some guidance for writing reports and expectations of evaluators when they communicate findings. These state that evaluators should

- guard against misconceptions, biases, distortions, and errors (see GP A8).
- communicate the evaluation's approaches, methods, and limitations accurately and in sufficient detail to allow others to understand, interpret, and critique their work (A3).
- conduct the evaluation and present results in a way that respects the stakeholders' dignity and self-worth (D4).
- make recommendations that maintain a balance between the client and other stakeholder needs and interests in a way that considering the public interest and good of socieity (E4 and E5).

Clearly, there is no one way to represent the good of society or the best interests of the public. However, sometimes we inadvertently allow our biases to cloud our judgment. Care should be taken not to advocate for a specific recommendation that is not fully supported by the data and does not consider the views and values of various stakeholders.

Conclusions Section

Discuss your conclusions in order of importance or by the evaluation questions**.** You need not summarize all the methods and findings again, but you should remind the reader of the evaluation's purpose and answer the evaluation questions. You may compare your results with those from other studies and provide possible explanations or reasons for any discrepancies should the results differ. Make sure to mention any inconclusive or inconsistent results and explain them as best you can (e.g., based on a negative case evaluation). You may suggest additional research that may be needed to clarify your results.

Briefly describe the limitations of your evaluation to show reviewers and readers that you have considered your study's weaknesses. At the end of the conclusion section, state or summarize your main conclusions once again before listing

Recommendations

The recommendations are perhaps the most critical part of the report and are most often an aspect of the report that will be criticized. Sometimes the recommendation can be avoided if the client wishes to decide for themselves how to proceed. However, most often, the client wants the evaluator's opinion.

Any recommendations you make must be ethical, practical, and realistic. Obviously, it would be best not to suggest the client do anything immoral or unethical. However, the recommendations must also be practical and realistic. Often recommendations are rejected, not because they would not solve a problem or meet a need, but because they cannot be implemented. The recommendation may be too costly, extremely challenging to implement, or may have undesirable consequences. For example, implementing a specific solution may address one specific need but cause additional problems. When possible, recommendations should be made in consultation with the client and other stakeholders.

Chapter Summary

- A common mistake evaluators make is failing to plan time for writing the final report.
- The main purpose of any report is to answer the evaluation questions and make recommendations.
- The format of your report will depend on the purpose of the report and the intended audience.
- Before you write your final report, you need to collect and analyze your data so you have something to say.
- The quality of your report will depend on the quality of the data your collect and your ability to analyze these data.
- It is best practice to discuss interim results with your client. When possible, include the client in interpreting the results and when developing recommendations. This can facilitate the evaluator's understanding and makes it easier to deliver negative results.
- The purpose of the results section is to organize the evidence you obtained in a way that the reader can understand.
- The discussion section should provide a contextual interpretation of the findings.
- The conclusion section summarizes the main findings and makes recommendations.
- The recommendations are often the most likely part of a report stakeholders will criticize.
- Recommendations need to be ethical, practical, and realistic. The recommendations must also take into account the possible ramifications of implementing the evaluator's suggestions.

Discussion Questions

1. Consider some ways a report might be biased. Describe what you might do to alleviate these problems?
2. Suppose that, based on the evidence you collect, you feel it is necessary to make extensive revisions to an instructional product you evaluated. Describe issues stakeholders might have implementing the suggestions.