Evaluation in ID

A Short Introduction

Apostolos Koutropoulos

Evaluation in ID: A Short Introduction

Initial version Written by: Apostolos Koutropoulos

Edited by/contributions by: Rebecca J. Hogue (Ed), Ezio the Cat (to be fair, I undid his edits)

Reviewed by: Lance Eaton, Linda Sudleski

Introduction

Evaluation is an important part of the instructional design process. The questions that instructional designers try to answer through an evaluation process are: Does this learning intervention achieve what it set out to achieve? And, how do we know that it achieves those initial goals?

There is a prevailing belief that evaluation, in the ADDIE process, takes place at the end of the process and it's a one-time deal. This cannot be further from the truth! A good evaluation process is always in the back of every designer's mind. At the end of the day, ADDIE is a cleaner acronym than AEDEDEIE, and even though (E) is at the end of this acronym, it's *never* the final word.

A good evaluation always has hooks to other parts of the design process that a designer can return to in order to improve the alignment of the outcomes and instructional deliverables. These hooks are visible in some visualizations of models like the Dick & Carey Model [D&C Model visual], the Morrison-Ross-Kemp Model [MRK Model visual], the SAM model [SAM Model visual], and even in some depictions of ADDIE [ADDIE Model visual]. As a designer, while you're engaged in the design process, you should always be asking yourself how you would go about evaluating both your overall deliverable and the individual components of that deliverable. As Margaret, one of my old instructional design classmates, used to say "You've got to bake it in! Not just sprinkle it on!"

Determining your Evaluables

OK! You caught me! Perhaps "evaluables" is not a noun that you find in a standard English language dictionary, but you probably get the gist of what I mean. Before you can evaluate something to determine if it's fit for purpose, you need to know what you're evaluating! Or do you? There are, of course, different schools of thought on this matter.

You don't need to go that far back in the history of our field to get a sense of the different views on evaluation. Some schools of thought dictate that designers should just examine the outcomes that they identified in the *Analysis* phase of their design work. That is certainly a way to do it, but it leaves designers painting within pre-established borders and in doing so you might actually *miss* a lot of good findings that come up serendipitously!

There are others, like Scriven, who propose *goal-free evaluations*. Scriven's goal-free evaluation (1974, as cited in Gagne, Briggs, & Wager, 1988) is an evaluation process that assesses the worth of educational outcome effects *beyond* the stated learning outcomes, those outcomes formulated in the Analysis phase. Scriven formulated 13 elements for his evaluation matrix (Gagne, Briggs, & Wager, 1988).

There are others, like Stufflebeam, who accept Scriven's underlying goals for goal-free evaluations but do not accept them as a substitute to *goal-based* evaluation; rather, they are seen as a valuable supplement to it. For Stufflebeam (as cited in Gagne, Briggs, & Wager, 1988), an evaluation should answer questions like what program is being evaluated, what are the main concerns of the audience, how will the information yielded by the evaluation be used? What will be the approach to conduct an evaluation? (contexts, input, process, output), and what questions will need to be addressed?

How you approach an evaluation, goal-based, goal-free, or a blend of both approaches, can vary based on your environment and the expected norms within that environment. For those new to instructional design, it's easier to get your feet wet by sticking with goal-based evaluation. As you become more expert in your instructional design practice, I advocate for switching gears and approaching evaluations with a hybrid approach. Stated simply, evaluate the outcomes that you've determined through your Analysis, but leave yourself open to a kind of serendipity. There are always things that come up that surprise and delight us, and we should always be paying attention for their emergence during our evaluations processes. There are also things that arise that give us pause, and prompt us to reevaluate our designs and implementations. If you are too goal-based you might miss these, but if you are too "loosey-goosey" (to use a non-technical term) in your evaluation approaches you might miss an important part of the process: getting a sense of whether your initial goals were met.

Conducting an Evaluation

Now that we've gotten some theoretical underpinning for evaluation, let's talk a bit about the nuts and bolts of conducting evaluations in your instructional design practice. There are two primary types of evaluation: formative and summative. I have an analogy that helps clarify the difference between formative and summative evaluation. Think of pottery. Formative evaluation is something that happens while the clay vessel is still on the wheel and you're still forming it. The potter makes decisions while they are in the process of forming that vessel. Any issues that stem from forming the vessel can be addressed while the clay is still malleable, and before the vessel is placed in a kiln and hardened. Summative evaluation, on the other hand, is the evaluation that occurs after the pottery is out of the kiln, finished off with glazing, and perhaps it's already purchased and used. At this point, the customer is filling out a customer evaluation card to let you know how they like it.

Formative Evaluation

Formative evaluation is defined by Dick, Carey, and Carey as "the collection of data and information during the development of instruction that can be used to improve the effectiveness of instruction" (2015, p. 283). Formative evaluation has been an important part of ensuring that instructional goals are met, and ensuring that money isn't being wasted. In the pre-digital and pre-connected days, there was considerable cost in materials creation and distribution. Not getting it right didn't just mean that you didn't meet your stated educational goals. It meant a tremendous amount of waste because paper-based materials, which took a considerable amount of time to create, reproduce, and ship, were just deposited in the trash. In a digital learning environment, you may not have the same exact levels of material waste, but the economic inefficiencies are still there. Formative evaluation saves money and your reputation!

Traditional approaches to formative evaluation take a stepped approach to the evaluation of designs and their deliverables. The typical starting point is to ask an external subject matter expert (SME) or a line manager that might have first observed issues with the existing training or performance levels. After that, you test your design with a sample learner, then with a small group of sample learners, and then with smaller field trials. In between the tests you improve your design based on your findings. Once you're satisfied with the results of these evaluations, and changes are made to address the shortcoming of the current, or in-process, design, you can ship a final product and eventually conduct summative evaluations.

There is a rhyme and reason why traditional approaches are so structured, mistakes on shipped products were really costly so the design and production team took their time before products saw the light of day. Even prototype materials took some time to create for smaller field trials, hence the stepped approach made a lot of sense. In today's world, if your context is appropriate, we can afford to be more agile in our design approaches. An example of this is the SAM Model for instructional design (Allen & Sites, 2021). We also want to encourage instructional designers to apply critical thinking at every stage of their work. Thus, before we even ask external SMEs for an evaluation of what we've designed as a preliminary draft, we ask all the "silly" questions in our own team of designers and SMEs, we play devil's advocate, and we indulge our inner two-year-old and keep asking "why?" (see: five whys). If our design contexts prohibit many of the elements in a traditional stepped approach, being a critical designer, and asking those whys, will provide you with a kind of formative evaluation that is invaluable.

Assuming that you've worked with your team to create materials that are ready to be seen by others, and in the process gather some more feedback, your next step should be to have your materials peer-reviewed you an external SME. Your team may have an SME on hand, possibly someone who handed you a stack of PowerPoint files and asked you to create training out of them. One potential issue with engaging with existing SMEs is that they see the training through pre-existing lenses and very specific blinders. They may have preconceived notions of what the training materials and assessments should look like. They may also content-stuff a course when less material is enough to meet the stated learning outcomes. An external SME can provide you with a sense of whether your materials are *factually* accurate, and whether there are any glaring holes in the content that should be covered.

Finally, once your materials are vetted by the various SMEs, it's time to get a sense of how they test with actual learners. In a traditional approach, your evaluation will begin with a one-to-one evaluation and a small learner sample size. The learners in your trials are indicative of the target learner demographic. The one-to-one aspect here relates to the fact that during this phase the designers interact directly with the learners on a one-to-one basis, and can ask learners clarifying questions while they are going through the materials. The goal of this phase is to address any obvious mistakes and errors. It also allows the designer to gauge the clarity of the materials, the impact that the instruction has on the learners, and how feasible the designed intervention is. Sometimes, despite our best efforts, we may design for constraints that we *think* we know, but we don't actually get to verify until we speak to our learners.

The next phase is a small-group evaluation. The main ideas behind small-group evaluations are to determine if the changes we made after our one-to-one evaluations were sufficient, to measure how long completion of learning takes, to determine the costs of offering this training to larger groups, and to determine the attitudes of training managers who will be the product owners after we hand off the final product. In instances where we're creating self-paced learning, this is also the phase where we determine if learners can use our interventions *without* a need for an instructor. In this phase of evaluation, the process is less guided compared to the previous step. Learners interact with the materials as if this is the final version. In this phase, you will typically also see pre-tests and post-tests given to learners who test the learning intervention.

The final stage of a formative evaluation is a field trial. In this phase, we want to see whether changes and updates made after the small group testing were sufficient to address any potential issues with a larger group of learners. We also want to test whether the instruction that we created is actually usable in the environments that we intended them to be used. Access to environments may not always be available (e.g., a busy factory floor), so any guesses or estimates we've made in previous, more controlled, testing phases are now put to the "real world" test. If the learning intervention

requires a course facilitator or instructor, the instructional designer is not visibly involved in this phase. Any instructor resources are handed off, and we, as instructional designers, await the outcomes.

Summative Evaluation

Dick, Carey, and Carey define summative evaluation as "the process of collecting data and information to make decisions about whether the instruction actually works as intended in the performance context; further, it is used to determine whether progress is being made in ameliorating the performance problems that prompted the instructional design effort. The main purpose of summative evaluation is to determine whether the given instruction meets expectations" (p. 343). In other words, summative evaluation, as the name might imply, takes place at the end and examines the sum changes that exist because of the application of the new learning intervention.

A summative evaluation benefits from being used in real-life contexts and is seen by large numbers of learners and subject experts. As with any product that's under scrutiny by a large number of users, summative evaluations can uncover things that were missed during formative evaluations. A summative evaluation can also determine whether the organizational goals were met by employing this newly designed learning intervention. It's important to note that a summative evaluation may go beyond the learning intervention itself. For example, if there is a new process or procedure that needs to be employed on the factory floor, and employees learn this procedure, but their line managers prevent employees from applying what they learned, the issue does not lie with the training, but rather lies in factors outside training.

In a traditional summative evaluation, you can expect to find at least two sections. One section would be an external review, and the other would be an impact analysis. In the external expert review section, you can expect to find reports on how congruent the materials, assessments, and learning processes are with the stated learning objectives, whether the content of the course is complete, up-to-date, and accurate, and whether the overall design makes sense. The impact analysis should be examining skills that learners have attained, often measured by pre- and post-tests, how that learning is transferred to the workplace, how frequently those skills are used, and if there are factors outside of training that prevent the use of those new skills and attitudes (think back to the previous example). Finally, is there evidence that the original issues were addressed with this training design? You will notice that the example contexts tend to revolve around workplace training. This is, in large part, due to the history of instructional design. If your own context is education (K-12 or higher education), your summative evaluation indicators would include different aspects. You might, depending on your context, examine the entry-level skills of students in subsequent courses, skills obtained in an introductory class for example, and examine how that impacts learner preparedness in those courses further downstream.

Again, I'll return to a point I made earlier: It's important to keep in mind that in traditional contexts the design of instruction was *costly*, so a summative evaluation was not something that was rushed. Formative means of testing were an important part of the process. Once designers were satisfied that the design and implementation were good, things were moved onto the next stage of production. A summative evaluation was a means of gauging the "final" output and testing it with many more users that would typically be available for a formative evaluation. If we think of our factory example, a field trial may be conducted in one factory in one region, but the full product may be deployed in all factories across all regions. Another thing to keep in mind is that *summative doesn't have to mean final*. I am reminded of the "paper_title final final final final final final.docx" example. Someone decided to call the document "final," but more edits and changes needed to be made, so it became "final final," and then more changes need to be made, so someone called it "final final final." *Don't fall for this trap*. You cannot just will something into being final. *Changes can be made, even after a summative evaluation*. If you, or your team, determine that changes should be made based on the evidence provided in a summative evaluation, you could go back and update things. In fact, some ID models, like the Dick & Carey model for example, specifically draw a line between summative evaluation and the beginning of the ID process.

The Five Levels of Evaluation

I would be remiss if I wrote an introductory chapter to Evaluation and did not mention the five levels of evaluation. The five levels should really be called the 4+1 levels. The first four levels were developed by Kirkpatrick (1994). The fifth level is an extension, and different people have different extensions to Kirkpatrick's model. The two that I will briefly summarize are Phillips (1997) and Kaufmann (1995) extensions. This model isn't without its criticisms, but it is well known in instructional designer and trainer circles, so it's worth introducing it. The Kirkpatrick model traces its development to training, so when it's used to evaluate other types of learning experiences those contextual factors need to be kept in mind.

Level 1 is *reaction*. This is the first level of Kirkpatrick's taxonomy and it tells you what the participants thought about the training. What were their overall impressions? This can be done at the end of a course, for example with a course evaluation, but it can also be done throughout the course. For example, there can be short evaluations at the end of each module that ask learners what worked and what didn't during that module. There can be different pieces of evidence, like private messages or forum messages with things that learners didn't understand, things that frustrated them, or things that delighted and surprised them. At the end of the course, learners can be asked whether they think that the course, as a whole, met their expectations and if they foresee applying what they learned. Separate to the *designed* course, they can also what they thought of the instructor, if the course was instructor-led. A common artifact you might see at this stage are so-called "Smile sheets." It's important to point out here that *positive responses to Level 1 evaluations do not mean that learning took place*. People may have had a great time in the course, and they may have thought that the instructor was the *bee's knees*, but nothing may have changed as a result of taking the course!

Level 2 is *learning*. This is the second level of Kirkpatrick's taxonomy. This level shows you what type of learning took place, if any learning took place. Did the participants learn something from the training? If so yes, what was it? The means of measuring learning can take many forms. You may see the results of post-tests, and compare them to pre-tests that learners might have taken. You may ask learners to reflect on their learning or have interviews with them; or depending on your context, you may have the ability to observe role plays and other presentations prepared by learners as their learning artifacts. Depending on your contexts, you may need to develop grading criteria that can be applied the same way across different cohorts that are led by different instructors, in the case of instructor-led learning. This should be done to avoid inconsistencies in assessment across cohorts.

Level 3 is *behavior*. This Kirkpatrick level shows you whether the training that learners undertook produced any on-the-job changes. Did the participants use the knowledge and skills they gained from the training when they went back to work? This level is sometimes called *transfer*. For example, if I had an anger management issue that was causing issues at work, and I took a workshop on managing my anger and passed it with flying colors, but still flew off the handle at work, my behavior has not changed; as far as my manager is concerned the learning intervention was not a success despite my high grades! To evaluate this level we need to step outside of the classroom and the learning intervention itself and evaluate the learners in their application context. This type of access may be something that instructional designers and talent developers don't have access to, so you may need to work with stakeholders to obtain access or data. Because application also takes time, this step starts a few months after the learning intervention has concluded.

Level 4 is *results*. This is the final Kirkpatrick level and may or may not be something that you conduct. The results level examines whether the expectations of the organization's stakeholders were met. In other words, did the training accomplish what they expected it to? Sometimes expected results are poorly understood at the onset of a project, even with an in-depth analysis. Sometimes it's not easy getting to the core of the issue. Let me share a personal anecdote. One of my first training jobs was as a trainer for various Microsoft Office products, with Word, Excel, and PowerPoint being the most popular. I used to work for an academic library and the reference desk had a fair amount of questions regarding how to accomplish certain things with these software products. An example of this type of question is "how do I create a histogram using Excel?" The powers that be, some decision level above my own, decided to offer training on these products. As a trainer, I was busy and helped quite a few students during my time located adjacent to the reference desk. However, the reference desk *still* got their fair share of calls for help from students who needed help

with these applications, so ultimately the volume of requests at the reference desk didn't go down. So, the question is: was this training intervention a success or not? How would you evaluate this?

Level 5 is where things diverge a bit. One type of level five evaluation is *Societal contributions*, and we see this in Kaufmann's model. Another type of level 5 evaluation is *Return on Investment (ROI)* in Phillips' model. For Phillips, the context is clearly corporate in nature. Phillips argues for the collection and analysis of data so that a determination can be made if the results of training include process improvements, productivity improvements, or an increase in revenue and profits. This is certainly *a* way of viewing outcomes of training, but not the only way. The bottom line isn't always financial. For Kaufmann, on the other hand, the results for this "mega-level," examine stakeholders broadly either society as a whole, or a company's clientele more broadly. For example, if a company is training their employees to use more environmentally friendly ways of disposing of waste, that kind of outcome impacts not only the learner and their company but those stakeholders close to company outlets that are handling this new type of waste removal.

Conclusion

To wrap things up, I hope to leave you with a few key ideas about evaluation. The first is that evaluation needs to be baked in, not sprinkled on top at the end. Let's call this Margeret's Principle. In order to determine if our learning interventions work we need to determine how we will evaluate that. As Larson and Lockee tell us, "continuous evaluation produces feedback that facilitates continuous improvement" (2019, p. 10). Second, we have a lot of stakeholders in our evaluations. Our learners should be first and foremost, but we also have organizational and societal stakeholders, depending on how far out you reach with your evaluation. Knowing how far you need to reach (or how far you *can* reach) early in the design process can help you gauge how much work evaluation will be. Finally, evaluation is an ongoing effort. You may have some natural punctuation points along the way with formative and summative evaluations, but your work products should always be evaluated for their fitness for use. If you design instruction as a contract worker, your natural punctuation point may actually be the end of your involvement in the process, but someone else may pick it up after you, so design hooks for ongoing evaluation, and document your design decisions. Even if you are going to remain the product owner of certain training documentation helps! Years may pass and you may not know why you did what you did. Document your work, and save your future self some headaches.

References

Allen, M. W., & Sites, R. (2012). *Leaving ADDIE for SAM: An agile model for developing the best learning experiences*. American Society for Training and Development.

Dick, W., Carey, L., & Carey, J. O. (2015). *The systematic design of instruction* (8th Ed). Pearson.

Gagne, R. M., Briggs, L. J., & Wager, W. W. (1988). *Principles of instructional design*. New York: Holt, Rhinehart, and Winston.

Kaufman, R., Keller, J., & Watkins, R. (1995). What works and what doesn't: Evaluation beyond Kirkpatrick. *Performance and Instruction, 35*(2): 8-12. Retrieved from https://edtechbooks.org/-pbSf

Kirkpatrick, D. L. (1994). *Evaluating training programs: the four levels*. San Francisco: Berrett-Koehler.

Larson, M. B., & Lockee, B. B. (2019). *Streamlined ID: A practical guide to instructional design*. Routledge.

Phillips, J. (1997). *Return on Investment in Training and Performance Improvement Programs*. Gulf Publishing Company.