The Learning-Transfer Evaluation Model:

Sending Messages to Enable Learning Effectiveness

by Will Thalheimer, PhD

The Learning-Transfer Evaluation Model: Sending Messages to Enable Learning Effectiveness

By Will Thalheimer, PhD

Citation:

Thalheimer, W. (2018). *The learning-transfer evaluation model: Sending messages to enable learning effectiveness.*
Available at https://WorkLearning.com/Catalog

This report is aligned with LTEM Version 11.

The report version is 002, incorporating fixes for some previous copy-editing errors.

Special Thanks:

To Julie Dirksen, Clark Quinn, Roy Pollock, Adam Neaman, Yvon Dalat, Emma Weber, Scott Weersing, Mark Jenkins, Ingrid Guerra-Lopez, Rob Brinkerhoff, Trudy Mandeville, Mike Rustici and various anonymous audience members at talks and workshops where I discussed the model, especially ISPI's 2017 Design Thinking Conference and the Learning Technologies 2018 conference in London. Without the thoughtful and extensive feedback I received, neither the model nor this report would be any good.

The Learning-Transfer Evaluation Model:
Sending Messages to Enable Learning Effectiveness

by Will Thalheimer, PhD

One of the most insidious and nefarious properties of scientific models is their tendency to take over, and sometimes supplant, reality.

Erwin Chargaff

While seeing any number of black crows does not prove all the crows are black, seeing one white crow disproves it. Thus science proceeds not by proving models correct but by discarding false ones or improving incomplete ones.

Byron K. Jennings

The Gist

The Kirkpatrick-Katzell four-level model—conceived by Raymond Katzell in the 1950s and enhanced and popularized by Donald Kirkpatrick over the next six decades—has left an indelible mark on the workplace learning field. The four-level model was influential in transforming the training-and-development field into the learning-to-performance field. We are indebted to Raymond Katzell and Donald Kirkpatrick for the model's emphasis on the instrumentality of learning, animating the idea that learning is intended to enable behavior change and organizational results. Unfortunately, the four-level model also sends messages—unintended by its authors—that undermine our efforts as learning professionals to improve our learning initiatives. Almost 60 years after the Kirkpatrick-Katzell model was introduced, the workplace learning field is long overdue for an improved model of learning evaluation—one that spurs us to greater learning effectiveness.

This report introduces a serious practical alternative to the Kirkpatrick-Katzell model. The proposed new model is aligned with the science of learning and is intentionally designed to catalog a more robust set of requirements than the Kirkpatrick-Katzell model—requirements targeted to overcome the most urgent failures in learning practice.

In short, the new model provides more appropriate guideposts, enabling us, as learning professionals, to create virtuous cycles of continuous improvement. The new model is

relevant to educational contexts as well, where learning doesn't necessarily focus on job performance per se, but on other future applications of learning.

Minds for Models

We humans build models—not only because it's fun and profitable, but because we need models to help us make sense of our complex world. In truth, we need models because our minds are small. Oh, certainly, with exhaustive concentration we can discern even the most puzzling phenomena. But all of us (if truth be told anymore) are pre-inclined to simplicity—to the least energetic uses of our minds. We like things simple. Nothing wrong with that! It's nature. It's natural. And it actually makes sense if you think about it. Who's got time to waste when we need to focus on what our boss wants, what show to binge watch, what monstrous outrage our political leaders will be caught emoting?

In the learning field, we build all kinds of models. We've got the four levels of Kirkpatrick, the Five Moments of Need, the five phases of ADDIE, the Nine Events of Instruction, the Decisive Dozen. We use these models because we need them. They help us know what to do. But what if they didn't? What if they pushed us toward poor practices? Shouldn't we then discard them or fix them or replace them? Of course!

A model is only as good as the messages it sends. If the messages are helpful, it's a good model. If the messages are harmful, it's time for a better model. Messages push us toward some actions and away from others.

Models create messages. Messages nudge actions. Actions produce results. If our models convey unhelpful messages, we end up with poor results. The more our models convey helpful messages, the better our learning results.

The bottom line—the thing I'm getting to in this report—is that it's time for a better learning-evaluation model. The four-level Kirkpatrick-Katzell model is a clear and present danger. It sends numerous harmful messages. It has pushed us to where we are today—enmeshed in an industry where too few of us are doing good learning evaluation, and hardly any of us are getting good feedback about the effectiveness of our learning.

Mostly, we measure Level 1. Level 1. Level 1. Level 1. It's insanity! Yes, I wrote a book on how to get feedback from learners—how to do a Kirkpatrick-Katzell Level 1—but getting good feedback from learners is the least we should be doing.

Let's think about this.

Learning-Evaluation Messages

In my book, *Performance-Focused Smile Sheets*, I introduced the concept of stealth messages to show how learner-feedback questions can send messages about factors that are critical to learning effectiveness. For example, a question about after-training support sends a stealth message to learners, trainers, and instructional designers that after-training support is critical for learning transfer. A specific focus on management support can send a message that involving learners' managers is vital to learning application.

Stealth messages are messages we send that aren't perceived as being attempts at persuasion. They are potent for two reasons—first because they send messages and second because their messages are more likely to be accepted as givens; less likely to be subjected to extended conscious deliberations in the minds of those receiving the messages. Stealth messages can be delivered with verbal communications, but they can also be delivered or strengthened through the use of symbolic, visual, or metaphoric nudges. The color red provides a stealth message of warning or danger in many human cultures. The ADDIE model, with the "E" standing for evaluation, sends a stealth message that evaluation should be done at the end of the learning development process. ADDIE, then, uses a hybrid of verbal and symbolic messaging. The word "Evaluation" is verbal. Its placement at the end of the model is symbolic, visual, and metaphoric.

I'm spending time on the stealth messaging notion because it explains almost everything about the sorry state of our learning evaluation practices. It also gives us a way out—a way toward an infinitely better model of learning evaluation.

Here's the bottom line: If our models push us toward bad evaluation practices, we're likely to create poor learning evaluations. If our models push us toward good evaluation practices, we're likely to create effective learning evaluations. So, let us start by examining the messages we might want to send.

What messages should we send—if we can—when we consider learning evaluation? Here's my short list:

1. Just because learners ENGAGE IN LEARNING doesn't mean they will have learned. Therefore, measuring attendance is an inadequate way of evaluating learning.

2. Just because learners COMPLETE A LEARNING EVENT doesn't mean they learned. Therefore, measuring course completion is an inadequate way of evaluating learning.

3. Just because learners PAY ATTENTION doesn't mean they learned. Therefore, measuring attention is an inadequate way of evaluating learning.

4. Just because learners SHOW INTEREST in learning doesn't mean they learned. Therefore, measuring interest is an inadequate way of evaluating learning.

5. Just because learners ACTIVELY PARTICIPATE in learning doesn't mean they learned. Therefore, measuring participation is an inadequate way of evaluating learning.

6. Just because learners say they LIKE A LEARNING EVENT doesn't mean they learned. Therefore, surveying learners on their general satisfaction—and on other factors not related to learning effectiveness—is an inadequate way of evaluating learning.

7. Just because learners REPORT THEY HAVE EXPERIENCED EFFECTIVE LEARNING METHODS doesn't guarantee they learned. Therefore, surveying learners on their experience with learning methods, while it can provide some good information on learning design, must be augmented with objective measures of learning.

8. Just because learners CAN RECITE FACTS AND TERMINOLOGY doesn't mean they know what to do. Therefore, measuring knowledge recitation is an inadequate way of evaluating learning.

9. Just because learners COMPREHEND A CONCEPT doesn't mean they will be able to use that concept in a work situation. Therefore, measuring knowledge retention is an inadequate way of evaluating learning.

10. Just because learners DEMONSTRATE A SKILL OR COMPETENCY during a learning event doesn't mean they'll remember how to use the skill or competency later. Therefore, measuring skills or competencies during or soon after a learning event is an inadequate way of evaluating learning.

11. Just because learners can MAKE RELEVANT DECISIONS doesn't mean they can perform successfully—they also have to be fluent, appropriate, and timely in the way they implement those decisions. Therefore, measuring for task competence (that is, measuring both decision making and action taking) is better than measuring for decision making competence alone.

12. Measuring DECISION MAKING COMPETENCE IS A BETTER METRIC than measuring knowledge recitation or retention. Measuring decision making is also better than gauging learner perceptions.

13. There are a NUMBER OF GOALS WE SHOULD HAVE as learning designers, including supporting our learners in building: comprehension, remembering, decision making competence, task competence, and perseverance in applying what they've learned to their job or other performance situations.

14. It is NOT SUFFICIENT TO SUPPORT LEARNERS ONLY IN COMPREHENSION, which is what too often passes for learning. At a minimum, we also need to support learners in remembering and in decision making.

15. MEASURING *ONLY* THE ULTIMATE GOALS of a learning intervention IS INSUFFICIENT because we will be left blind to factors that enable those ultimate goals.[1]

16. It's important to measure BOTH THE POSITIVE AND NEGATIVE impacts of our learning interventions.

17. In measuring the effects of learning, it's CRITICAL TO SEPARATE OUT FACTORS that are not related to the learning.

A good evaluation model will send messages highlighting these issues—or at least hinting at most of these issues. Let's take a look at the messages sent in the Kirkpatrick-Katzell model.

The Kirkpatrick-Katzell Model's Successes and Failures in Messaging

The four-level Kirkpatrick-Katzell model sends many important messages.[2] It tells us we should focus not only on learning, but on creating on-the-job behavior and organizational results. It hints that learner opinions are not sufficient and should be a lower priority than the other three levels. It tells us the four levels are interconnected. These are critically important ideas—especially the concept that our end goal is not learning; that learning is, instead, a means to an end. We owe a debt of gratitude to Donald Kirkpatrick and Raymond Katzell for this stealth message—that learning is not our end goal.[3] This is one of the most important ideas ever conveyed in the workplace learning field. The Kirkpatrick-Katzell four-level model was instrumental in getting the industry to accept this truth.

Unfortunately, the model also conveys—or is often understood as sending—other messages that are problematic. It sends the message that learner feedback is related to learning results, job performance, and organizational results—but research shows that this is not

---

[1] So, for example, if our ultimate goal is to produce improved customer satisfaction, it's not enough to measure how learning impacted that satisfaction. We also need to know whether the learners could make good decisions, whether they could remember what they learned, and whether they comprehended what they learned in the first place. To see why this is essential, suppose we did our homework and we determined that customer service is worse than it should be because employees have a knowledge problem—one that can be helped with an effective learning intervention. Suppose we train the employees but customer service does not improve. If we haven't evaluated decision making, we won't know whether our training was deficient in helping learners make good customer service decisions. If we haven't evaluated remembering, we won't know whether our training was deficient in supporting remembering. If we haven't evaluated comprehension, we won't know if our training was deficient in supporting comprehension. Given that there is a causal chain from comprehension and decision making to remembering to on-the-job application to customer service, we have to evaluate the full causal chain to know what happened.

[2] Raymond Katzell originated the idea of separating learning evaluation into four parts (Kirkpatrick, 1956). Donald Kirkpatrick enhanced Katzell's formulation by creating labels for each of the four levels. He also, for five decades, tenaciously popularized the use of a four-part model.

[3] It will certainly surprise many that Raymond Katzell formulated the four-part notion of learning evaluation. You can read more here: https://wp.me/p9hau4-1bj, which can be found on my blog: https://www.worklearning.com/wills-blog/)

true. Traditional smile sheets have been shown to be virtually uncorrelated with learning results and work performance.[4]

The Kirkpatrick-Katzell model also supports the connotation that measuring Level 3 (job performance) and Level 4 (organizational results) is more important than measuring Level 2 (learning results)—even though we learning professionals have much more leverage over learning results than we do over job performance and organizational outcomes. This idea is especially problematic when we interpret the model to mean we ought to measure *only* Level 3 job performance or Level 4 organizational results. It's problematic because it ignores the causal pathway starting at learning and moving to job performance and then organizational results. If we measure only Level 3 and 4 and get poor results, we will have no idea how our learning design contributed to these failures.[5]

The four-level model is centered on training—hiding the role prompting mechanisms (job aids, performance support, etc.) and on-the-job learning might play in learning and performance. If what gets measured gets managed, then we're not likely to even consider prompting mechanisms or on-the-job learning as targets for our evaluations. Indeed, until recently these non-training modalities were largely ignored in the workplace learning industry.

The Kirkpatrick-Katzell model ignores the causal chain from learning to remembering to performance to results. More specifically, it has absolutely zero messaging about the critical importance of remembering—or that learning interventions can be good at supporting comprehension and poor at minimizing forgetting. Ideally, we need a model that pushes us to use learning designs that support long-term remembering.

The Kirkpatrick-Katzell model pushes us to focus on weighing outcomes, while it is largely silent about measuring the various learning supports critical to good learning outcomes. Most pointedly, the model ignores the role non-learning professionals—especially managers—can play to support the learning-to-performance cycle.

The four-level model provides no leverage points for learning-design improvements. Indeed, with its focus on learner perceptions as a causal factor in learning results, it has distracted us from learning factors that really matter. Ideally, we'd want our learning evaluation model

[4] Alliger, Tannenbaum, Bennett, Traver, & Shotland, 1997; Sitzmann, Brown, Casper, Ely, & Zimmerman, 2008.

[5] I am somewhat intrigued by Roy Pollock's idea (personal communication November 2017) that maybe we should measure work results routinely—and look at other learning results only if something goes wrong. I'm still uncomfortable with this as I think we need more feedback on our everyday performance as learning professionals, not less. But maybe there is a compromise that might work—perhaps a performance-focused learner feedback survey and a valid measure of work performance as Roy suggests.

to encourage thoughtful reflection on how to improve our learning designs—pointing us to the importance of remembering and decision making competence, for example.

The Kirkpatrick-Katzell model also fails to warn us about some of the worst evaluation practices—common practices like giving learning interventions recognition of achievement when learners attend, pay attention, complete, or participate in learning. For too long, many in the learning profession have used these irrelevant metrics as indicators of learning. A good learning evaluation model would disavow these ineffectual evaluation practices.

No model is ever perfect, and—given the gargantuan complexity of human learning—no learning evaluation model is likely to contain all the messaging we might want. In the new model, *The Learning-Transfer Evaluation Model*, some of the above limitations are overcome, but many are left for us to consider outside the parameters of the model. Raymond Katzell and Donald Kirkpatrick, working in the 1950s, were not privy to today's advances in the science of learning, nor could they have foreseen the evaluation practices that would be popular in the next century. They built their model in response to practices of the 1950s. The key insights baked into the Kirkpatrick-Katzell model must be retained in any new model, inappropriate messages must be removed, and critical missing messages must be added where possible.

The following two lists summarize the benefits and harms of the Kirkpatrick-Katzell model:[6]

Beneficial Messaging Inherent in the Kirkpatrick Four-Level Model:

1. Tells us we should focus not just on learning, but we should also aim for on-the-job performance and organizational results.
2. Implies that learning is instrumental in achieving improved behavior and results.
3. Hints that learner opinions of the learning are not sufficient on their own, and should be viewed as a lower priority than the other outcome measures.
4. Implies that learner opinions, learning results, on-the-job behavior, and organizational results are interdependent.[7]

[6] The Kirkpatrick New World Model introduces some minor tweaks, but generally conveys the same messaging as the original. Indeed, with its recommendation to think first about results and work backward, it's not so much an evaluation model as a learning development model.

[7] This is mostly a beneficial message except that learner perceptions have been shown *in practice* to be virtually uncorrelated with learning results and on-the-job behavior. To be specific, the beneficial message is that Level 2 (learning results), Level 3 (on-the-job behavior), and Level 4 (organizational results) are interrelated. On the other hand, perhaps new methods—like the performance-focused learner feedback approach—may enable a reliable positive relationship between learner perceptions and the other learning outcomes.

Harmful Messaging Inherent in the Kirkpatrick Four-Level Model:

1. It fails to disabuse us of ineffective learning evaluation practices such as giving credit for attendance, completion, attention, and participation.
2. It ignores the causal chain from learning to remembering to performance and results.
3. It ignores the critical importance of supporting remembering and minimizing forgetting in our learning designs.
4. It ignores and fails to highlight the advanced competencies of decision making and task performance.
5. It prompts us to value learner ratings as predictive of learning, work performance, and organizational results—something proven to be false as typically practiced.[8]
6. It implies that Level 3 and Level 4 (job performance and organizational results) are more important than measuring learning—even though we, as learning professionals, have greater leverage over learning.
7. It is also too easily interpreted to mean we don't need to measure learning results because job performance and organizational results are more important—thus ignoring the fact that learning is on the causal pathway to job performance and organizational results.
8. It pushes us to focus on learning outcomes and ignores the measurement of learning supports, including such things as situational cues, resources and time, cultural support, and management support.
9. The model is training centric, ignoring prompting mechanisms (like job aids and performance support) and on-the-job learning.
10. It conveys the idea that the only goal of learning should be organizational results—ignoring learning outcomes that affect learners directly and ignoring other potential beneficiaries and injured parties.

[8] But see the caveat in the previous footnote.

The New Model

The new model, *The Learning-Transfer Evaluation Model* (which we might pronounce "L-tem" and write LTEM), is designed specifically to help organizations and learning professionals determine whether their evaluation methods are effective in providing valid feedback. Previous evaluation models have not done this sufficiently.

LTEM is composed of eight levels—starting from completely inadequate methods of learning evaluation all the way through to the effects of learning transfer. The model is made available in a single-page, tabular format to ease comprehension. It is annotated and color coded to enhance the clarity of its messaging.

The model is designed to be relevant for all learning interventions, including classroom learning, elearning, mobile learning, on-the-job learning, self-study learning, etc. Where learning occurs, we can evaluate it. We may have to change terminology or modify our evaluation practices, but the new model is designed to help bring wisdom to our efforts.

The model, while an exponential improvement over the Kirkpatrick-Katzell four-level model, still leaves some important messages unstated. For example, it does not focus on evaluating prompting mechanisms like job aids or performance support. This is intentional. When models get too complicated, people don't use them.

Each of the levels of the model will now be described.

Level 1—Attendance

Believe it or not, many organizations will certify learning by checking attendance. This is a terrible idea because attendance does not guarantee that learning occurred. Learners have been given credit for signing up for courses, starting courses, sitting through learning events, and completing learning experiences. None of these attendance metrics is acceptable as a measure of learning.

Level 2—Activity

Too often, we use learner activity as a gauge of learning. We do this formally and informally. For example, we may formally measure the number of words a learner uses in an online forum. We may informally gauge learning success by the "energy in the room" when learners are involved in small-group discussions.

There are three main ways we measure learner activity: (1) attention, (2) interest, and (3) participation. All three are unacceptable as measures of learning because our learners may pay attention, show interest, and participate actively—yet not learn.

What are some examples? Some organizations give credit in elearning programs when learners are periodically presented with a button to click to show they are paying attention. Other elearning programs measure attention by monitoring timing of mouse clicks. Measuring attention is an unacceptable metric because attention does not guarantee

learning. Instructors, whether online or in a classroom, often give credit for class participation. Sometimes peers are asked to evaluate each other on various activity criteria. Rarely, but sometimes, outside observers evaluate learner activities. Note that activity metrics go beyond mere attendance; instead, they attempt to gauge such things as how active, how engaged, how helpful, or how cooperative a learner appears to be during learning. While learner activity may seem to be indicative of good learning design, it is inadequate because activity does not demonstrate learning results. Learners may fail to learn. They may learn the wrong things. They may learn things poorly. They may focus on low-priority ideas instead of critical constructs. They may fail to be prepared to use what they've learned. Some activities are useful. Some are a waste of time—even though they seem useful. Some activities are harmful.

With today's digital technologies, it is tempting to count learner activity using objective metrics (for example, the number of words written on a discussion board, number of social-media posts, or quality of written concepts as gauged by artificial intelligence). While these measures may remove the subjective element, they still suffer from the same problem as other activity metrics. People can be objectively active in learning, yet still fail to learn, or learn poorly, or learn the wrong things.

Level 3—Learner Perceptions

Learners are often queried about their perspectives on the learning experience. We do this both formally and informally. We provide learners with formal surveys (aka smile sheets, happy sheets, feedback forms, etc.). We also might ask a classroom of learners to share what they think about the learning they've just completed. Even more informally, we might wait to hear what the "word on the street" is about the learning—either from the learners themselves, the learners' managers, or even from other employees.

Most commonly, we formally survey learners about their satisfaction—about whether they'd recommend the course to others, whether they felt the instructors were competent, whether the course met their expectations, etc. Unfortunately, such traditional smile-sheet questions have been found to be virtually uncorrelated with learning results.[9] While such surveys are useful to gauge the reputation of the learning experience, these smile sheets cannot be recommended as a measure of learning.

Learners can be asked questions that more fully gauge learning effectiveness. In my book, *Performance-Focused Smile Sheets* (https://SmileSheets.com), I detail new kinds of questions that can be asked—questions that focus on research-based factors such as whether the learning supported comprehension, remembering, and motivation to apply what was learned, and supports for after-training follow-through. Despite these

[9] Alliger, Tannenbaum, Bennett, Traver, & Shotland, 1997; Sitzmann, Brown, Casper, Ely, & Zimmerman, 2008; Uttl, White, Gonzalez, 2017.

improvements in survey methods, surveys focused on learning effectiveness create proxies of the constructs they target. For example, surveying to determine the level of realistic practice learners receive is just a proxy for whether the learning is supporting long-term remembering. Ideally, we'd measure remembering directly—as encouraged in the higher levels of this model. Surveying for effectiveness is designed based on learning science findings. More research is needed to verify that surveying for effectiveness is related to the targeted constructs in the many contexts where learner feedback is gathered. As of now, it seems fair to say that surveying for effectiveness may provide adequate data—at least for further hypothesis testing—but it should be periodically augmented with objective measures of learning effectiveness.

Level 4—Knowledge

Knowledge is often tested to ensure learners have learned. Knowledge can be tested soon after learning or after a substantial delay—after several days or so. This dichotomy is critical because remembering is critical to most real-world learning. When learners are tested soon after learning, we can call this process *knowledge recitation*. When learners are tested after a substantial delay, we can call it *knowledge retention*.

Knowledge recitation is focused on asking learners to answer questions related to facts and terminology—during or right after they learn the information, when the information is highly accessible from long-term memory. Information tested under these conditions is highly retrievable for two reasons. Learners have not had time to forget and they are likely to be triggered for retrieval by the learning context.

Knowledge recitation has two crippling problems as a learning evaluation measure. First, knowing facts and terminology is unlikely, on its own, to enable decision making and task competence. Knowledge does not necessarily enable performance. Second, reciting information after just learning it is an untenable metric because immediate recitation does not guarantee the ability to retrieve information later. For both reasons, knowledge recitation is an inadequate method for evaluating learning.

Both knowledge retention and knowledge recitation focus on facts and terminology, but knowledge retention also requires learners to remember the learned information over a substantial time period, which I've set at several days or more.[10] Knowledge retention, because it uses a more authentic retention interval than knowledge recitation, is a

[10] Over the years, as I studied the research on learning, memory, and instruction, I noticed that many learning experiments distinguishing between short and long retention intervals used week-long delays as the longer interval—and, more importantly, that relatively long delays didn't show much differentiation after a week or more. Certainly, people forgot more over time, but the experimental findings don't tend to differ significantly for, say, a one-week delay versus a two-week delay (when both are compared to a one-day delay, for example). To be practical, I've set longer durations at "several days or more"—meaning three or four days or more—because it's my assumption that many workplace learning stakeholders would balk at routinely using week-long or longer delays in their learning designs and/or evaluations.

marginally better evaluation method, but it still suffers from the fatal flaw of focusing on knowledge of facts and terminology—knowledge that is likely to be largely insufficient on its own in enabling decision making and task competence. For this reason, knowledge retention is an inadequate metric for learning evaluation, except in those rare cases when it's a person's job to recall facts and terminology. For example, supermarket cashiers who need to memorize product codes might be properly evaluated using the ability to remember that knowledge.

Knowledge, of course, is not all one thing. Some knowledge is composed of facts and trivia. Other knowledge is composed of more meaningful constructs, including concepts, principles, generalizations, processes, and procedures. Indeed, we could break down knowledge into myriad subcategories. The various Bloom's Taxonomies, including the original from 1956 and the various updates in the 2000s, show the multitude of ways knowledge can be subdivided.[11] Interestingly, Benjamin Bloom and his colleagues created their taxonomy based on actual evaluation tools that were used in schools and universities.

None of the Bloom's Taxonomies—not Bloom's, not Marzano's, not Anderson and Krathwohl's—are useful for our purposes. First, they are too complicated; second, and more importantly, they don't focus on the future utilization of learning. That is, they don't focus on transfer. By keeping a laser-like focus on transfer, we are better positioned to have an evaluation model that is practical and relevant to our targeted outcomes.

While we don't have to overcomplicate our evaluation of knowledge, we should focus on the most meaningful aspects of the knowledge that is targeted for learning. Instead of focusing on definitions and terminology, we should focus on the concepts, principles, and wisdom that relate most directly to the learners' future use of the information. We should also keep in mind that we can evaluate procedural knowledge in the decision making and task competence levels of the LTEM framework—Levels 5 and 6.

A Special Note on Realistic Remembering

I'm making an intentional dichotomy between short-term remembering and relatively long-term remembering. This distinction is one of the most critical concepts in learning. If we prepared our learners only for the short term, much of their training would be wasted. They simply wouldn't remember what they learned; and, when they needed the information in their work, they wouldn't retrieve it from memory. Too much training is built just like that—creating only short-term effects—which is another reason this dichotomy is so essential.

You may wonder if this applies when we teach people and they begin to apply what they've learned the very next day or shortly thereafter. Let's say we have a one-day workshop and we teach people 20 key concepts. Let's say the learners take our workshop on Monday and begin to apply what they've learned on Tuesday. How many of those 20 learning nuggets are

[11] Bloom, 1956 (or Bloom, Engelhart, Furst, Hill, Krathwohl, 1956), Marzano, 2001, Marzano & Kendall, 2007; Anderson & Krathwohl, 2001.

they likely to use that first day? Maybe 10. How many the rest of the week—on Wednesday, Thursday, and Friday? Maybe five more. Already we can see—even for this seemingly just-in-time learning intervention—that learners had to remember 50% of what they learned longer than a day, and 25% of what they learned longer than a week. The conclusion from this thought experiment should be clear. Only rarely do learners immediately use everything they've learned—mostly when they learn only one or two things and utilize them right away. The truth of it is, for almost all our learning interventions, our learners will have to remember a significant portion of what they learned.

This dichotomy between short- and long-term remembering plays out in three places in the LTEM model: In the distinction raised in Level 4 between knowledge recitation and knowledge retention, in Level 5 between decision making and remembered decision making, and in Level 6 between task competence and remembered task competence. In each case, the remembered competence is closer to the time lags learners will experience in their real work situations—when remembering is essential to their performance.

To be practical and somewhat specific about this, testing for short-term remembering should involve measuring within the same day in which learning took place (i.e., within about eight hours). Long-term remembering can be operationalized to involve delays of three days or more after learning—to be practical—although delays of a week or more might be more realistic.

Level 5—Decision Making Competence

In the four-level Kirkpatrick-Katzell model, Level 2 was designated as "Learning." Too often, "learning" was interpreted as "knowledge recitation"—asking learners to regurgitate facts and terminology they had just learned. But learning results are so much richer. At a minimum, learning results can constitute comprehension, the ability to make decisions, and the ability to successfully engage and complete realistic tasks.[12]

Decision making is clearly an essential result of any learning initiative that intends to support subsequent behavior. When we teach managers to be better supervisors, we hope they'll make better *decisions* when they are interacting with their direct reports. When we teach employees to avoid engaging in sexual harassment, we hope they will make better *decisions*

[12] Once, intending to lovingly augment the Kirkpatrick-Katzell four-level model, I designated a Level 2.5 as learner decision making. Of course, decision making is not the only augmentation needed. I've already mentioned task competence, and perhaps there are other learning results we might consider as well. NOTE that, while we might consider adding "motivation to apply" as another learning result, there are good reasons to avoid making it a separate evaluation target. For those interested, I didn't add a separate level to the new model focused on "motivation to apply" because motivation (1) is really hidden inside human cognition, (2) is thus not directly measurable, and (3) can be indirectly measured through learner surveys, which are already represented in the new model.

about what to say and what to do. Even in purely educational settings, we should design our learning to help learners make better *decisions*. When we teach algebra, we hope our learners will be able to *decide* when to reduce like terms to one instance in an algebraic expression. When we teach history, we hope—or we should hope—our learners will be able to make better *decisions* when they enter the voting booth.

Training that is aimed only at creating awareness—without any expectation that behavior should change—is training that is likely a waste of time. Learners may become attuned to important concepts, but they are likely to (1) forget what they've learned (because of a lack of realistic practice), (2) be unprepared to use what they've learned (because cognitive links have never been made between real-world situations and the learned concepts), and (3) be insufficiently motivated to apply what they've learned (because they haven't considered the real-world implications).

At a minimum, we should design learning to support realistic decision making and we should evaluate our learning interventions based on how well they support learners in gaining decision making competence.

Given the imperative to take short- and long-term remembering into account, we can measure decision making competence in the short term (within the day it was learned) or in the long term (for example, after three or more days).

Measuring short-term realistic decision making is a giant leap from earlier levels, but it is still insufficient as a learning metric because it does not evaluate whether learners can maintain their newly-learned decision making competence over time—to support them in using their skills on the job.

Measuring long-term realistic decision making is an important benchmark for learning professionals. Indeed, moving up the model from the bottom, this is the first measurement method we can rightly use to certify a competence related to work performance! This is huge! And having such a designation in our evaluation model is critical to ensuring we are measuring and designing for real work performance.

Examples of How to Measure Decision Making Competence

To evaluate realistic decision making, we need to present learners with realistic situations and prod them to make decisions that are similar to the types of decisions they will have to make on the job. There are two (or three) ways to do this.[13] We can present learners with realistic scenarios and ask them to make realistic decisions based on their interpretation of the scenario information. We can ask learners to make decisions in high-fidelity simulations. Or we can ask learners to make decisions in real-world situations—for example, on the

[13] Here we are indebted to Sharon Shrock and Bill Coscarelli, whose evaluation taxonomy—introduced in their classic text, *Criterion-Referenced Test Development*—utilized three categories of realistic performance: (1) Real-World, (2) Simulation, and (3) Scenario-Based Decision Making.

job.[14] Note that real-world decisions on the job will often push learners into the next evaluation level—Level 6 (task competence)—because, in the real world, we don't usually just make decisions; we usually make decisions and take actions together.

Scenarios can be text based or can involve video and audio. They can involve forced choice responding, as with multiple-choice questions. They can involve multiple-option responding. They can involve open-ended responding or some other simple response method. Scenarios can be short and simple or they can be long, or threaded, as if in a story, or they can involve multiple layers of analysis before decisions are made—like in case studies.

Simulations go beyond scenarios by providing more realistic contexts and often more realistic responding as well. Note that there is not always a definitive line between scenario-based metrics and simulation-based metrics. Often, in practice, things called "simulations" are really fancied-up scenarios. It doesn't matter much for our purposes. The key is (1) whether learners are provided with realistic background contexts, (2) whether they must decipher what the context means (without help or hints—because we're assessing here!), (3) whether they have a range of options from which to choose to indicate the decisions they are making, and (4) whether they are compelled to actually make a decision.

Bridge between Decision Making Competence and Task Competence

The distinction between decision making competence and task competence is complicated and important, so I'm going to provide a long discussion to introduce it.

Task competence comprises decision making, as exemplified in Level 5, but it also adds task performance. Here is an example. Suppose we teach managers to give performance feedback to their direct reports and then we assess their decision making ability in knowing what to say and when to say it. If we do that, we're measuring their decision making competence. But then, suppose we go further and assess their ability to voice the feedback (what tone of voice and body language to use). If we do that, we're measuring their ability to take action after they've made a decision. By measuring both decision making and action taking, we're measuring task competence.

When we teach, we often separate decision making and task competence. For example, if we're teaching electricians, we can give them practice trouble-shooting circuit problems using a diagram that represents electric circuitry. This is decision making practice. They are not taking action. That is, they are not finding real wiring to test, using tools in a realistic

[14] Note that, whereas scenario and simulation decision making are decidedly learning situations, if we ask our learners to make real-world decisions, this might seem to jump our evaluation up to Level 7 in the model, Transfer. However, although there is some ambiguity in the case of learners performing in real-world situations, asking our learners to engage in real-world decision making for the purpose of learning has a different psychological character and can therefore be appropriately separated from regular on-the-job decision making (where learners have left the formal learning experience and have transferred their learning without our direct intervention).

way, et cetera. Similarly, we can give salespeople practice deciding what to say when faced with customer objections without having them actually voice the words they would use or the tone of voice they would employ. We can give nurses practice deciding what to do when faced with simulated patient information, without having them actually take a person's blood pressure and other vitals. If we measure this decision making ability alone, we are doing Level 5 work. Again, Level 5 measurements ask learners only to make decisions, not to implement those decisions.

Conversely, we can also give people practice on task performance without asking them to make decisions. We can say, "Okay—role play what you'd say to an angry customer who is upset at having to wait in line." In such a situation, we have essentially made the decision for the learners. They haven't had to decide that the customer is angry because of a long line. Similarly, we can train a nurse on how to listen through a stethoscope to the sounds of a person's airways without having them decide when to use such a procedure.[15]

To reiterate, we can focus on decision making, action taking or both. In LTEM, when only decisions are measured, we're in Level 5; when decisions and actions are measured as full tasks, we are in Level 6.

Special Note to Introduce the SEDA Model

The SEDA Model helps us clarify the distinction between decision making and task performance. Originally developed to help learning designers write powerful scenario-based questions, the SEDA Model is a practical tool for multiple learning design purposes. It works because it is based on a simple truth about human behavior.

Consider what people do in their real-world experiences. They are constantly faced with *Situations*. They *Evaluate* those situations—in other words, they make sense of them. Once they understand a situation, they make a *Decision* about it. Then they take an *Action*. This can be diagrammed as follows:

Situation → Evaluation → Decision → Action

SEDA is the acronym formed by the first letters of the words "Situation," "Evaluation," "Decision," and "Action." The SEDA Model, then, nicely simplifies what people do in the real world. Each action creates a new situation. Each new situation provides feedback on the previous action, decision, and evaluation.

[15] I should note, for precision, that—when we attempt to determine what constitutes a decision and what constitutes an action—there will always be some subjectivity. In truth, there are different levels of abstraction at play. For example, I labeled "a nurse using a stethoscope" as an action, but the nurse may have to decide how to hold the stethoscope or where to place it, et cetera. Don't worry; in practice, the distinction between decisions and actions is usually straightforward.

When we design learning, the SEDA Model can be our guide. We know—from research on how context influences the learning-to-remembering process—that real-world situational cues trigger thinking and action. Human cognitive processing is reactive largely to triggers, whether those triggers come from a person's environmental context or from their internal working memory. SEDA, although a simplification of human cognitive processing, captures the basic elements needed to guide learning design. The key for learning is ensuring learners practice every aspect of the SEDA Model. We need to place learners in realistic Situations (whether they be real-world situations, simulated situations, or scenario-based situations). We need to require our learners to Evaluate those situations, ensuring they make sense of those situations on their own. We need to prompt learners to make Decisions based on their evaluations, and take Actions based on their decisions.

This sounds simple, but it's often easier to comprehend the value of SEDA by looking at failures in learning design. I will do this quickly so we can get back to our focus on evaluation.

I'll start with one of my own failures. I used to teach simulation-based workshops on leadership topics. I taught content, then put learners in elearning-based simulations where they made decisions based in realistic leadership scenarios. Let's examine this by using the SEDA Model. I presented my learners with Situations, which they Evaluated (made sense of) before making Decisions—decisions about what to do in those simulated situations. What they *never* did in my workshops was practice the Actions they would have to undertake to implement their decisions. They never practiced what words to say, what tone of voice to use, what body language to emanate. We might say the learning design was a SEDA (with a grayed-out A to signify that learners did not practice Actions).

What about an elearning-based simulation that teaches learners how to diagnose a medical episode? *"An older middle-aged man enters the ER. He's an alcoholic and is complaining about lower abdominal pain."* Learners are asked what to do. Using SEDA, we would be presenting a Situation, but would NOT be giving learners a real chance to Evaluate that situation. Did you notice that the simulation told the learners the man was an alcoholic? The simulation basically evaluated the situation for the learner. We might grade the experience as a SeDA, using a lowercase "e" to signify a partial Evaluation and a grayed-out A to signify that no actual actions were practiced—that is, learners never had to actually examine a person's body, call in a prescription, et cetera.

What about a statistics class that teaches one topic a week for several weeks? So, for example, we might teach Chi-Square in the first week, t-tests in the second week, f-tests in third week, and so on. During the week, we give learners extensive practice calculating the statistics. At the end of each week, we give learners a test asking them to calculate the statistic taught that week. Using SEDA, we clearly have given learners plenty of practice in taking Actions—doing the math to calculate statistics. What we have *failed* to give them are any real-world Situations that need to be Evaluated. Nor have we given them practice in

making Decisions about which statistic to use—and when. In the real world, people using statistics are faced with some sort of research effort to work through. They have to Evaluate those Situations and Decide which statistic to use. We would have to grade this statistics class as SEDA, with a grayed-out S, E, and D. Note how harmful this design is! Learners learn procedures for calculating statistics—something they will *never* do in the real world, by the way, because statistical programs do this grunt work. Moreover, they never actually practice the most important skill—figuring out which statistic to use given a variety of background scenarios depicting different research situations.

One more—and this one's the killer. What about lawyers learning the law or students in a marketing class learning marketing concepts or employees being onboarded by learning about the history and values of their new company? Let's say these learners are provided with scintillating speakers who cover concepts and highlight important principles. Using SEDA, we can see that the learners were *NEVER* presented with Situations—and, of course then, they never had to Evaluate situations, make Decisions, or take Actions. They got no practice in any of the SEDA components! An empty SEDA—a ghost SEDA—and, incidentally, a very poor learning design: one that leaves learners with zero skills.

I'm not arguing that every learning event must incorporate all four components of SEDA. While that may be ideal—and, while SEDA may help us in aiming for full learning effectiveness—we can also use SEDA to help us keep track of the SEDA components we are providing and those we are missing. For example, going back to my leadership simulation days, I might have used SEDA to see that I was missing the Action component. I could have then redesigned the course to include more action practice or I could have ensured learners had job aids or performance support or additional practice before they attempted to implement the decisions they had learned to make.

With that brief introduction to the SEDA Model, it should be clear why, at a minimum, we need separate levels for Decision Making Competence and Task Competence.[16] We want our evaluation model to prod us to create learning designs that are fully effective. By focusing on decision making in Level 5, the new model promotes Evaluation and Decision Making in our learning designs. By focusing on Task Competence in Level 6—and remember: Task Competence includes both decision making and action practice—the new evaluation model promotes the use of all four components of SEDA. In other words, Level 6 provides practice in all the real-world behaviors needed for our learners to be fully effective in their work or in other targeted performance situations.

[16] In the *Learning-Transfer Evaluation Model*, I'm including the SEDA Situation, Evaluation, and Decision components within the Decision-Making Competence levels. I'm adding the SEDA Action component to create the Task Competence level. Therefore, the Task Competence level of the model includes all four SEDA components. It presents learners with Situations and has them Evaluate those situations, make Decisions about what to do, and take the Actions determined by those decisions.

Level 6—Task Competence

We can now focus on Level 6—Task Competence.

Let's start with task competence that is demonstrated soon after learning. If our learners can demonstrate task competence right after learning, they are well on their way to transferring what they've learned. Demonstrating full task competence is a very strong learning result. But there is one catch. If learners demonstrate task competence soon after learning, we still don't know whether they can remember that competence over time. Measuring task competence during or soon after learning is not sufficient to enable us to certify learners as prepared for transfer.

On the other hand, if task competence can be demonstrated several days or more after learning, we can fully certify learners as task competent. Again, this doesn't mean they have transferred learning. They haven't. They've merely shown they can perform tasks while situated in a learning context. Actual transfer is represented in Level 7.

Examples of How to Measure Task Competence

To measure task competence, we have to provide learners with a full SEDA experience. We need to present them with realistic Situations, have them Evaluate those situations (again, without help or hints—because this is an assessment!), enable them to make Decisions, and have them take Actions in line with those decisions.

If we're teaching people to use PowerPoint® so they can make effective presentations to their learners, we'd provide them with PowerPoint, some content to teach, some information about the learners, and some specific learning goals to achieve. We might then have them create a set of PowerPoint slides and have them make a training presentation to some simulated trainees (maybe their fellow learners, class instructors, or other learning professionals). In doing all this, we've presented them with a realistic situation (i.e., PowerPoint, training goals, learners) and had them Evaluate all the requirements, Decide on the design of the slides and the learning flow, and take Action in using what they've created.

If we're teaching supervisors how to run a meeting, we could evaluate their performance running a simulated meeting with actors. If this is too cumbersome or expensive, we could also use the SEDA Model to design alternatives. If we had to keep costs to the bare minimum, we could create scenario-based questions that described meeting situations and have our learners evaluate those situations and make decisions about what to do. We would have a SEDA result that provided Situations, Evaluation of those situations, and Decisions. What we wouldn't have was an assessment of how well people could take Action. So, we might then decide to add another evaluation element. We could develop short video clips of meetings with a point-of-view perspective that makes the learner part of the meeting and requires them to voice their responses at certain times during the simulated meetings. Responses could be video recorded and evaluated based on criteria taught in the learning.

Ideally, we might use experts, including instructors, to evaluate these action performances; but, in some cases, it might be acceptable to use peer evaluations to save time and money.

As you can see, evaluating task competence is much more demanding than measuring at lower levels, but also more aligned with real-world competence.

Level 7—Transfer

Learning transfer is defined by two criteria. First, people have had to previously engage in some sort of learning experience. Second, they have to use what they've learned on the job (or in some other targeted performance situation).

Special Note on the "Targeted Performance Situation" Idea

It's easy for us to think about someone putting their learning into practice on the job. But people can put what they've learned into practice in other ways as well. A person who takes a leadership class may use what they've learned in helping to organize a local charity or run a youth soccer team. These transfer successes may not have been the goal of the original learning investment, but they do represent transfer nonetheless.[17] For this reason, I've tried to use the word "work" instead of the word "job" in the optimistic hope that "work" can be broadened to include other endeavors besides on-the-job activities. Aren't we doing work in calculating an algebra problem, in practicing a foreign language, in learning to bake bread?

Not all learning investments are designed with job performance as the goal. Education settings are the clearest example of this. College students may learn algebra so they can learn trigonometry and calculus. High school students may learn history and social studies so they can make better life choices or vote for better elected officials.[18] So, for an algebra class, the targeted transfer situation might be the students' future trigonometry class. For history, one targeted transfer situation might be the voting booth; others might include future history classes, discussions with friends related to history or politics, or reading historical fiction.

I've developed the *Learning-Transfer Evaluation Model* to work for both workplace learning and education settings. One warning here—because my career has focused mostly on workplace learning, I may have unintentionally framed things more readily for workplace learning. Still, as I worked on the model—and this report—I did keep one eye open to the world of education. I'm confident the model has relevance for education as well as workplace learning. Indeed, I think students would benefit if more teachers and professors kept their learners' future performance situations in mind.

---

[17] We should probably separate transfer that was targeted in our learning design from transfer that is incidental or a lower priority. We typically invest in learning to gain some sort of benefit. We can certainly have multiple goals, but we might want to be clear about what our bottom line is for successful transfer before we design and develop our learning initiatives.

[18] Insert your own joke here.

By designing for transfer, we create learning that is more likely to be remembered and utilized. What's the value of learning something if that knowledge is locked in our brains, never used—fading into fuzziness over time? Educators who want to use the model may have to occasionally reframe some of the words or the emphases, but the model is intended to have the same benefits for education as it does for workplace learning and performance.

When we measure for transfer, then, we must select a relevant performance situation to target. We don't willy-nilly want learners to use what they've learned. We really hope they use what they've learned in certain situations or contexts. When we provide leadership training, we hope supervisors will use what they've learned in situations in which they are interacting with their direct reports. When we teach PowerPoint to help people craft more effective presentations, we hope our learners will use what they've learned while they are using PowerPoint to make presentations.

There are two types of transfer: assisted transfer and full transfer. Assisted transfer connotes situations where a person transfers their learning to the job, but does so with significant assistance, support, or prompting. A person who transfers their learning to the job because their manager strongly urges them to do so would achieve assisted transfer. A group of learners who transfer their learning to the job only because their course instructor is acting as an after-training coach would achieve assisted transfer.

Of course, most transfer is assisted in some way—the learner is given permission or time or resources—but the distinction between assisted transfer and full transfer is more subtle. Assisted transfer represents situations where the learner would have been unlikely to engage in transfer on their own without assistance, support, or nudging. Assisted transfer and full transfer achieve the same ends, but assisted transfer is more tenuous, more dependent on significant help from other people.

Full transfer occurs when a person takes what they've learned and successfully puts it into practice in their work—without *the need* for significant help or prodding. Even if they got assistance, support, or nudging—if they would have put their learning into practice even without these interventions—we would designate the result as full transfer.

Let me emphasize again that both assisted and full transfer represent complete success in transfer! Indeed, we might argue—when an individual is forced to work with others to achieve transfer—that *more* benefits are created, including such benefits as team building, social impetus to persevere, interpersonal bonding, and socially-mediated learning. Full transfer gets a slight edge in status because it indicates the learner was fully prepared to transfer learning to work without additional intervention.

Level 8—Effects of Transfer

Transfer is the penultimate goal of learning, as learning is almost always instrumental—a means to other ends. Even when we learn for the pure joy of learning, we can consider joy to be an outcome of learning. Given learning's instrumentality in creating other outcomes, our evaluation efforts should assess learning outcomes.

For almost all workplace learning efforts, we invest in learning because we hope it will be instrumental in achieving other ends. We train nurses not just so they transfer their skills to their work, but because we expect to get better patient outcomes. We train managers in leadership skills not just so they perform better as managers, but because we hope their teams will achieve better results (for example, increasing product or service quality, reducing costs, and increasing employee morale and productivity). Even when we just provide training as an employment benefit—one that is intended to directly benefit our employees—we hope such organizational generosity might translate into increased loyalty, lower turnover, recruiting benefits, etc. The notion that learning should be instrumental is one of the great features of the Kirkpatrick-Katzell model. Level 4 (Results) is a clear statement that learning ought to achieve other ends.

Some have modified Kirkpatrick-Katzell's original Level 4 label "Results" and changed it into "Business Results." Of course, business results are not the only outcomes we might hope for. Roger Kauffman, in particular, has emphasized the potential for assessing learning's impact on society. I've emphasized—in my *Learning Landscape Model*—the importance of acknowledging the outcomes that affect the learner. In my mind, it seems we've had a blind spot to this essential stakeholder in the learning process. This is especially apparent when we consider the psychological investment learners often make—first, challenging themselves to overcome their own weaknesses and limitations, then dealing with anxiety and resistance and blowback as they champion new practices in the workplace.

Transfer can affect many people and things. If we're conducting an evaluation, we ought to at least generate a list of the potential stakeholders and impacts. Here's my list. Learning transfer can affect (a) learners, (b) coworkers/family/friends, (c) the organization, (d) the community, (e) society, and (f) the environs.

Learning transfer can produce benefits and harms. Too often, we focus only on the benefits of our learning interventions. We should also examine the downsides. We ought to look at time usage and productivity loss. We ought to look at the environmental damage from flying people across the country or the globe. We ought to look at the damage caused when we teach people nonsense. In the learning field, we harmed a whole generation of learning professionals (and their learners and organizations) because we taught learning-styles snake oil. Our exuberance in the early days of social media (for learning) led to bad information being shared online between our employees. We ought to look at the damage caused when we create compliance training for legal purposes that unintentionally sends messages that the organization is just covering its ass—that maybe having learners click next through an

elearning course is all we need to do. We ought to look at the harm we might be doing to our fellow citizens when we teach union-busting tactics or teach managers how to squelch employee requests for better working conditions. We ought to examine the dangers of content that has no scientific backing or no vetting by true experts. We ought to calculate the waste that occurs when we throw training at non-training issues.

Enough! I made the point. We have a responsibility to think broadly about our learning impact and consider both benefits and harms.

Creating beneficial transfer effects usually requires more than just our efforts as learning professionals. Our learners have to take responsibility for transfer. Their managers and coworkers have to lend support or be cooperative. Organizational units might have to change practices and policies to enable success. Where training is part of a larger strategic initiative, many other factors will come into play.

To fully vet transfer effects, we have to use rigorous methodologies to ensure we can separate out the learning effects from other factors that may influence the target results. This is not always easy or even possible.

Let's say an organization provides training to call center reps with the goal of (a) reducing the length of phone calls and (b) increasing customer satisfaction with those calls. At the same time, the organization redesigns its product line to make it simpler and more user friendly. The result: The number of calls to the reps drops, as does the time needed in the calls—and customer satisfaction soars. Did training make a difference? Unfortunately, it will be impossible to tell which factors produced the results unless we had earlier decided to use a control-group design and withheld training from some of our reps. By comparing those reps who had gotten the training to those who hadn't, we'd be able to tell what effect the training had. This type of control-group design might be worth it—particularly in some strategically-important initiatives or to test our general training methodology—but, in most cases, organizations will not take the risk of withholding training from employees.

To summarize the key requirements of Level 8, Effects of Transfer, we should (a) consider learning outcomes that affect an array of stakeholders and their environs, (b) look for both benefits and harms, and (c) use rigorous methods to ensure we're able to draw conclusions about the causal impact of the learning experience itself.

Using the Eight Levels

We've now looked in depth at each level of the new model—the *Learning-Transfer Evaluation Model*. Some of the levels are self-explanatory. Some may have required extensive explanations to convey their importance in the model. For many of us, the new model will take some time to get used to. For more than half a century, we've lived and breathed within the cozy confines of the Kirkpatrick-Katzell four-level model. The

Kirkpatrick-Katzell model's great advantage—aside from its seminal notion that learning must produce results—is its simplicity. We can make sense of the four levels in a minute. The new model is more sophisticated, in keeping with the new professionalism needed in the rapidly advancing learning field. The new model, because of its in-depth sophistication and its power to convey many more important messages, will require more from us. It will require us to study it until we comprehend it fully. It will also, when we use it, push us to create much better learning interventions.

Messages Inherent in the New Model

The new model is better than the old model because it sends more sophisticated messages about learning effectiveness. Here are the messages intended to be inherent in the *Learning-Transfer Evaluation Model*.

1. Measuring attendance is an inadequate way of evaluating learning.
2. Measuring course completion is an inadequate way of evaluating learning.
3. Measuring attention is an inadequate way of evaluating learning.
4. Measuring interest is an inadequate way of evaluating learning.
5. Measuring active participation is an inadequate way of evaluating learning.
6. Surveying learners on factors not related to learning effectiveness is an inadequate way of evaluating learning, but it's slightly better than measuring activity or attendance.
7. Surveying learners on factors related to learning effectiveness is better than surveying them on factors not related to learning effectiveness.
8. Surveying learners on factors related to learning effectiveness can provide hints about learning effectiveness, but such surveys should be augmented with more direct measures of learning outcomes.
9. Measuring knowledge recitation is usually inadequate because knowing facts and terminology typically does not fully enable performance.
10. Measuring knowledge retention is better than measuring knowledge recitation, because remembering is a stronger outcome than recitation.
11. Measuring knowledge retention is usually inadequate, because remembering facts and terminology does not fully enable performance.
12. Measuring decision making competence is better than measuring knowledge recitation or retention.

13. Measuring remembered decision making competence (as assessed several days or more after learning ends) is better than measuring decision making competence as assessed during or immediately after learning.

14. Demonstrating remembered decision making competence is a significant milestone in supporting learning transfer.

15. Task competence embodies both decision making and action taking and is, therefore, a fuller measure of preparation for transfer than decision making alone.

16. Measuring remembered task competence (as assessed several days or more after learning ends) is better than measuring task competence during or immediately after learning.

17. Demonstrating remembered task competence is a significant milestone in supporting learning transfer.

18. Demonstrating transfer, whether it is assisted or not, is better than measuring any of the lower levels—because these levels, at best, signify only that a demonstration of competence has been made, not that actual transfer has been achieved.

19. Demonstrating full transfer may be a more difficult accomplishment than demonstrating assisted transfer, because it shows that a person is fully prepared to engage in transfer without significant support, assistance, or prodding.

20. Learning is undertaken to achieve both learning transfer and to obtain beneficial effects of transfer.

21. Learning transfer can affect many stakeholders and outcomes, including (a) learners, (b) coworkers/family/friends, (c) organization, (d) community, (e) society, and (f) the environs.

22. We ought to examine both the positive and negative effects of transfer.

23. To certify transfer effects, we need to use rigorous methods that can assess transfer's causal impact.

I'm sure I've missed a few messages and have inadequately conveyed some of those that are included. The big picture is that *The Learning-Transfer Evaluation Model* sends many messages that are critically important in our work as learning professionals. These messages enable us to get better feedback about our learning evaluations and our learning designs. By using LTEM, you will gain greater clarity in thinking about learning and learning evaluation, your organization and learners will achieve greater results, and the learning industry will have a better way to benchmark learning evaluation strategies and learning designs.

How to Use LTEM

The Learning-Transfer Evaluation Model is designed to be practical. It does this first by sending clear messages—messages that should prompt you and your organization to action. So, for example, if your only evaluation metric is learner satisfaction, you can see in the model that what you're doing is inadequate—so you should aim to do more. Simple!

Note that LTEM offers four levels of evaluation in which some level of validation is warranted (Levels 5, 6, 7, and 8).

- ***Level 5: Decision Making Competence***
  Asks the question, "Do learners know what to do?"
- ***Level 6: Task Competence***
  Asks the question, "Can learners actually do what they learned how to do?"
- ***Level 7: Transfer***
  Asks the question, "In targeted work situations, are learners being successful in using what they learned?"
- ***Level 8: Transfer Effects***
  Asks the question, "If learners have been successful in transfer, what effects has that transfer had on targeted and untargeted outcomes and stakeholders?"

Note that LTEM has two levels that are deemed completely inadequate in validating our learning results—Level 1 (Attendance) and Level 2 (Activity). Let's be careful here. This doesn't mean we should avoid taking attendance or measuring learner activities! There are other reasons to pay attention to these metrics. For example, taking attendance can encourage learner engagement and also give us an indirect indication of learner interest, motivation, and engagement—attributes we can measure to gain insights about our learning designs. Measuring learner attention, interest, and participation can also be useful in helping us diagnose strengths and weaknesses in our learning programs. So, while measuring at Levels 1 and 2 may not enable us to validate the success of our learning, such measurement can give us useful insights for our learning designs.

Note that Level 3 (Learner Perceptions) is divided into two approaches, one of which is deemed completely inadequate as a validation metric and the other of which is deemed largely inadequate. Querying learners about constructs that are unrelated to learning effectiveness is clearly inadequate—because learning effectiveness is our ultimate goal. On the other hand, querying learners about constructs related to learning effectiveness can give us some indication of whether the learning is likely to be successful, but learner feedback on its own is not sufficient to validate learning success.

LTEM is a roadmap. You don't have to take the whole journey in a day! Use LTEM to see where you are in terms of learning evaluation and see what you're missing. Make a plan to do better. We all work in situations that have constraints and goals that may limit our options. Your circumstances may not allow you to measure everything. Indeed, most of us

(most of the time) will not be evaluating at all levels in the model. We won't even be evaluating all the green-is-good validation-enabled levels in the model. That's okay! Use LTEM to improve the feedback you get—to enable you and your organization to improve the learning you're creating.

If you're responsible for a large number of learning interventions, you don't need to evaluate each one fully every time you roll it out. This is too costly and, unless the learning designs are completely unique for each learning intervention, the insights you gather will be redundant.

When should you invest in deeper, more complete evaluations—that is, evaluations utilizing approaches from Levels 5 through 8? You should utilize deeper evaluation methods with strategically important learning interventions, costly learning interventions, learning interventions that will be deployed through many deployments or many years. You should use deeper learning evaluations when you are testing new learning designs—and periodically for all your important learning interventions.

You do not need to engage in deeper learning evaluations when you just recently evaluated a learning intervention using deep evaluation methods. If you basically use the same learning design in all your learning programs, you don't need to evaluate each one deeply. You might instead set up a rotating schedule of deep evaluations.

What if you're a vendor—for example, you create classroom or elearning programs for other organizations or individuals? The best vendors build evaluations into their learning solutions. They don't ask; they just incorporate deep evaluations into their projects. They do this to ensure they are providing maximum value to their customers. They create cycles of continuous improvement not only to benefit their customers but also to benefit their own organization. By enabling valid feedback, they enable themselves to create demonstrably superior learning programs—ones that gain them more and better customers.

Deeper evaluation can also be used to benefit your colleagues. When we work in a situation where feedback is missing (or worse, where it is invalid) we don't learn the right things. For example—and I've seen this firsthand—if we work for an organization that only looks at learner satisfaction data, we get faulty information and we learn to design and deliver learning that gets good satisfaction ratings, but that almost certainly fails to be effective in supporting comprehension, remembering, and on-the-job application. By using deeper evaluation methods, we get more accurate information about the success of our learning designs. We begin to learn the right things—what really works and what doesn't. LTEM, then, can be used to educate you and your team on a continuous basis. When your team is smarter, they'll not only create better learning; they'll like their work better. When people like their work, they stay with their organization.

One of the dirty little secrets of the workplace learning field is that there is a wide disparity in the quality of those of us who are in it. Some are brilliantly effective because they use

research-based wisdom and years of experience in looking at valid evaluation results. Some are dangerously ineffective because they follow recipes to guide their design work, tend to follow fads instead of evidence-based recommendations, and learn from faulty feedback derived from poor evaluation methods. If your organization is not using deeper evaluation techniques to gain validated feedback, you won't attract the best people, your team's conversations and interactions will focus on the wrong things, and you'll create a vicious cycle that reinforces poor learning design thinking.

Using LTEM for Learning Design and Development (Not Evaluation per se)

I'm a big believer in working backwards from ultimate goals to learning design decisions. Many others before me have advocated for this idea as well—most famously in recent times was Stephen R. Covey, who suggested that highly successful people "*Begin with the End in Mind.*" Yet, this wisdom may go back to time immemorial, as evidenced in the biblical quotation (Ecclesiasticus 7:36), "*Whatsoever thou takest in hand, remember the end, and thou shalt never do amiss.*" In learning evaluation, Donald Kirkpatrick said, "*Trainers must begin with desired results and then determine what behavior is needed to accomplish them. Then trainers must determine the attitudes, knowledge, and skills that are necessary to bring about the desired behavior.*"[19]

LTEM can be used as a guide to work backwards from the ultimate goals of learning. When it is used in this way, the process is not an evaluation process, but a project management process used specifically to develop learning programs. In using it in this way, LTEM makes clear the ultimate goals toward which we are working.

To use LTEM as a learning design and development tool, here is a process I recommend:

1. Start by looking at Level 8. What ultimate results are you hoping to achieve? Consider all the possible results, not just "business results." Look at all the impacts you might have on a wide range of stakeholders. Look for both positive and negative effects. Look also for effects on the environs—on the situations, objects, and entities that might also be impacted by your learning programs.

2. Afterward, look at Level 7. How will transfer be manifested? What transfer results are you hoping to achieve? Consider whether you need full or assisted transfer.

3. After you have a clear sense of your Level 7 and Level 8 goals, create a list of Evaluation Objectives—a clear statement of the various measures you will use to evaluate your program's success. For example, if you are training sales managers on how to coach their direct reports, you might specify a 5 percent improvement in the sales managers' direct reports' sales numbers as one evaluation objective (a Level 8

[19] Kirkpatrick, D. L. (1994). Evaluating training programs: The four levels. San Francisco: Berrett-Koehler. Page 26.

transfer objective). You might also specify an acceptable rating of the sales manager's coaching performance (from a survey of those they were assigned to coach) as another evaluation objective (a Level 7 transfer-effect objective).

In addition to these transfer evaluation targets, develop targeted measures for the other levels in LTEM as well. Specify how you will evaluate Level 6 (Task Competence), Level 5 (Decision Making Competence), Level 4 (Knowledge, where meaningful), and Level 3 (Learner Perceptions, being sure to focus on learning effectiveness, at least). Where Level 1 (Attendance) and Level 2 (Activity) metrics are useful to get feedback that can lead to learning-design improvements, you should specify those in advance as well.

4. After negotiating the acceptability of the aforementioned evaluation objectives with key stakeholders, begin the instructional-design process by determining the targeted performance situations and the targeted performance objectives within those situations. In other words, specify what the learners should be able to do and in what situations they should be able to do those things.

5. Once the performance situations and performance objectives are okayed by key stakeholders, you can begin designing and developing the learning program.

6. Preferably, you should use an iterative rapid-prototyping approach—quickly building and testing parts of your learning program AND utilizing Level 1 to Level 6 evaluation methods to formatively evaluate and improve your learning designs.

7. After sufficient iterations of development and testing, you can deploy your learning program and then evaluate its effectiveness using at least Level 3 learner-feedback queries focused on effectiveness, Level 4 (Knowledge, where meaningful), Level 5 (Decision Making Competence), Level 6 (Task Competence), Level 7 (Transfer), and Level 8 (Transfer Effects).

8. Gather and analyze the data; share it with key stakeholders—especially the relevant instructional designers, managers, and trainers; determine plans for improvement; and then push forward with those plans while reinforcing lessons learned.

9. For strategically important learning programs, evaluate routinely. For other programs, evaluate periodically to ensure continuous improvement.

For brevity, the process outlined above is a foreshortened version of a fully specified project management process. Still, it incorporates several important principles. Start by focusing on the transfer levels, Levels 7 and 8. From there, build a separate and distinct set of Evaluation Objectives—the metrics you will use to hold yourselves accountable. Consider all levels of the LTEM model as you brainstorm evaluation options. Involve your stakeholders to negotiate agreements on your key objectives. Don't start instructional design work until

you've clearly outlined your evaluation objectives and your targeted performance situations and objectives. Iterate and test, involving formative evaluation where appropriate. Evaluate continuously (or at least periodically) after your learning program is up and running.

Afterword

I am grateful for the opportunity to engage in working on learning research and learning measurement issues. Without the help of my family, my clients, and those who have volunteered to collaborate on this work, it would be impossible. I'd also like to acknowledge all those who have laid the groundwork over many years for the work we do now.

Will Thalheimer
February 2018

Bibliography

Alliger, G. M., Tannenbaum, S. I., Bennett, W. Jr., Traver, H., & Shotland, A. (1997). A meta-analysis of the relations among training criteria. *Personnel Psychology, 50*, 341-358.

Anderson, L. W., Krathwohl, D. R. (Eds.). (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives.* New York: Longman.

Bloom, B. S. (1956). *Taxonomy of educational objectives: The classification of educational goals*, 1st ed. Harlow, Essex, England: Longman Group.

Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H., Krathwohl, D. R. (1956). *Taxonomy of educational objectives: The classification of educational goals. Handbook I: Cognitive domain.* New York: David McKay Company.

Chargaff, E. (1986). Quoted in J. J. Zuckerman, 'The Coming Renaissance of Descriptive Chemistry,' *Journal of Chemical Education*, *63*, 830.

Jennings, B. K. (2007). On the Nature of Science, *Physics in Canada* 63, 7.

Kirkpatrick, D. L. (1959a). Techniques for evaluating training programs. *Journal of the American Society of Training Directors, 13*(11), 3-9.

Kirkpatrick, D. L. (1959b). Techniques for evaluating training programs: Part 2—Learning. *Journal of the American Society of Training Directors, 13*(12), 21-26.

Kirkpatrick, D. L. (1960a). Techniques for evaluating training programs: Part 3—Behavior. *Journal of the American Society of Training Directors, 14*(1), 13-18.

Kirkpatrick, D. L. (1960b). Techniques for evaluating training programs: Part 4—Results. *Journal of the American Society of Training Directors, 14(*2), 28-32.

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

Marzano, R. J. (2001). *Designing a new taxonomy of educational objectives.* Thousand Oaks, CA: Corwin Press.

Marzano, R. J., & Kendall, J. S. (2007). *The new taxonomy of educational objectives, 2nd ed*. Thousand Oaks, CA, US: Corwin Press.

Pollock, R. (2017). Personal Communication, November 2017.

Sitzmann, T., Brown, K. G., Casper, W. J., Ely, K., & Zimmerman, R. D. (2008). A review and meta-analysis of the nomological network of trainee reactions. *Journal of Applied Psychology, 93*(2), 280-295.

Thalheimer, W. (2016). *Performance-focused smile sheets: A radical rethinking of a dangerous art form.* Somerville, MA: Work-Learning Press.

Uttl, B., White, C. A., Gonzalez, D. W. (2017). Meta-analysis of faculty's teaching effectiveness: Student evaluation of teaching ratings and student learning are not related. *Studies in Educational Evaluation, 54*, 22-24.

The Learning-Transfer Evaluation Model

Abbreviated as LTEM (Pronounced "L-tem")

1	**Attendance**	**Learner signs up, starts, attends, or completes a learning experience.** *A metric inadequate to validate learning success—because learners may attend but not learn.*
2	**Activity**	**Learner engages in activities related to learning.** • Measures of Attention *A metric inadequate to validate learning success—because learners may pay attention but not learn.* • Measures of Interest *A metric inadequate to validate learning success—because learners may show interest but not learn.* • Measures of Participation *A metric inadequate to validate learning success—because learners may participate but not learn.*
3	**Learner Perceptions**	**Learner is queried in a way that does NOT reveal insights on learning effectiveness.** • Examples: Measures that target Learner Satisfaction, Course Reputation, etc. *A metric inadequate to validate learning success—because such perceptions are not always related to learning results.*
		Learner is queried in a way that reveals insights related to learning effectiveness. • Examples: Measures that target Learner Comprehension, Realistic Practice, Learner Motivation to Apply, After-Learning Support, etc. *Such measures can hint at outcomes but should be augmented with objective outcome measures.*
4	**Knowledge**	**Learner answers questions about facts/terminology.** • Knowledge Recitation—during or right after learning event. *Usually inadequate because knowing terminology does not fully enable performance.* • Knowledge Retention—after several days or more. *Usually inadequate because remembering terminology does not fully enable performance.*
5	**Decision Making Competence**	**Learner makes decisions given relevant realistic scenarios**. • Decision Making Competence—during or right after learning event. *Not a fully adequate metric because learners may forget decision making competencies.* • Remembered Decision Making Competence—after several days or more. *ADEQUATE TO CERTIFY DECISION MAKING COMPETENCE.*
6	**Task Competence**	**Learner performs relevant realistic actions and decision making**. • Task Competence—during or right after learning event. *Not a fully adequate metric because learners may forget their task competencies.* • Remembered Task Competence—after several days or more. *ADEQUATE TO CERTIFY TASK COMPETENCE.* *NOTE: "Tasks" comprise both decision making and action taking. For example, a person learning to write poetry could decide to use metaphor, could act to use it, or could do both.*
7	**Transfer**	**When learner uses what was learned to perform work tasks successfully—as clearly demonstrated through objective measures.** • Assisted Transfer—when performance is substantially prompted/supported. *ADEQUATE TO CERTIFY ASSISTED TRANSFER.* • Full Transfer—when learner demonstrates full agency in applying the learning. *ADEQUATE TO CERTIFY FULL TRANSFER.*
8	**Effects of Transfer**	**Effects of Transfer: Including outcomes affecting (a) learners, (b) coworkers/ family/friends, (c) organization, (d) community, (e) society, and (f) the environs.** *Certification at this level requires certification of transfer plus a rigorous method of assessing transfer's causal impact—including positive and negative effects.*