Training is a technical term, applicable to interventions that result in a performance outcome, however, the term is often used inappropriately to elevate conglomerations of content to training status even though no performance improvement results, or is ever likely to result. Myths about evaluation, that it is optional, or too expensive, or somehow unfair to learners (all false statements) persist because of a lack of competence in evaluation skills on the part of a majority of practitioners. Actually, evaluation is a critical component in effective and efficient (and therefore cheap in total cost) training, and it is unfair to learners to expect performance improvement on the job unless the training has been proven through rigorous evaluation. The current rates of change in both technical and non-technical arenas, as well as a slowing economy and tightening budgets contribute to a reality where fewer and fewer people can afford to throw away hours of instruction or training on non-functional conglomerations of content. Rigorous, full evaluation is simply not optional for Instructional Technology professionals. (Contains 15 references.) (Author/AEF)
Evaluation of Instruction (Training): It is Not Optional For Professionals (or, Who Screened the Baggage on Your Flight?)

By: JennyLynn Werner
Evaluation of instruction (training): It is NOT optional for professionals (or, who screened the baggage on your flight?)

JennyLynn Werner
SixSigma Performance

Abstract

Training is a technical term, applicable to interventions that result in a performance outcome, however, the term is often used inappropriately to elevate conglomerations of content to training status even though no performance improvement results or is ever likely to result. Myths about evaluation, that it is optional, or too expensive, or somehow unfair to learners (all false statements) persist because of a lack of competence in evaluation skills on the part of a majority of practitioners. Actually evaluation is a critical component in effective and efficient (and therefore cheap in total cost) training, and it is unfair to learners to expect performance improvement on the job unless the training has been proven through rigorous evaluation. The current rates of change in both technical and non-technical arenas, as well as a slowing economy and tightening budgets contribute to a reality where fewer and fewer people can afford to throw away hours of instruction or training on non-functional conglomerations of content. Rigorous, full evaluation is simply not optional for Instructional Technology professionals.

Introduction

Although calls for data in the field of Instructional / Performance technology continue from every direction (e.g. Anglin, et al, 2000; IBSTPI, 1998; Gery, 1999; Merrill, et. al, 1996; Shrock, 1999), the number of publications which include meaningful evaluation is not increasing, and may be decreasing (Werner & Klein, 2000). Evaluation is often considered to be optional or too expensive, or is simply not a part of the project from the beginning. Some customers and clients have been misinformed in the past that evaluation is expensive, or unnecessary, or somehow unfair to the learners (all these assumptions are false). In fact, evaluation is not only not expensive, it is the cheapest step in the process and the best practice for all involved. It's also not at all time-consuming if the intervention involves observable, measurable performance outcomes and the practice is sufficient and aligned to those outcomes (inherent in the definition of training, so an intervention without observable, measurable outcomes is simply not training). In fact, the most expensive choice anyone could make is to forgo evaluation and release dysfunctional or non-functional interventions that then consume vast amounts of resources in terms of organizational resources, material costs, and learner time and effort. In addition, most unmeasured “training” sets learners up to fail - thus impacting their subsequent learning efforts far into the future.

A case study, utilizing a technology-based intervention to address a need for skill development by new hires in a semiconductor fabrication facility (Fab), is provided to illustrate the evaluation process, including summative evaluation, and demonstrates the necessity for evaluation as well as the total value (costs and benefits) of routine and systematic evaluation of interventions. The application of Level 1 (trainee response) data, Level 2 (posttest performance) data, and Level 3 (on-the-job performance) data to revise the intervention until it functions at the required level to meet the business need are described.

Foundations

Instructional technology is the application of expertise (in the areas of human information processing, learning, performance, cognition, etc.) to solve real-world gaps in skills and knowledge through specially designed interventions. It is this expertise, along with effectiveness and efficiency data from previous interventions, that enables us to develop instruction that works for the target audience. Although it seems to be lost on a large number of practitioners, the profession of instructional technology is rooted in observable, measurable performance outcomes, under conditions where the consequences of failure are serious; after all, the application of learning theory to solve real-world problems effectively and efficiently really took off as experts worked with the military to prepare soldiers to fight and survive in many types of warfare (see Gagne, 1989; Dick & Carey, 1985; Anderson, 1995; National Research Council, 1991).

As we work to improve the performance of others, we must role model performance to business indicators ourselves, and demonstrate that discipline around assessment and accountability are fundamental to continuous
improvement. Using our own tools while demonstrating efficiency and effectiveness and performing at least to the
levels we demand of others are both critical components of our credibility. Our customers are nearly always held to
some level of minimum performance; can you imagine a factory manager who is only measured according to the
amount of raw materials going into the factory? For this scenario, actual production of working units is not
measured, ever. Training which is not measured according to actual performance outcomes of trainees after training
(i.e., training tracked by a body count or attendance, and with opinion surveys) is very analogous to the Factory
Manager who is measured only by how much raw material goes into the factory. Our customers and clients deserve
better; the standards of our profession demand better (see standards at AECT site, http://www.aect.org and at
IBSTPI at http://www.ibstpi.org; consider reviewing the new ISO requirements for training also). It’s impossible
for measurement to be superfluous, to be expensive in terms of total cost, or be time consuming in terms of total
resources. One basic premise of both Total Quality and Continuous Improvement is that, if you can’t measure what
you’re doing then you don’t know what you’re doing (measurement is critical to improving both process and
product; remember, it’s possible to change for the worse). This is so very critical because an intervention isn’t
training because the word training appears in the title. Training is only training if it functions to enable a permanent
(or near permanent) performance change that is observable and measurable.

The Premise

There are clearly issues with the present status of the profession of performance or instructional
technology. Training is treated as a joke, as a necessary evil, as too expensive or too difficult to do well, and the low
expectations of end users (trainees) and performance owners (managers) enable the practitioners to perpetuate
content masquerading as training. The same problems have been discussed over and over in the past twenty years of
Training & Development (an American Society of Training & Development publication), as well as in articles
published inside and outside the field. The complaints are consistent across the years, and are detailed in articles
like: Furnham’s Fire the training department, (1997), Kruger, & Dunning’s Unskilled and unaware of it: How
difficulties in recognizing one’s own incompetence lead to inflated self-assessments (1999), and Armour’s Big
lesson: Billions wasted on job-skills training (1998), just to name a few. For example, in a USA Today article,
Armour (1998) explains provides some insights, “Faced with crippling skill shortages, employers are spending
skyrocketing amounts of money training workers. The problem? Many programs just don’t work. Billions of
dollars are spent on wasteful training courses, experts say . . . $5.6 billion to $16.8 billion is wasted annually on
ineffective training programs,” and even provides a solution, “. . . experts say businesses should check for results to
separate effective programs from costly gimmicks. ‘American industry is spending billions and billions on training
programs and doing no evaluation of their effectiveness,’ says Cary Cherniss, a professor at Rutgers. ‘You have to
measure it.’ ”

Case Study

Semiconductor Equipment Operation Training

The intervention was designed to meet a business need; new hires needed skills training to work in a
semiconductor wafer Fab. SortSoft was a new station controller software interface for the Schlumberger 9000
(S9K) tester. The SortSoft Computer Based Training (CBT) package is based on a full, real-time simulator of the
SortSoft interface, and therefore was expected to reduce certification time for level 1 Technicians in Sort areas
across the company, as well as reducing the trainer time (for technician trainers or engineers/equipment owners) and
S9K tester (production) time required for training functions.

The project was unique because it was the company’s first full, real-time simulator built to run on a PC and
supported by a fully-aligned CBT. In addition, concurrent development was required as the training was to be
available to support the implementation of the product at each Fab site. The objectives of the training package
accurately reflect the certification checklist from the factories. In fact, the team nearly gained approval for the
course completion to stand for certification. The equipment owner engineers (those responsible for the operation,
up-time, maintenance, training and certification for the equipment they “own”) acknowledged the alignment of the
training and felt that the posttest did reflect the actual operations skills required. Although several additional alarm
sequences were added to the wafer sorting segment of the posttest, the engineers did not accept it as full certification
because they wanted to check for themselves that each operator could demonstrate all the skills on an actual tester.

At the time the team began initial design, a checklist was created covering all the skills required for
certification on the SortSoft interface. The checklist was used for updating the skills of technicians certified to
operate the existing interface at the development facility. That checklist was also used as a basis for another recording instrument; equipment owner engineers were asked to carefully track the time it took to re-train the technicians on the new interface and to train new technicians as well. Seven new technicians were trained at the development Fab, and the recorded times for training new technicians were used as a baseline for comparison in measuring the efficiency of the CBT/Simulator training intervention. It is interesting to note that the equipment owner engineers had previously asked to provide training time requirements by the programmer initially assigned to the project and their responses bore no relationship to the actual times tracked during technician training in the development Fab. One engineer reported training times as much as 50% in excess of the actual time required, all other engineers reported training times more than 50% below the actual time required. The instructional designer, aware of the reliability and value of self-reported data, created the time tracking instrument and carefully communicated the purpose and importance of tracking time spent on training. Although the engineers did not match the specificity of time tracking achieved with the CBT/Simulator intervention, the times reported were more accurate and therefore a meaningful basis for comparison.

Usability / Functionality Testing

Basic usability testing, even for paper-based manuals, precedes any performance testing; functionality testing can easily run concurrently with usability testing. Usability means that people can follow the directions, use the documentation or intervention successfully, not get lost or frustrated while trying to use the materials, and not have trouble figuring out what to do next. Functionality testing means the materials work consistently like they are supposed to: the exercises or practice items are aligned to what the trainees are learning to do and anyone who has completed the instruction should be able to complete most or all of them correctly; the answers as well as steps for successful completion are provided as appropriate feedback when required; scenarios or simulations are provided where appropriate to learning outcomes; and the thing works (application runs, links link, manual is accurate and has complete table of contents, index, and sufficient and accurate instructions/directions to users). Once people can get through it successfully, then a tryout or pilot test will provide critical feedback on effectiveness and efficiency of the intervention.

In the SortSoft case, content experts and the test engineers, using an explanatory guideline and checklist created for that purpose, also reviewed the intervention. In some cases, the equipment owner engineers requested that the content be restructured or reorganized in ways that made it more useful for experts (they themselves) but NOT for the novices who made up the target audience, and those changes were refused (with extensive communication about why).

During usability and functionality testing, we identified necessary revisions. We installed a dialog box to pop-up when the trainee selected Exit from the application menu, so a trainee could choose to return to the CBT if they did not actually want to exit the CBT. Also, information and content were presented in a different colored text than actual instructions for using the simulator and directions for completing practice items. Some small instructional changes were made to revise confusing practice items or add detail to feedback at the end of several modules. At the request of the equipment owner engineers, two different alarms which routinely occur during typical wafer processing were added in two the final practice scenario and also into the posttest simulation of three wafer lots.

Most surprising of all was the behavior of all the testers in regard to the interactivity level of the training. Trainees were required to complete each module correctly before continuing on to the subsequent module. Directions at the beginning of the CBT and the beginning of each module explained that clearly. However, every single trainee ignored the requirements for interaction and practice, and simply clicked the Next button until they reached the end of the first module; they were then dismayed to see instructions directing them back to the beginning of the module so they could complete the required interaction with the simulator that provided the practice. To address this astonishing outcome, a short module was designed to introduce these trainees to performance-based training. By requiring them to actually complete a brief simulation of the Log In process (no different than logging into their network) prior to beginning the first module, we began to change their expectations about interactive, performance-based training; they all acquired bad habits from completing electronic page-turners (simple presentations) mis-labeled as training.

The CBT development schedule followed the SortSoft development schedule closely, so after additional tryouts at several sites with members of the target audience, and revisions required by software changes or lessons learned in tryouts, the pilot was scheduled at the newest Fab. The entire population of new hires training for Sort Operations (all shifts) reported to a small computer lab for a dedicated shift, and each completed the CBT/Simulator intervention. Trainees worked at their own pace. Trainees could log out to take breaks as each chose; the total time
did not include breaks. Time stamps throughout the program tracked the completion time for each module and each scenario, and all the performance data (answers to questions, selections made when operating the simulator, etc) was also tracked electronically. The completion time varied from 4 hours 10 minutes to 7 hours 50 minutes, with 92% of the trainees finishing in less than 6 hours 30 minutes.

Quantitative (performance) Data

Most quantitative assessments in training involve both skill and knowledge components; these trainees had to know some rules regarding safe and effective operation of the tester itself in addition to knowledge about the operations of the software. The skills section of the posttest utilized the SortSoft simulator, and when the trainee could demonstrate the capability to process three lots of wafers flawlessly while correctly responding to all alarms, the equipment owner engineer conducted the Skills Certification test. The Cert test simply required that the trainee flawlessly sort three lots of wafers while correctly responding to all alarms, although the checklist included the components of that goal in more detail, which enabled engineers to determine areas of failure for those who did not complete the task flawlessly; those components are also part of the SortSoft posttest and performance test.

The standards for the CBT/Simulator intervention were set so as to align to the actual performance goal in the factory. Each equipment owner reported the same results: for the first time in his/her experience, all of the trainees achieved certification on the first attempt, which they described as a huge saving in their time during the most critical time period when a new wafer Fab is ramping up to begin production.

The Level 2 included both the Posttest and subsequent Fab Certification for each trainee:

Posttest (two sections):

Knowledge – trainee can:
- Select SortSoft's purpose / objective, environment and benefits from a list
- Select the correct description and procedures for SortSoft operations
- Label SortSoft’s Main Screen and the information displays
- Access / utilize commands in SortSoft’s Menus necessary to perform basic SortSoft tasks
- Recognize SortSoft’s alarms and state appropriate responses to each

Skills – trainee can:
- Log in to the tester / SortSoft
- Run a correlation wafer and determine next steps
- Set up the tester for a specified wafer lot and / or probe card
- Introduce a wafer lot
- Modify a wafer lot
- Begin, pause and end testing (with and without saving the data)
- Respond to / resolve typical operations alarms in order to continue or restart testing

As noted, all trainees passed the posttest, and all passed the Certification Test on the first attempt. Initial competency levels, performance consistency, and response to alarms were among the factors listed by our clients when we requested that the equipment owner engineers assess the costs / benefits of the intervention. Two to three shifts (12 hour shifts) were dedicated per trainee for instructor-led training (conducted one-on-one by Technician Trainers or equipment owners), the CBT/Simulator freed up all that time on the part of the Trainer, and all but five or six hours on the part of the trainee, fewer Cert Tests also freed up time for everyone. More importantly, that time was freed up during the most intense times in semiconductor Fab work, the ramping up stage. Equipment owners noted benefits of hands-on training without risking the multi-million dollar S9K tester or the $25,000 probe card used to conduct the sorting tests. Down time, even for training, is an expensive proposition because the Fab is so expensive to operate even aside from the capital costs of the testers. Best of all, according to the engineers, the hands-on training did not subject wafers (nearly complete in fabrication when sorted, and therefore very expensive in terms of processing done) to damage and destruction in the hands of novice trainees.

Unfortunately, however, the engineers had developed such a profound distrust of content and applications mis-labeled as training that we were never able to persuade them to allow the simulator-based posttest to qualify as certification for technicians who passed flawlessly. Even the successful re-certifications of 100% of those trained using CBT/Simulator over the next year was not sufficient to overcome the additive experience of so many bad training experiences. Their responsibility as equipment owners includes expensive equipment and supplies as well as expensive product, and although they enthusiastically acknowledged the success of the training and posttest, they wanted to see trainees perform at required levels before they certified technicians to run the equipment without supervision.
Qualitative (opinion) Data

The CBT/Simulator intervention was very different from the typical electronic application provided to the trainees during New Hiring training (page-turners organized by content experts), and the novelty effect likely contributed to the 100% Level 1 response rate. Also, the Level 1 instrument was very specific to the content of the CBT and to the skills required for the posttest and certification, so there was a clear link between the responses trainees gave and the training they had just completed. Several trainees commented that they appreciated the difference between the generic surveys used for other training and one designed to elicit meaningful inputs for the specific training intervention.

SortSoft CBT Level 1 Description

The Level 1 instrument was specifically designed to measure perceptions of motivation, continuing motivation, satisfaction, and confidence. The survey items were specific enough to both content and skill outcomes to enable revisions where necessary. Not all comments resulted in revisions; for example, trainees consistently complained of too many practice items, that they were required to solve more problems than they needed to in order to acquire the skill. However, during tryouts trainees had consistently failed the posttest performance test on the simulator until we increased the number or complexity of the practice problems. Level 1 data has very little value until it is correlated with Level 2 (performance) data; the opinion of the trainee is irrelevant if there is no measurable, observable improvement in job skill – after all, this is on-the-job training in usually competitive businesses, not entertainment. All of the trainees agreed or strongly agreed to the statements on confidence about their performance in the Fab, and that the intervention provided enough information, instruction and practice. All but one trainee would recommend the training to their colleagues and all but one would choose a CBT/Simulator to learn to operate other Fab equipment. The trainee who expressed a preference for instructor-led training was the trainee who took the longest to complete the intervention; language skills (reading comprehension, specifically) were likely a factor in both the completion time and the preference for instructor-led training.

Trainees responded to a list of items using a 4-point Likert-type scale attached to each: Strongly Agree, Agree, Disagree, and Strongly Disagree; after this much time expended on training so critical to the trainee keeping a job, the trainee can reasonably be expected to have an opinion about each statement (an online, 4-point Likert-type scale was implemented with course release).

The Level 1 data was reported out as counts (the number of trainees who agreed with a statement, the number who disagreed strongly with the statement, etc; all four categories for each statement). I specifically coach my clients to pay attention to how practitioners roll up qualitative data, as descriptive statistics cannot be meaningful for this data. The numbers assigned to the categories (typically 1-4 or 1-5) are arbitrary, or artificially determined, so the categories remain Strongly Agree, etc. You cannot take an average of one Agree and one Disagree, and the distance between Agree and Disagree is not numerically equivalent to the distance between Disagree and Strongly Disagree, nor equal to twice the distance between Agree and Strongly Disagree. Reporting out a mean score for opinion data is just meaningless, and it destroys a practitioner's credibility with anyone who has a basic understanding of measurement.

Level 3 (performance on the job) data was captured subsequently when each S9K operator was successfully re-certified to current specifications at appropriate intervals across the next several years. A summative evaluation conducted after three years yielded a request for one module to be deleted due to a change in procedure for processing wafers. Unfortunately, the only other request made at that time was a request to disable the data tracking and writing functionality as no one was using the data even though these trainees all require training in numerous other areas and the data is valuable for learning about that critical target audience who routinely handles and processes wafers so late in the expensive process.

Conclusion

Without using data from the Level 2 and Level 3 evaluations of the SortSoft CBT/Simulator, we could never have achieved the level of performance described above. In the final analysis, we did a better job of training new technicians than the equipment owner engineers; because ISD applied along with expertise in learning and information processing is more powerful than subject matter expertise alone for training. If we had released the intervention in its initial form, the performance levels of new and retrained technicians would have been affected. The company would have paid a huge price when utilizing less efficient training that required far more technician training time over the years, and using less effective training requiring more technician training time adds higher
costs due to more engineer training time and certification time. The costs in damaged equipment and products would be added on also. Overall, it is always cheaper to make the training effective and efficient; instructional technologist time is much cheaper than the long-term costs of bad training for all the trainees who are affected. There are an infinite number of excuses, mostly mythical and all of which result in a lack of accountability for a practitioner. A conglomeration of content, even if carefully formatted, does not constitute training; a batch of information handed out (with or without a PowerPoint presentation) is almost never training in any meaningful sense of the term. In fact, bad training is such a routine occurrence that we had to add a special section teaching trainees to actually complete the interactivity before attempting to continue with the CBT. Good training can even help people become better learners; bad training damages profitability, attitudes, and credibility and causes trainees to develop bad habits.

As professionals, we have a responsibility to deliver what we promise to deliver and what we get paid for. The current rates of change in both technical and non-technical arenas, as well as a slowing economy and tightening budgets contribute to a reality where fewer and fewer people can afford to throw away hours of instruction or training on non-functional conglomorations of content. The consequences of failure are serious; whether the training is for baggage screeners in Boston's Logan airport, or the government employees who help families fill out required forms, or the people who repair combat equipment for Special Forces; the commitment is to ensure learners achieve performance capability. Evaluation is simply not optional for Instructional Technology professionals.

**One Thing To Take Away With You**

The most important thing anyone practicing in the performance profession can learn is the power of data if applied as part of the continuous improvement cycle. The following paragraph is used with permission of the authors (available at [http://www.whidbey.com/frodo/isd.htm](http://www.whidbey.com/frodo/isd.htm) (emphasis added):

Perhaps the **greatest strength** of the ISD process is the evolutionary nature of the prescriptive, research-based model itself. While the practice of ISD still retains the strengths of the empirical evaluation and revision cycles, to the extent research and experience permit, it is prescriptive. That is, rather than depending extensively on the test-revision cycle to generate effective instruction in an iterative manner, every attempt is made to incorporate research findings and past experience into the detailed procedures and supporting ISD documentation to ensure that the instruction developed comes as close to the mark as possible the first time. This improves the **validity of the process while also improving reliability.** This has proven to be a **powerful tool** in large scale ISD. In addition, as the process provides more data from the constant evaluation process, the procedures can be continually improved.

**References**


Anglin, G. J., Dick, W., Morrison, G., & Richey, R. (2000, February). Reflections on Instructional Technology from some who have been around a few (Ha!) years. Symposium held at the International Convention of the Association for Educational Communications and Technology (AECT), Long Beach, CA.


NOTICE

Reproduction Basis

This document is covered by a signed "Reproduction Release (Blanket)" form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.

This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").