A Critique of Kirkpatrick’s 4 Levels of Evaluation

Evaluation is one of the keys to the success of the human resource development field, and an important one at that. Measurement and evaluation (M&E) gives HRD practitioners a tool to be able to see whether or not the performance improvement efforts are working, successful, or simply a waste of time and money. In their article looking at M&E, Wang and Spitzer (2005) state,

“To us, the debate is not whether M&E is important or whether HRD researchers and practitioners should conduct evaluations. We believe the core of the debate should be the validity of some of the M&E methods commonly used by HRD practitioners, whether we should be content with what we have, and what we are doing to advance the field” (p. 5).

In this article, Kirkpatrick’s four levels of evaluation framework, which is one of the most widely used M&E methods, will be assessed for its effectiveness in today’s HRD field.

First published in 1959, Donald Kirkpatrick’s four articles on the four levels of evaluation have made a huge impact on the field of AHRD. Other, similar frameworks have been designed, yet Kirkpatrick’s remains the most popular among training departments and evaluators. Considering the framework is still utilized by many practitioners today, Kirkpatrick’s four levels of evaluation continue to be very significant to the AHRD field.

Summary of Kirkpatrick’s articles
The four levels of Kirkpatrick’s framework include reaction, learning, behavior, and results. He developed these four levels to provide a starting point and guideline in how to evaluate a training program (Kirkpatrick, 1959a). Stated in his very first article published in 1959, Kirkpatrick says, “These articles are designed to stimulate training directors to increase their efforts in evaluating training programs. It is hoped that the specific suggestions will prove helpful in these evaluation attempts” (p. 3). In his articles, Kirkpatrick (1959a; 1959b; 1960a; 1960b) provides not only definitions of the purpose of each of the four levels but guidelines on how to evaluate programs at each of those levels. He also discusses the importance of each of the levels.

The first level, reaction, is used to find out people’s feelings and attitudes towards a program. Kirkpatrick (1959a) states that one of the reasons measuring people’s reactions to a program is important is because; “Decisions by top management are frequently made on the basis of one or two comments they receive from people who have attended” (p. 8). The author also extends that in order for “maximum” (p. 8) learning to take place, participants must be interested and like the program.

Learning is the next level. As Kirkpatrick (1959b) discusses, learning is a bit harder to evaluate compared to evaluation of reaction, especially when evaluating “principles and facts” compared to “techniques” (p. 23). For the purposes of Kirkpatrick’s (1959b) level two of learning, learning is defined as “What principles, facts, and techniques were understood and absorbed by the conferees” (p. 22). It is mentioned that in order for learning to effectively be measured, the measurement tool must go over all of the information covered in the program, otherwise “it will not be a valid measure of the effectiveness of the learning” (Kirkpatrick, 1959b, p. 24).
Level three, behavior, goes a step further beyond reaction and learning. Behavior is even more difficult to measure compared to reaction and learning. For one, it requires more than just the person being evaluated; the employee’s management team and his or her subordinates are just some of the people that are highly suggested to be a part of the evaluative process as well when looking a person’s behavior (Kirkpatrick, 1960a).

Although the process of evaluating behavior may be time consuming, according to Kirkpatrick (1960a), “it is worthwhile and necessary if training programs are going to increase in effectiveness and their benefits made clear to top management” (p. 18).

Behavior is a good evaluative piece to not only decide whether or not to keep a program but to also gauge how much the participants chose to take away from it the training.

Finally, level four of Kirkpatrick’s framework looks at results. Kirkpatrick states (1960b), “From an evaluation standpoint, it would be best to evaluate training programs directly on results desired. There are, however, so many complicating factors that it is extremely difficult if not impossible to evaluate certain kinds of programs in terms of results” (p. 28). Clearly it is shown that out of the four levels, results are the most difficult to measure, but often times the most important one to measure. Results are what management really cares about.

**Criticisms and recommendations to the framework**

One common criticism among many practitioners who study Kirkpatrick’s framework is that the framework is incomplete (Kaufman & Keller, 1994; Holton III 1996; Watkins, Leigh, Foshay, & Kaufman 1998; Bates, 2004). The field itself has come a long way since the early 1960’s and a lot has changed since then. Because of this, it is often questioned whether Kirkpatrick’s framework is still an effective tool to produce an
effective and wholesome evaluation of a program. The only catch here is that it is still being used and is still one of the most popular tools to use for evaluating training. Just because it was designed and developed over five decades ago does not necessarily mean it can no longer be an effective evaluation tool. That being said, because the field has grown and developed so much, perhaps the framework can be adjusted a little bit to incorporate the changes and new material that has been learned about good evaluation and the pieces of it, rather than being completely thrown away.

Holton’s (1996) two main criticisms of Kirkpatrick’s framework are that it is indeed not a model and that it is ultimately a “flawed” (p. 5) evaluation model. He takes the idea of the 4 levels being a “model” too literally. Kirkpatrick’s (1996) response to Holton is, “The funny thing is that I personally have never called my framework a ‘model,’” (p. 23) and that “Personally, I don’t care whether my work is called a model or a taxonomy” (p. 23). Kirkpatrick (1996) asserts, “as long as it helps to clarify the meaning of evaluation in simple terms and offers guidelines and suggestions on how to accomplish an evaluation” (p. 23-24), is all that matters to him. If Holton were to really look at Kirkpatrick’s framework from its actual intent rather than what other practitioners may or may not have turned it into, then he would not have adequate support for his analysis of it not being a model. Even in his very first article, on the very first page, Kirkpatrick (1959a) provides the intent of the four levels to be used as guidelines for evaluation, not a perfect technique, as cited earlier in this article.

Holton goes on in his article to not only point the flaws of Kirkpatrick’s levels but also proposes a new model that goes beyond Kirkpatrick’s. After providing much background research to support his creation, the complete proposed model is presented.
This proposed model, just by looking at it, is very overwhelming. Holton himself even identifies how much time and dedication would be required to carry out the actual model. Though there is support for the model, the flaw Holton makes is that he proposes a model that is more suitable for “research” (p. 17) purposes, rather than a model for an everyday business worker. Holton takes his time to dumb down and attempts to show how inadequate Kirkpatrick’s framework is just to come up with a sophisticated and highly detailed model that ultimately does not seem to have the same purpose as the framework he is criticizing and trying to fix in the first place.

Yes, it is agreed that if Holton’s proposed model was heavily studied, used, critiqued, and eventually narrowed down to be smaller, it could stand to be a good model to use as an everyday evaluation tool. What Holton fails to see, though, again is the ultimate purpose of Kirkpatrick’s framework: to do exactly that, to provide a useful, simple guideline for anyone to be able to use as at least a starting point in an evaluation attempt.

Not everyone needing to evaluate a program, whether it be training or not, is going to be an expert on this particular process for evaluating training. This is one of the strengths of Kirkpatrick’s framework; it provides a starting point, and often more, for those needing assistance on how to go about beginning an evaluation of a program.

Kaufman and Keller (1994) are additional researchers who have taken a look at the Kirkpatrick’s framework. They provide sound suggestions on how to further strengthen the framework itself. One suggestion they make for the framework to be altered so it could be used for all types of “performance improvement interventions” (p. 377), rather than to just be used for training programs. As it stands now, Kirkpatrick’s
framework is most conducive only to training. This recommendation would be ideal because not only would it help training departments evaluate a training program, but it would also give other departments that same useful outlet to evaluate other performance improvement efforts.

Another recommendation of Kaufman and Keller’s (1994) includes adding a fifth level to the existing four levels of evaluation. A level to cover the social aspects, those people, things, places, etc. that are affected outside of an organization, is the fifth level that would extend the current framework to go from a “microlevel” to a “megalevel” evaluation tool (p. 377). Organizations cannot just think of themselves and their employees, but the people who they serve as well. This could be an effective measurement because it would get not only the trainers in the mindset of thinking beyond the internal effects of the company’s actions but the trainees as well. The only downfall to this idea is that the authors do not really address why and how it would affect whether or not a training program is effective or not, which is the whole purpose of the four levels of evaluation.

Level one, reaction, of Kirkpatrick’s framework is one area that is commonly looked at as a flaw of the framework and one that should be revised. Alvarez, Salas, and Garofano (2004) stated that, “new methods for constructing reaction measures are needed so that they provide evaluators with information regarding the appropriateness of training content and design” (p. 407). Rather than asking participants about how they liked or did not like the training, it has been suggested that this level also target more towards the “transferability or utility of the training” (Alliger, Tannenbaum, Bennet, Traver, & Shotland, 1997). Rather than simply target how much a participant enjoyed the training,
Alliger et al. (1997) also suggested that finding out how useful a participant thought a training program will be for them in the future would be a better judge to help figure out how successful the program was.

Holton (1996) also agrees that reaction is not sufficient as it stands in Kirkpatrick’s framework. He provides yet another take on what level one really should be measuring, and that is motivation. Holton (1996) provides adequate support to show that multiple researchers have found that there is no direct correlation between reaction and learning. He then goes on to say show that, “Motivation to learn has a direct relationship with learning,” (Holton, 1996, p. 11) and begins his argument that instead of reaction being measured, motivation is what really should be measured in terms of how much a participant is going to learn from a training program.

As it stands right now, level one poses to be insufficient on its own. The question that researchers seem to face is what direction level one should go. Should reactions to the program continue to be measured and simply add measurements on the usability of the training? Or should asking participants how much they liked a program be scrapped altogether and be replaced instead with how motivated were they to learn and what was their ability levels before entering the training program? Or neither? Just asking how much a participant liked or enjoyed the program is simply not enough. It also should not be a tool to gear whether or not a training program should be scrapped altogether.

Regardless of what direction level one goes, the author of this article agrees that level one is an area that holds for much room for improvement.

Kirkpatrick’s model is clearly one that deems popular for its use among HRD practitioners and training departments, even still today. The field of HRD has heavily
evolved and many practitioners are concerned with the validity of the framework and whether it is still an effective tool to measure evaluation. Since it is still heavily used today, it deems to hold some strength for its effectiveness but that does not mean it is 100% effective and no changes are required. Further investigation and research on whether to “complete” Kirkpatrick’s framework or to start over with a completely new model should be continued.
References


