Four levels, one judgement: Determining the relative importance of Kirkpatrick’s four levels for evaluating training programs

Kirkpatrick’s four-level framework remains one of the most widely used approaches for evaluating training programs. However, to use it rigorously, evaluators need to be able to determine the relative importance of the four levels within their particular program context. How important is participants’ reaction compared to changes in their behaviour? What if the timeframe is too short to observe broader organisational results – what effect would this have on our overall evaluative judgment?

In this article Lirata’s Evaluation Specialist Fran Demetriou provides a guide to importance determination of the four levels, and explores some of the challenging issues raised by this process.

The research that supported this work was undertaken as part of Fran Demetriou’s Masters in Evaluation degree (University of Melbourne), and supported by Mark Planigale (Lirata’s CEO). Fran presented this research and its application within a Lirata evaluation project, at the Australian Evaluation Society 2018 conference. Fran is currently undertaking further research on this area, and welcomes feedback to support the development of this work.

Introduction

Donald Kirkpatrick published his four-level framework for evaluating training programs in his 1994 book Evaluating Training Programs: The Four Levels. His objective was to ‘provide a simple, practical four-level approach for evaluating training programs’. Since then, the framework has been applied extensively in evaluating training and development programs.

Kirkpatrick proposed that the four levels a training program should be evaluated on are:

- **Reaction**: participants’ satisfaction with the program, including engagement, relevance and quality of content.
- **Learning**: changes in participants’ knowledge, skills, attitudes and confidence.
- **Behaviour**: changes in participants’ behaviour as a result of the program (the application of their developed knowledge, skills, attitudes and confidence)
- **Results**: the broader results that have occurred because of participation in the program (such as better consumer outcomes).

Lirata recently used this framework to evaluate a leadership program. While recognising that the four levels do not provide a complete framework in which to evaluate a program (there are other criteria of merit one may wish to evaluate on), and it is incomplete as a theory of change (the earlier levels are not sufficient to cause changes in latter levels), the framework does provide a clear way to organise an evaluation and communicate findings.

Despite much discussion on the benefits, challenges and flaws with the Kirkpatrick framework, it has endured as a widely used framework for evaluating training programs.

Synthesising findings

Lirata’s consulting team have recently been strengthening our methods for bringing together the findings in our evaluation projects so that we
are presenting clear and defensible judgements about the programs we are evaluating.

Being able to simply and explicitly present this helps our stakeholders understand how we have concluded how well their program is performing and what we have considered when providing recommendations.

Not all of our evaluations need to come to an overall judgement of the performance of the program. The need for an overall synthesis depends on the purpose of the evaluation and the client’s requirements.9

Where an overall judgment is required, preparing for a synthesis requires two key steps:9

• Merit determination: Defining what constitutes different levels of performance (e.g. poor, adequate, good, excellent)
• Importance determination: The process of assigning labels to dimensions (in the Kirkpatrick framework, this would be the four levels) to indicate their importance

To be able to determine importance requires a better understanding of the relationship between Kirkpatrick’s four levels, and deeper thinking on contextual factors that would impact their importance determination.

Factors for determining importance

Importance determination is often strongly influenced by the intuitions of stakeholders about what matters most in their context. However, there are also a range of technical factors that can affect this determination, and that vary from evaluation to evaluation. We identified four factors that would have implications on the relative importance of the four levels: timeframes, causal relationships, the program’s sphere of accountability to barriers and enablers, and the implications of differences in Reaction (the first of the four levels).

1) Timeframe expectations

As we progress through the levels, the time lapse for effects to occur increases. Participants react to a program immediately and they learn immediately or shortly after having time to digest the content. The time it takes for participants to apply their learning, and then for organisational results to occur because of changed behaviour is often considerably longer. The time lapses between outcomes in Learning, Behaviour and Results are often difficult to predict.

This presents a challenge when considering whether missing data on effects at latter levels should be considering detrimental to the program’s success, or simply an issue of evaluation design and inaccurate assumptions about when effects will occur.

However, clients or stakeholders may have expectations on when they want these outcomes to happen by, and these timeframe expectations may be critical or non-critical to the program’s success.

The questions we considered when weighing up how important each level was on this factor were:

• What is the expected timeframe for effects at this level?
• How critical is this timeframe for program success?
• How sufficiently does the evaluation design factor in expected and unknown timeframes?

2) The causal relationship between the levels

The framework is not a comprehensive causal model (the prior levels don’t necessarily lead to the latter levels, and other context-specific factors will contribute to effects in these), but certain levels are at least necessary for latter levels to be achieved as a result of the program. For instance, research has contested the relationship between Reaction and Learning;2, 4, 7, 8 while Learning does not cause Behaviour
change (as other factors are also needed), in the context of training programs it is necessary for Behaviour change to occur – as is the case for Behaviour and Results.

The question we considered when weighing up how important each level was on this factor was:

- How essential is this level for effects on latter levels?

3) Sphere of accountability to external barriers and enablers

The first two levels occur within the program, but as a participant progressing to applying their learnings, and then achieving results for the organisation, external barriers and enablers come into play.

For example, in a program aiming to build financial management capability, a participant may learn how to budget effectively, but their workplace may not provide them with an opportunity to apply this knowledge. The inability of the participant to apply this learning should not result in a poor verdict of the program overall, if the program was not responsible for ensuring application occurred beyond the program.

The questions we considered when weighing up how important each level was on this factor were:

- How susceptible is this level to external barriers and enablers?
- What is the program’s intended level of control/influence over these barriers and enablers?

4) Other implications of Reaction

While the Reaction level might often be considered less important than the Learning or Behaviour levels, there can be significant implications of different levels of performance at Reaction level, including reputation and subsequent enrolment into the program. This is influenced by the context the program is delivered in, such as the level of competition, whether the training is mandatory or optional, and whether the program is a one-off or intends to enrol further cohorts.

Consider the difference in importance of participants’ reactions in a mandatory internally delivered workplace health and safety training course versus a locally delivered public speaking course with several competitors. The latter has more riding on participants’ reactions than the former, in terms of program viability.

Consider the nature of the content of the program – how might negative reactions impact participants’ wellbeing, and therefore how critical is a minimum standard of performance for this level? In training programs with a strong personal element, reaction may have a more significant impact on wellbeing than in programs with a more technical focus.

The potential for reputational consequences, impacts to subsequent enrolment and program viability and participant wellbeing are some of the implications we considered when weighing up the importance of Reaction.

Our final question for considering the relative importance of the four levels was:

- What are the implications of the performance of Reaction in this context?

Considerations in determining importance of the four levels

The table below summarises the questions for weighing up relative importance:

<table>
<thead>
<tr>
<th>Questions we considered when determining relative importance</th>
<th>Implication for importance determination</th>
</tr>
</thead>
<tbody>
<tr>
<td>Timeframe expectations:</td>
<td>Levels that have more immediate effects are relatively more important to determining the success of the program.</td>
</tr>
<tr>
<td>• What is the expected timeframe for effects at this level?</td>
<td></td>
</tr>
<tr>
<td>• How critical is this timeframe for program success?</td>
<td></td>
</tr>
</tbody>
</table>
• How sufficiently does the evaluation design factor in expected and unknown timeframes?

More critical timeframe expectations for latter levels increases their relative importance.

Causal relationships:
• How essential is this level for effects on latter levels?

Levels that are essential to effects in latter levels are relatively more important.

Sphere of accountability to external barriers and enablers
• How susceptible is this level to external barriers and enablers?
• What is the program’s intended level of control/influence over these barriers and enablers?

Levels within the program’s sphere of accountability are relatively more important.

Other implications for the performance of Reaction
• What are the implications of the performance of Reaction in this context?

Reaction increases in relative importance when there are other critical implications of its performance.

Concluding (for now)

This article has explored a range of issues which can influence the importance determination of the four levels within Kirkpatrick’s framework for evaluating training programs. Considering these issues and arriving at a decision on relative importance is part of the preparatory stage for synthesis. The questions presented enable a more explicit consideration of context and evaluation design constraints when judging the overall performance of the training program using Kirkpatrick’s four-level framework.

In an ideal world, an evaluation design and the findings presented would comprehensively cover elements discussed within these issues such as barriers and enablers (and the program’s impact on them if they intend to influence these), and factor these into the performance determination.

The reality is that evaluations must provide sufficiently well-justified judgements about the performance of a program within real-world constraints – with the available time, finances and data. One area for further development within this research is when more comprehensive findings are collected, how these can be incorporated in the importance determination guide.

Separating the values issues from the design issues within the importance determination process is one step to consider.

A further consideration is the interplay between the issues and how this impacts importance determination. For example, barriers and enablers will have implications on time lapses between effects at the latter levels.

This article presents some of the issues that could influence importance determination of Kirkpatrick’s four levels that were surfaced through the literature and through application of the framework. It is hoped that through sharing this work, other evaluators who have experience using the Kirkpatrick framework for evaluating training programs will share their examples and insights of its application across different contexts to support the development of a more comprehensive synthesis guide.

Further information

Lirata applies evaluation theory and extensive real-world experience to deliver innovative, context-driven and rigorous evaluations. If you would like to talk further about this article or about your evaluation needs and Lirata’s capabilities, please contact Fran Demetriou at Lirata Consulting:

Mobile: +61 (0)456 574 250
Landline: +61 (0)3 9457 2547
Email: fran.d@lirata.com

LIRATA LTD
contactus@lirata.com
www.lirata.com
References


