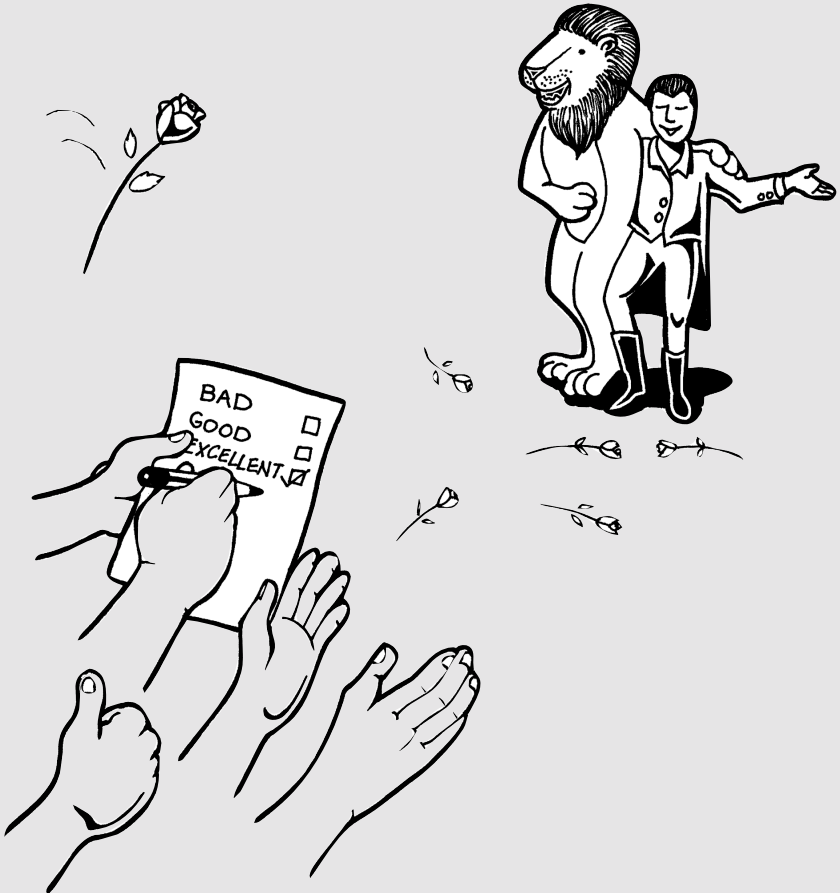


SECTION SEVEN

Evaluation





SECTION SEVEN

Evaluation

Let's begin with a simple definition of training evaluation. Properly conducted, training evaluation is a systematic means of analyzing the learning of our participants. By this we mean the how, what, and the how much of participants' learning. Evaluation should not be limited to handing out a questionnaire at the end of a training program. In this section we will describe how you can use results of evaluation to create better training programs and how to improve your sessions in future presentations. But there are additional possibilities. For example, other forms of evaluation can be used to justify your organization's continued support of your training department by showing that those who have completed its programs perform more effectively and help the organization achieve its broader goals.

In many publications on training, a typology of evaluation developed by Donald L. Kirkpatrick is mentioned. He described four levels of evaluation, with each building on the previous level: The first is "reaction." At this level, evaluation measures the reception of training programs among participants. At the second level evaluation investigates "learning" that occurred among the participants—how much did the program change their attitudes, increase their knowledge, or improve their skills? Evaluation at the third level, "behavior" asks whether those who attended the training program really did change their behavior when they returned to their work units. Finally, at the fourth level, "results," tries to assess whether a training program has benefited the organization (Kirkpatrick 1998, pp. 19-24). Of these, the most important is the first, the participants' reactions to the program. As Kirkpatrick notes, while measuring reaction to a program (level one) is something we should always do, "trainers should proceed to the other three levels as staff, time, and money are available" (Kirkpatrick, 1998, p. 24).

Recognizing that many who read this manual will face severe limitations in resources to conduct evaluations, our approach here will be to concentrate on the relatively simple things you can do on your own. The starting point is to recognize that you need to design a strategy for evaluation. In particular you want to decide before doing anything else what information you want and can use. It is wasteful to collect data that either do not answer the questions you want answered or focus on issues that are beyond your power to control.

In this section of the manual we will distinguish between two forms of evaluation. The first—formative evaluation—happens during the creation of a new training program and is aimed at improving its design. Summative evaluation can occur

either during a program or after it has been completed. The purpose in summative evaluation is to improve subsequent presentations of that training program. In the following discussion, we will review some of the methods you can use and describe some issues you may want to consider, such as providing participants with anonymity. Since questionnaires are frequently used we will also consider aspects of questionnaire design that can help you generate responses that genuinely reflect the trainees' experiences.

Formative evaluation

The purpose of formative evaluation is to improve a training program's design during the planning stage. As such, it builds on some of the things you discovered during the needs analysis described in section two. The fundamental question you will ask is whether the program you are designing will satisfy the needs you discovered. You will also want to assess whether the program would be likely to achieve the specific objectives you set—that is, those objectives you formulated following the process described in section four of this manual. In the first section we proposed that planning of training is a circular rather than a linear process, and this becomes very apparent during the formative evaluation process. The results from the needs assessment, the objectives you set yourself, and anything you discover through formative evaluation all influence each other. For example, you may conclude that one of the objectives was unrealistic. Or perhaps one objective is so different from the others that a separate training program would need to be developed.

But how do you conduct formative evaluation? You certainly cannot give a questionnaire to trainees who have taken the program, since it has not yet been presented. One approach is to ask for the responses of those who already possess the skills you are trying to develop. This can either be done individually or by bringing together a group. Show them what you are proposing to do and ask them for their observations. Do they think this would rectify the deficiency discovered in the needs analysis? Is there anything they think you have missed and that should be added? Alternatively, is there anything that they think could be omitted.

Sometimes people are hesitant to undertake formative evaluation because they worry that the results of pre-testing may not be valid if the program plan is not in its final form. This concern is unnecessary. We can illustrate this with a simple analogy. Many of you reading this will be familiar with the way advertising agencies regularly use "storyboards" to pretest ideas for a new television commercial. Since it is so expensive to produce the commercial advertising executives customarily present clients with a series drawing showing each element of the commercial and

how they will be sequenced. In the same way that advertising clients can make intelligent decisions based on storyboards, so too can you make good decisions about training through formative research.

Several years ago, one of the authors of this manual was working on a university undergraduate distance education course. A proposal had been made to use a series of audiocassettes as one component of the course and several ideas for innovative ways to structure them had been made. Nobody knew whether these ideas would work. Therefore, the author and a specialist in curriculum development sat in an empty office and recorded a sample cassette of what we thought might work. It was very crude. Indeed, I remember hitting the side of a glass with a pencil to produce the signal that we wanted to use to tell the student that this was the point to stop the tape and complete one of the accompanying written exercises. Later, when one of the audio engineers heard the cassette he was appalled by the poor quality and made us promise that we would make it clear to everyone that neither he, nor any of his colleagues, had anything to do with its production. But the crude cassette proved invaluable. We distributed copies to a number of the academics who were writing other parts of the course and even contacted a few students who had previously taken similar courses. Their responses were very helpful in planning the finished structure of the cassettes, and best of all, the whole process cost very little.

Collecting feedback during a training program

Questioning trainees who are currently taking a training program can serve both formative and summative evaluation purposes. Reactions at various points throughout the program can lead to changes in that presentation. For example, as part of a program lasting several days, participants could be asked to provide feedback at the end of the second day. The responses could be used to decide what should receive more emphasis in subsequent days. The other advantage of taking this approach is that, as we described in section five, you can enhance adult motivation by demonstrating that you value the knowledge and opinions of learners through your efforts to build a learner centered approach into your activities.

The same information you collected to decide whether adjustments were needed during presentation of a training activity can also be used later to determine whether to make permanent revisions in subsequent presentations of the program. It has been a common practice in some AIBD training courses to offer a mid-course review. This has proved particularly valuable in longer training courses where suggestions acquired in the review can be accommodated as the course continues to its conclusion.

Summative evaluation

As noted at the beginning of this section, evaluation should not be limited only to questionnaires collected from trainees at the end of a program. We say this recognizing that the most common way to collect feedback is through various types of questionnaires. However, alternatives do exist. For example, if training has been designed to develop a particular skill, you might want to incorporate an exercise into the later stages of the program that will both provide practice for the trainees in that skill and also give you some information on how much they have learned.

There is no single ‘correct’ way to design an evaluation questionnaire or the individual questions it contains. Regardless of how it has been constructed, a questionnaire is only successful if it supplies the kind of reliable information you need to make informed decisions while not placing undue demands on the respondents. As a result, questionnaire design is at least as much an art as a science.

For instance, there is no simple answer to the question: “How long is too long?” The number of questions you can ask, and the time you can ask respondents to take completing the items depends very much on the situation. When respondents are committed to the training activity, the questionnaire can ask for more than if they see it as something incidental to their lives. This is why in a training situation the questionnaire distributed at the end of the last day has limitations. Not only will those completing them probably be anxious to leave, they will also know that it is now too late for any information they give to make any difference, at least for them. Because questionnaires are so routinely used in training evaluation, participants may be skeptical of them, perhaps doubtful that anybody ever reads the responses. That is why evaluation at various points throughout the training program can underscore the trainer’s interest in gaining the learners’ honest reflections on their experience. Although it may initially sound strange, you may even want to consider evaluation before the training program begins. It might be very helpful to know what expectations the trainees are bringing to the first session.

Creating items for questionnaires

The first decision you need to make is whether to use open-ended questions, closed-response items, or a combination of the two. An open-ended question places no limits on the ways a respondent can answer. A blank space is provided into which the respondent enters his or her response. This type of question is particularly useful when you do not know the sorts of answers you might get and do not want to miss any possible thoughts or suggestions. The disadvantages begin to appear when you try to analyze the responses. First, it is very time consuming to read and organize all of the answers. From these responses, you have the task of discerning

patterns in the reactions of participants. This is not to say that the time it takes to analyze open-response questions makes them something you should avoid. Rather it is a warning that, while they are usually easier to write, the work simply moves to a later stage when you try to analyze them.

Closed-response questions force the respondent into selecting from a predetermined range of possible responses. There are several familiar formats for closed-response items. One of the most familiar is the multiple-choice question. Respondents are given a series of alternative answers to a question and asked to choose one, perhaps by circling a letter preceding it. Another version of this type of question asks for a “yes” “no”, or “true” “false” response. Checklists are also popular. Here the respondent is given a list of items and asked which would apply in a particular case, perhaps by checking a box next to everyone that applies. Ranking a list of items is also possible. In this kind of questionnaire the respondent is asked to order a list of items perhaps by placing a one (1) next to the item she or he considers most import and then, if there are five items, ranking the others two (2) through (5).

Trainees’ opinions about a topic can be by investigated by asking them to respond to a series of statements. For example: “It is important for a manager to regularly visit all of the units he or she supervises.” The respondents are then asked to indicate the strength of their agreement or disagreement with the statement by selecting from options such as: “I strongly agree”, “I agree” “I am undecided”, “I disagree”, “I strongly disagree.” Here, as in other forms of closed-response questions, a category of “other” can be included. Often this is accompanied by a request for a brief written explanation. However, if a large number of participants select this “other” option interpreting the results will be difficult and time consuming. In the worst scenario, you may find many respondents have selected the “other” option but have failed to explain the basis for their answer. Unless you can re-contact the respondents and ask for clarification, you will be left in the dark on that question. The advantage of closed-response questions is that, if properly constructed, the answers can be easy to tabulate and analyze. If you find yourself dealing with large groups, say many dozens or hundreds of respondents, computers can be used to scan the responses and analyze the data.

Although the techniques listed above are among the most widely used, this brief list certainly does not exhaust the formats available. Many introductory textbooks on social science research methods include examples of more. A more specialized source on questionnaire design, interviewing, and attitude measurement can be found in a book by A. N. Oppenheim (1992).

Because of the foregoing list of pro and con issues, we urge you to consider using a mixture of open and closed-response items. The combination of the two

can complement each other in useful ways. A particularly clear example of the value of combining the two types of questions is recalled by one of the authors of this manual. Several years ago he was conducting an evaluation of a series of television programs that were broadcast as part of a university undergraduate distance education course. One of the questions asked the students to rate each of the television programs in the first half of the course on a scale ranging from “very helpful” to “not at all helpful.” There was, of course, also a response option of “did not view” for each program since every student would not have seen every program. The second part of the question asked the students to explain how they had decided which programs were helpful. In other words, they were asked to describe on what basis they answered the question about “helpfulness.” The responses to this second part of the question were very time-consuming to analyze but the program producers, and the academics who had written the accompanying written materials found the students’ explanations very valuable in deciding what revisions should be made before the next presentation of the course.

A follow-up inquiry about results obtained using an open-response item can often provide insights into the way respondents have interpreted a question. This is helpful for questionnaire designers, since we cannot assume everyone will interpret our questions in the way we intended. It is impossible to entirely eliminate problems with questionnaire items, such as unexpected interpretations, but some techniques for avoiding some of the pitfalls are described below.

Avoiding problems with questionnaires.

The aim with all questionnaires, regardless of whether they are being used for formative or summative evaluations, is to collect data that reflect as accurately as possible the views of all members of the relevant population. In this case the term population is being used not to refer the residents of a city or country but rather the entire group of persons who have been trained (or in the case of formative evaluation, might be trained). Thus if you organized a workshop for twenty trainees, the population would be every one of the twenty who attended. Similarly, if you were interested in the views of all executive producers in a broadcast television organization, the population would be “all executives producers working for that organization.” For the purposes of this manual we will assume that you will always be trying to get information from everyone in the intended population. This is known as a census. When there are too many members of a population to contact everyone, the alternative to a census is to use a sample. However, to be successful and collect information from a sample that truly reflects the views of the whole population, including those who were not questioned, an appropriate sampling method needs to be employed. Constructing such a sample is beyond the scope of this manual, but you may be able to gain the assistance in your organization’s research department if you have one, or alternatively someone from a local university might be willing to help.

There are, however, several steps you can take to ensure that your questionnaires will gather the kind of information you need. Four techniques will be described here: the value of pretesting, improving the response rate on mailed questionnaires, the issue of respondents' anonymity, and avoiding problems with questionnaire wording.

Unless you have used a questionnaire previously and found it to be successful, it is always better to pretest (sometimes also called "pilot") a questionnaire. Even when you are using questions that others have used—perhaps taken from a book on training, or similar publication—it is still better to conduct a pretest whenever possible. This is especially true for courses offered in an international setting. Differences between cultures can mean that a question that worked well in one context may be interpreted very differently elsewhere.

One simple technique for pretesting is to find a few people who are representative of the population who will eventually be completing the questionnaire. Then ask those individuals to work their way through the questionnaire. Afterwards, ask each one whether they experienced problems answering any of the questions. You can also use the opportunity to ask questions that probe the way they had interpreted each question. Sometimes you will be quite surprised to discover that what seemed like a very straightforward question can be interpreted in very different ways. Based on the results of a pretest, however, revisions can be made to the questionnaire.

In the training workshop situation getting a high response rate to your questionnaires should be easy. However, when you are mailing questionnaires to people in other locations it may be quite difficult to get a large percentage of those to whom you have sent questionnaires to complete them and mail them back to you. People often ask what percentage you need to get back to know that the results reflect the views of the whole population. This can be a highly complex question to answer, but for our purposes a good rule of thumb is that the lower the percentage of questionnaires returned the more cautious you should be about the results. Consider a situation where only fifty percent of the questionnaires have been returned. Can you be certain that those who did not return the questionnaires would hold the same views as those who did? Or is it likely that those who responded were in some way different? In most cases, it is probably the latter.

A good example of this problem occurred a few years ago in the United States. An organization sent questionnaires to all television stations across the United States asking them how much free time for public service announcements (PSAs) their station had provided. The response rate was relatively low but the organization simply assumed that those who failed to respond would have answered in a way similar to those who did reply. Clearly, this was a highly flawed assumption. Here is why: A station that emphasizes PSAs and is proud of its record is far more likely

to respond than another that grudgingly transmits a few. If a station saw this as an important part of their public service commitment, it would probably have someone on staff designated to deal with PSAs. The questionnaire would have been routed to that person as soon as it arrived. At the less committed station it might well be passed from desk to desk with each hoping that someone else would complete it. Indeed, the questionnaire may merely have been discarded.

How then can you increase the likelihood that a larger percentage of questionnaires will be returned? First, keep the questionnaire as short as possible to collect the information you need, and indicate, honestly, how long you think it will take to complete. As discussed above, there is no simple answer to the question: "How long is too long?" But as a general rule the shorter the better.

Always explain to the participants why you need this information and what you will do with it. Include in the original mailing a copy of the questionnaire and a cover letter explaining the significance of the information they are being asked to provide and why they were selected. If you can be certain that the questionnaires will reach everyone within a week, consider sending a second mailing two weeks after the initial mailing. This can just be a short letter reminding them of the survey and asking them to contact you, if for some reason they had not received it. Finally, after a further two weeks send another copy of the questionnaire together with a new cover letter to all who have not responded. If someone has not replied in a month they may have lost or discarded the questionnaire, or at least by then it may be hard to locate, and a new copy of the questionnaire may prompt a response.

A third thing to consider in an effort to collect the best data from questionnaires, whether used in the training situation or distributed by mail, is to consider the issue of providing respondents with anonymity. We can all think of situations where our response to a question was influenced by the knowledge that a record of that response was being made and that our name would be attached to it. In an organization's internal training setting this is certain to be an issue. Since your aim in distributing questionnaires is to collect the most useful information, you should always consider whether you really need to know the identity of the respondent.

Perhaps there is some characteristic of the respondent that you do need to know. For example, you might believe that the responses to a questionnaire would be influenced by the number of years an individual had worked for that organization. But that does not mean you need to know the identity of each respondent. Perhaps it would be adequate to know approximately how many years they had worked. You could ask for this information in fairly broad categories such as 0-5 years, 6-10 years, more than 10 years. Provided the group is large enough, individuals will be less concerned that a particular response might be traced back to them, and thus tempted to write only what they think their superiors would want them to say.

Some readers may wonder how one can provide anonymity in a mail survey, if you need to know who has not replied in order to mail them reminders. Although not a perfect method, you can assign anonymous respondent numbers to each individual in the mailing list and include no other identification on the questionnaire. If you use this tactic, assure respondents that only you have access to the list of persons and their number assignments. Obviously, some will not believe you, and the success of this approach depends on the culture of the organization, but in some situations it may help to get the responses you need.

Finally, we will review a few of the potential problems that arise in questionnaire wording. While some may be specific to questions written in the English language we provide them because what we want to emphasize is the importance of thinking how a respondent might misinterpret a question. Again, pretesting can help to identify problems you have not previously thought of.

If there is a particularly significant word in the question that could lead to misinterpretation, if the respondent misses it, consider underlining it or using italics. A frequent example is the type of question that asks the respondent to agree or disagree with a statement that something should *not* be done.

It is also important to avoid “double-barreled questions. These are ones that ask for two pieces of information in the same question, leaving the respondent with a choice of which to answer. For example: “Do you agree that there is a problem with staff morale and that counseling sessions could improve the situation?” As you can see, the respondent might agree that there is a problem and yet believe the proposed remedy could make things worse. In this case, it is impossible to know how a respondent might answer this question.

Be careful when using words like “frequently”, “occasionally”, and “rarely.” There is no agreement on what these terms mean. Thus, the responses to a question asking how often someone reads a staff newsletter could be quite misleading. One person’s “occasionally” could be another’s “rarely.” Provide examples of the frequency of reading that would apply in each category.

Avoid providing inadequate alternatives in the response options. For example, at times you may want to force respondents into selecting between “agree” and “disagree” on a series of questions. But what if “don’t know” is a valid response? Or, if respondents are being asked to rate the performance of other units within an organization, what happens if someone has never had any dealings with a particular unit? A response option of “no experience on which to base a judgment” might be useful here.

Double-negatives should also be avoided. Consider the possibility for misinterpretation if a questionnaire included a statement like: “Increased resources

should *not* be given to units who have *not* demonstrated improved efficiency during the past year.”

Finally, consider the layout of the questionnaire as an important part of its design. The questionnaire should be straightforward, attractive, with clear directions. If someone answering “no” to the first question will be unable to answer the next four, include an instruction sending them directly to question six. Do not make them read items that are not relevant to them. This will either confuse or frustrate your respondents and it might reduce your response rate. Try to present questions in a logical sequence. Moving from general questions to the more specific is often appropriate. However, if you need to collect some general information that only provides a background to the main focus of the questionnaire you may want to leave this until last. Starting with those background questions may make some respondents decide that the questionnaire is not important. In this case, beginning with questions that clearly relate to the topic of the questionnaire would be better.

Evaluating learning outcomes

Apart from gaining the thoughts of your trainees on the planning of programs, evaluation should additionally investigate whether learning objectives have been satisfied. This matches Kirkpatrick’s second level of evaluation mentioned at the beginning of this section. You may remember that in section four we made the point that training is expected to do the following things:

- (1) to develop cognition or to improve the knowledge of participants in specific job situations;
- (2) to develop skills to complete the required psychomotor tasks efficiently; and
- (3) to develop attitudes or beliefs which make participants more effective in their work.

In order for you to determine whether a course has met its learning objectives, you must decide whether participants have developed the skills, knowledge and attitudes laid out by the training objectives. Measuring the terminal behavior in each of the three kinds of learning requires different techniques. So in the following discussion, we will consider the evaluation of cognitive, skill, and attitude development separately.

First, let’s look at the evaluation of cognitive learning. Cognitive learning is the increase of knowledge—the learning of facts and the relationships among facts. To use another familiar term, one might say that cognitive learning means gaining new information. Information gain is the result of cognitive learning—one knows more as a result of cognitive learning. Objectives for cognitive learning usually state that participants will be able to demonstrate knowledge of specified terms, concepts, or ideas. Consider this then—the words “will be able to demonstrate” tell us what we must do to evaluate training aimed at this kind of objective. We

must arrange a test that allows participants to *demonstrate* their knowledge. The most common way to evaluate cognitive learning is to allow participants to demonstrate their knowledge through what is known as an objective test. An objective test is one in which a trainer gives a participant a problem or a question for which there is only one correct or acceptable response. Note our stress that in objective measures there can be only one proper answer; all others must, by definition, be wrong. The objective test can take several different forms. Which of these types of objective measures you select is not really important to evaluation of learning. Your choice will probably be based upon your own preferences. If you favor a specific type of objective test, then use it.

A checklist is the most common way to evaluate psychomotor skill development. A checklist contains all of the required operations or steps necessary to demonstrate mastery of skills stated in the training objectives. The required order of the steps are indicated in the checklist. To develop a checklist one begins by breaking down a task into the smallest practical set of specific actions. The manner of stating each step on the checklist should imply how to determine whether the action was properly performed. The checklist may be similar in form to ones developed in the job analysis carried out during training needs analysis. It is important to remember that many skills require steps or operations that a skilled individual might take for granted. The purpose of a checklist is to make certain that all of the physical actions required by a psychomotor skills task have been carried out. Through the checklist, it should be possible to pinpoint problem areas in skill performance.

Attitude development is likely to be more difficult to evaluate than other types of learning. Part of the problem is that attitudes are apt to be less clearly defined than cognitive learning. One of the many difficulties in measuring attitudes is that often learners' feelings on certain subjects are hidden, even from themselves. Or ever if they are aware of their attitudes that may not fully understand them. Haven't you heard someone say that he or she prefers a particular thing, but when that person is actually required to make a decision, the choice is different? Because attitudes are often made up of contradictory or confusing notions, people may seem to be unaware of their own beliefs. A further complication in measuring attitudes is the difficulty in being completely honest with others about our beliefs. We all want to be accepted by others, so it can be awkward to admit we hold an attitude that others might find disagreeable or peculiar. As a result, we tend to claim attitudes that are conventional or ones we believe others are likely to accept. For this reason, individuals tend to give respond in ways they think we would like. And this usually occurs without any consciousness on their part. Finally, attitudes are difficult to measure because we are not only concerned with what attitudes are held, but also with the strength of those attitudes. We all may agree that it is important to report to our office on work days at the scheduled time. But just how strong is this belief in punctuality? Is it important to be on time just when it is

convenient? Or when traffic does not interfere? Or, should we be on time even when there is a serious illness in our family? This issue in attitude measurement obviously makes the evaluation task different from tests of skills or knowledge, where the objective is merely to observe whether learning goals have been met.

Although there are many ways of measuring attitudes, the most common and the easiest to use is the multiple choice question asking for a response to an attitude statement as described earlier. Questions of this type consist of two parts. The first is a statement about an attitude or opinion; the second is a set of five (occasionally fewer or more) choices: agree strongly, agree, neutral, disagree, or disagree strongly. The object is to have persons select one of the five choices that is closest to their own thinking about the matching statement. By doing this, you will have gained information about both respondents' attitudes and the strength of those attitudes. If the number of questions about a set of attitudes is large enough, it is possible to gain an overall impression of respondents' attitudes. This is done by combining or adding together the results of all the questions.



References

Kirkpatrick, D.L. (1998). *Evaluating Training Programs: The Four Levels*. (2nd ed.). San Francisco: Berrett-Koehler.

Oppenheim, A. M. (1992). *Questionnaire Design, Interviewing, and Attitude Measurement*. (Rev. ed.). London: Pinter.