NINE

SYNTHESIS METHODOLOGY

Like merit determination, synthesis is another task that is very specific to evaluation. It is the tool that allows us to draw overall evaluative conclusions from multiple findings about a single evaluand.

> **Synthesis** is defined as "the process of combining a set of ratings or performances on several components or dimensions into an overall rating" (Scriven, 1991, p. 342).

Synthesis is most relevant to the Overall Significance checkpoint in the Key Evaluation Checklist (KEC) (Exhibit 9.1). This is where the evaluation team needs to combine all of the evaluative information gleaned from looking at Checkpoints 6 through 10 (Process Evaluation, Outcome Evaluation, Comparative Cost-Effectiveness, and Exportability) to draw overall conclusions about the evaluand.

Exhibit 9.1 The KEC Checkpoints Where Synthesis Methodology Is Used

> **11. Overall Significance**
>
> Draw on all of the information in Checkpoints 6 through 10 to answer the main evaluation questions such as the following. What are the main areas where the evaluand is doing well, and where is it lacking? Is this the most cost-effective use of the available resources to address the identified needs without excessive adverse impact?

The form of synthesis covered in this chapter is not to be confused with meta-analysis or literature reviews. These involve summarizing or combining the findings of multiple research or evaluation studies (about different evaluands) to draw overall conclusions about the relationships among variables. Meta-analysis uses a very specific statistical technique to give a weighted average of effect sizes across multiple studies. As such, it can handle only quantitative studies. In contrast, a literature review uses the reviewer's judgment, rather than an explicit technique, to synthesize studies.

There is a substantial overlap between the merit determination and synthesis steps in an evaluation. Many readers likely noticed that the rubrics we used to combine a mix of data in the merit determination chapter are in fact a very simple synthesis methodology. In this chapter, we take that basic logic further with some more systematic methods that can handle more complex data.

SYNTHESIS: WHAT AND WHY

Nearly any evaluand has a range of strengths and weaknesses—some more important than others—that we need to consider when we draw evaluative conclusions about quality or value on a particular dimension or component or about the evaluand overall. After all, doing poorly on some aspect of minimal importance is less serious than doing poorly on something crucial. This is why we need synthesis methodology—to have a systematic way of taking into account the pluses and minuses uncovered when the evaluation team draws evaluative conclusions.

Erroneous Arguments Against Doing Synthesis at All

At this point, it is worth presenting again the typical argument against the use of synthesis:

> This book [*Assessing Organizational Change*] is largely silent on the issue of combining outcomes from different domains in order to reach an overall conclusion about the effectiveness of a change effort. This is by design. The decision was made early on to simply report how the organization had changed on a wide array of outcome measures. No common metric was developed, nor was a weighting system developed that argued that gains in some measures are more important than gains in others. The rationale for not doing this is simple and to us persuasive. It is that different constituents value outcomes

> differently, and thus it is best to let interested parties reach their own overall conclusions. There are also practical problems in trying to translate diverse outcomes to a common metric. (Lawler, Seashore, & Mirvis, 1983, p. 542)

Hopefully, the holes in Lawler and colleagues' (1983) argument are becoming more apparent to the reader as we progress through this book. In Chapter 6, we learned that the definition of "value" in a well-designed evaluation is derived from multiple defensible sources, including the needs of impactees, ethics, the law, and relevant professional standards. Therefore, the claim that "different constituents value outcomes differently" is much less problematic than it first appears because good evaluation does not rely on personal values.

In Chapter 7, we took this further and showed why the relative importance of certain dimensions of merit can and should be determined using much more than just the opinions of individual stakeholders. Certain outcomes, for example, can often be shown objectively to be of greater or lesser value (e.g., to the organization or the community) than certain other outcomes.

In Chapter 8, we learned a method for translating diverse outcomes to a "common metric," that is, merit ratings (excellent, very good, good, acceptable, or poor). In this chapter, we tackle the task of *combining* multiple ratings of merit in a way that takes their relative importance into account.

The Need for Synthesis at Multiple Points in the Evaluation

In the course of doing an evaluation, there are multiple points where some form of synthesis is required. One that we have already encountered in the chapter on merit determination is when multiple sources of data (often both qualitative and quantitative) are combined with quality or value "standards" (definitions of "how good is good") to provide an explicitly evaluative rating on a particular dimension. From that point, there may be another one or two steps (or perhaps more) required to generate quality or value ratings on broader dimensions, on components of the evaluand, and/or on the evaluand as a whole.

An important point to note is that some synthesis is always necessary, whether the evaluation is formative or summative in nature. If an overall conclusion about the quality or value of the entire evaluand is needed (this is always the case for summative evaluations and is quite often the case for formative evaluations), a full synthesis will be needed. If the evaluand's quality

or value simply needs to be reported on several dimensions or components, synthesis will stop short of the final step. But however the evaluative conclusions are reported, there is still a need to combine multiple findings or sources of data to draw those conclusions.

The synthesis step involves one additional piece beyond combining the data we have collected about the current evaluand (process, outcomes, and cost). We also use comparisons in the synthesis step to help place the evaluand's performance in a wider context. This is where the Comparative Cost-Effectiveness checkpoint of the KEC comes into play.

THE ROLE OF COSTS AND COMPARISONS IN SYNTHESIS

Every evaluand requires resources to be created and maintained. And whenever resources are allocated to something, this is always at the expense of whatever else might have been done with the same resources (i.e., opportunity costs). Therefore, whether the evaluation is formative or summative, the question is not just "Did the value of the outcomes outweigh the value of the resources it took to achieve them?" Rather, it is always "Is/Was this evaluand the best possible use of available resources to achieve outcomes of the greatest possible value?"

For evaluations where the primary evaluation question is a "ranking" one (e.g., Which of these three innovative pilot programs best meets the needs in this community? Which job candidate should we hire?), the evaluation team needs to go into considerable detail on each alternative and make very explicit comparisons. We cover some methodologies for doing this when we look at synthesizing for "ranking" later.

In cases where the primary evaluation question is a "grading" one (e.g., Was this executive training program worth implementing?), the comparisons are used to put the evaluation findings in context to allow better interpretation of merit rather than to make explicit and detailed comparisons with all possible alternatives (in most cases, this would be "paralysis by analysis"). Several methodologies for doing this are covered in the next section on synthesizing for "grading."

In some cases, comparisons do not play a large part in the synthesis step, whereas in other cases, they are extremely important. For example, when selecting someone for a job from a short list of three persons, comparisons are central because this is a ranking task. In performance appraisal, an employee's

performance can often be interpreted purely in terms of the value of his or her contributions to the organization without needing to compare these to what others have achieved.

For program evaluation, comparisons in some form are almost always necessary. They can either be worked into the synthesis steps very explicitly or be used to place the synthesized findings in a broader context for interpretation.

SYNTHESIZING FOR "GRADING"

Rubrics are one of the simplest methods for blending (or synthesizing) data. But in some cases, the nature of the data is a little more complex, making it difficult to use a rubric as the only tool. For example, the data about evaluand performance that are gathered might not all be equally important or reliable, and we might need to have some way of taking this into account when we use the rubric. In addition, there might be so many different sources of data with different nuances and combinations that it becomes extremely difficult to determine merit reliably using a rubric and the limited powers of the human brain.

As mentioned previously, there are several different options when it comes to synthesis methodology. Which one is used depends, first, on whether the main evaluation question is an "absolute" (grading) or a "relative" (ranking) one. (For a review of these terms, refer back to Chapter 2.) The second consideration when selecting the right synthesis methodology is whether a qualitative or quantitative synthesis method is to be employed. A quantitative synthesis methodology is one that uses numerical weights that are applied using multiplication. A qualitative synthesis methodology is one that uses qualitative labels that are applied without the use of multiplication. In the following sections, an example is given for four different evaluation tasks that answer a grading or ranking question using quantitative or qualitative methods.

When the primary evaluation question is one of absolute quality or value, we are seeking to answer questions such as the following:

- How well did the evaluand perform on this dimension?
- How effective, valuable, or meritorious is/was the evaluand overall?
- Is/Was this component worth the resources (e.g., time, money) put into it?

Quantitative (numerical) Weighting Example With "Bars"

The first example is drawn from **personnel evaluation** (specifically, performance appraisal). It uses a quantitative (numerical weighting) methodology with some twists to generate an overall performance rating for an employee for the review period. This overall rating is then used to determine performance-based rewards.

The setting is a small accounting firm. The process began with a clear definition of the main tasks performed by employees. For this particular organization, 13 separate tasks were defined (e.g., telephone and reception, data entry for particular types of client support packages, tax agency database management). Each employee typically had responsibility for approximately 4 to 6 of these tasks in any one quarter.

Next, each task was given an importance weighting using Strategy 2 (drawing on the knowledge of selected stakeholders) from Chapter 7. After an in-depth discussion, the business owners agreed on a definition of "importance" at four levels as follows:

1. Minor task (worth doing but not particularly important for the success of the business)
2. Normal-priority task
3. High-priority task (very important for the success of the business)
4. Extremely high-priority task (crucial for the success of the business)

Note the small number of levels of importance here. In general, it is best to define approximately three to five levels. Anything more fine-grained than that makes it extremely difficult to get good agreement on importance. With a 10-point scale, for example, it is possible to waste hours arguing whether a particular task should be weighted 6 or 7. In reality, there is seldom any need for this level of precision, so for practical reasons, the three- to five-level rule works well in most cases.

Having established importance weightings, the next step was to draw up rubrics for each of the 13 tasks. Here, approximately four to six levels were usually sufficiently fine-grained to capture the variation in performance without wasting a lot of time deciding which category should apply in a particular case. An example of one of those rubrics was shown in Table 8.4 in the previous chapter.

Each employee was rated on four to six tasks (depending on which ones fell within the employee's job responsibilities), receiving a score from the following scale:

1. Totally unacceptable performance
2. Mediocre (substandard) performance
3. Good performance (expected level)
4. Performance that exceeded expectations
5. All-around excellent performance

Each task had an importance weighting ranging from 1 (minor task) to 4 (extremely high-priority task), as outlined earlier.

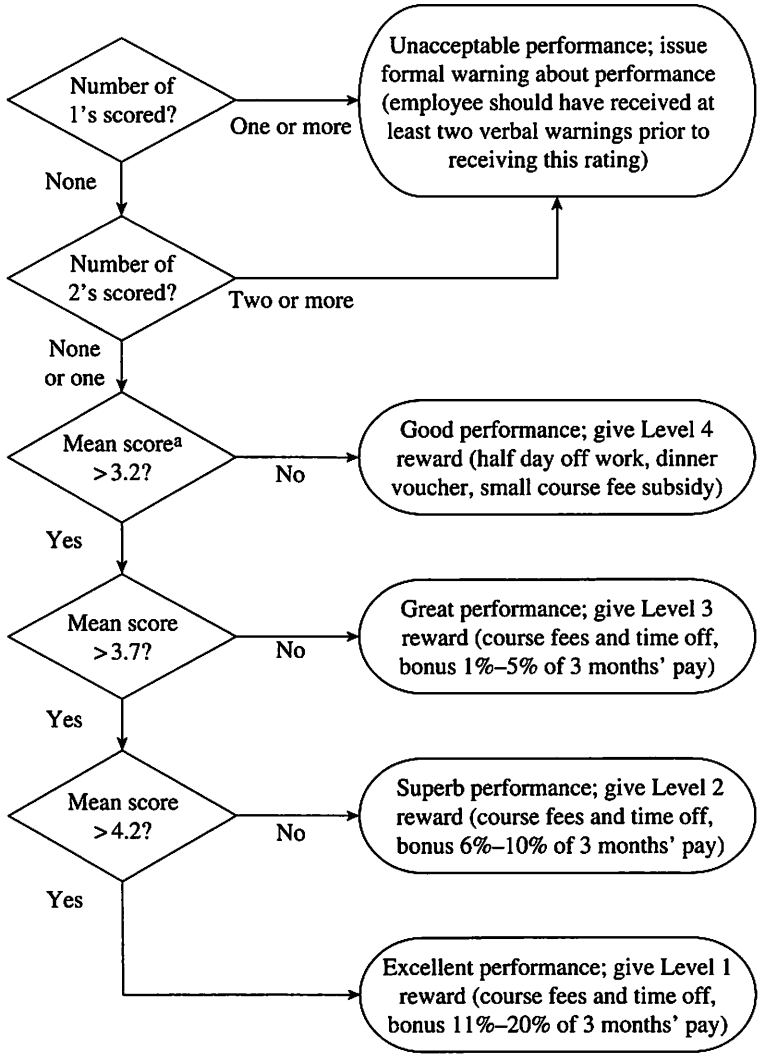
The synthesis methodology used to generate an overall performance rating for each employee incorporated both a numerical weighting strategy (weighted average of performance ratings) and bars (i.e., minimum acceptable levels of performance, in this case, on each task). A rating of 1 on any task was defined as totally unacceptable performance regardless of how well the person did on his or her other tasks. Any more than one rating of 2 was also defined as unacceptable. The synthesis algorithm is shown in Exhibit 9.2.

Qualitative (nonnumerical) Weighting Example 1 (with no "bars")

A quantitative weighting strategy works well in simple cases provided that bars are included to ensure that very poor performance on a particular dimension is not inappropriately masked by better performance on other dimensions. But numerical weighting systems often lead to conclusions that have the evaluation team staring at a conclusion that seems not to be quite right, thereby leading to the temptation of juggling the weights until the answer looks right. Such cases call for a qualitative (i.e., nonnumerical) weighting strategy.

The qualitative synthesis strategy presented in this section is a simple, step-by-step method for synthesizing mixed evaluation information that was developed by the author and a colleague for an evaluation of a school-based health program (Mersman, 1999; Mersman & Davidson, 1999). Those readers who are used to working with large-scale evaluations might find this example somewhat unimpressive. This was a real shoestring evaluation—short timeline, low budget, and little access to data. We too were amazed to find that even the

Exhibit 9.2 Synthesis Algorithm for a Performance Appraisal and Rewards System in a Small Accounting Firm



NOTE: a. All mean scores are weighted means.

relatively simple data generated by such a small-scale evaluation project presented a synthesis problem that we found could not be solved satisfactorily using any existing methodology. The good news is that once we developed the methodology, it was clear that it could be applied even to situations where there were many more criteria. So, the intent here is to provide a small simple

example that is easy enough to grasp as well as to deliver some principles and methods that could be applied to a more elaborate evaluation.

The evaluand in this case was a school-based health program with approximately nine different components that were aimed specifically at students: nutrition education, mental health services, case management for pregnant and parenting teens, safer sex, legal services, and several others. The client needed very quick and approximate answers to the question of how well each of these components was meeting important needs of the students and their families.

As just mentioned, this was the proverbial shoestring evaluation. Because of time and budgetary constraints, apart from limited observation and one or two interviews, the main data collection device was a short survey of students who had used the services. The survey was devised in both Spanish and English and was filled out in class in the presence of a bilingual teacher who explained each question in Spanish after the lead evaluator had explained it in English. Three questions about each program component were asked: two quantitative (using 4-point response scales) and one open-ended. Students were asked, "How useful was the [nutrition education] program to you?" "How satisfied were you with the program?" and "What other changes or events, good or bad, have happened to you or someone you know because of [receiving the service]?" A brief summary of the responses for the nutrition education component of the program is shown in Table 9.1.

The content of both the open-ended answers and the first quantitative item addressed the extent to which the program met student *needs,* which is a more important consideration than **satisfaction,** the other quantitative item (which deals with *wants*). Therefore, we had two sources of information about needs and one about satisfaction. Of the needs-related information, the open-ended responses contained rich information, but we obtained responses from only a third as many people as completed the quantitative items. The question was, How could we combine the results systematically in a way that took account of both the centrality to the needs issue (which was the main evaluation question) and the fact that some of the data were more representative of the users of the program due to the higher rate of response?[1]

Taking into account both of these considerations (and after much debate between us), we prioritized the three different sources of data as follows (1 = strongest data, 3 = weakest data):

1. Ratings of usefulness (directly related to needs and reasonably representative of the students who used the program)

Table 9.1 Summary of Responses to Questions About the Nutrition Component of a School-Based Health Program

How useful was the program to you?	Not at all useful 1.6%	Somewhat useful 23.8%	Useful 57.7%	Very useful 15.9%	$N = 63$
How satisfied were you with the program?	Not at all satisfied 1.6%	Somewhat satisfied 11.3%	Satisfied 69.4%	Very satisfied 17.7%	
What other changes or events, good or bad, have happened to you or someone you know because of [receiving the service]?	"I have lost weight and I am more healthy, I believe." "When I was pregnant, I was avoiding junk food and eating more nutritious food." "I found that I am anemic and got all my shots." "Now when I eat something from a bag, I check the nutrition. It helped me a lot." "My sister is taking it seriously and losing weight. I am too." "First I was trying to eat healthy, but then I didn't care anymore and went back to junk food."				$N = 20$

2. Responses to open-ended question about effects of the program (directly related to needs, rich and descriptive data, but weaker on repesentativeness)
3. Satisfaction ratings (useful information to add to the mix, strong on representativeness, but addresses met wants or satisfaction rather than needs, which were more important)

Having identified which criteria were to be considered primary, our next task was to come up with a way in which to convert each of these three pieces of raw data into explicit determinations of merit. For example, should the usefulness ratings for the nutrition program be considered satisfactory, good, or excellent? What should we make of the mix of positive and negative statements collected in the open-ended responses?

The greater challenge was presented by the quantitative data because they lacked the richness of evaluative content in the open-ended comments. What spread of ratings on the usefulness question should be considered poor, marginal, satisfactory, good, or excellent? Obviously, this is by no means a precise science, but we needed to have at least some sort of broad-brush evaluative rating to be able to clearly answer the client's questions.

To convert the usefulness ratings into determinations of merit, we used two strategies. The first was to look across the ratings for the nine different program components and to see how the distribution of scores stacked up against the others. As we looked across the different ratings of usefulness, there seemed to be some natural splits. The lowest-rated program component had only 57% of students rating it as useful or very useful, the highest-rated programs were close to 90% or higher, and the other program components were bunched in the middle between 70% and 80%.

To work out how we should characterize the performance of programs that fell into each of these three "clumps," we went back to the qualitative data to see what sorts of comments were associated with these categories. It was clear from this that the kinds of outcomes being produced by even the lowest-rated services were by no means indicative of dismal performance. On the other hand, having only approximately half of the respondents consider a service useful meant that this could not rightly be called a good outcome either; this judgment was based purely on what we believed was reasonable common sense. Balancing these considerations, we labeled the bottom category "adequate" (or, in shorthand form, we called it a "C"). At the other end of the scale, the comments associated with some of the services with very high usefulness ratings (90% or higher) were clearly indicative of extremely valuable outcomes for the students. Accordingly, we labeled the top category "excellent" (or shorthand "A") and the middle category "good" (or shorthand "B"). Based on similar logic (and due to similar rating distributions), we used the same rubric for the satisfaction ratings (Table 9.2).

Table 9.2 Rubric for Converting Quantitative Data to Determinations of Merit

Merit Rating	*Evidence*
Excellent (A)	Approximately 90% or more of respondents rated the service as useful or very useful (or said that they were satisfied or very satisfied)
Good (B)	Approximately 70% to 90% of respondents rated the service as useful or very useful (or said that they were satisfied or very satisfied)
Adequate (C)	Approximately 50% to 70% rated the service as useful or very useful (or said that they were satisfied or very satisfied)

When the usefulness ratings were converted to merit ratings, we used the preceding categories as broad guides but applied grade adjustments depending on where the evidence fell within the defined range. For example, approximately 74% of students rated the nutrition component as useful or very useful, so this component was assigned a rating of B/B– on usefulness.

Note that this is an extremely simple example due to the broad-brush information needs in this particular evaluation. For evaluations that require more factors to be taken into account, it is advisable to create a more sophisticated rubric. In the next chapter, we discuss an example of a synthesis rubric used for another project where the author took many more factors into account. But for now, let's stick with the simple example and follow it through.

Next, we had to convert the qualitative responses to explicit determinations of merit. Using the fundamental principles underlying the basic rubric in Table 8.2 in the previous chapter, we created a rubric of our own to convert the qualitative data into merit ratings (Table 9.3). The ratings were defined in terms of both the strength of the evidence and the magnitude of the impact described in the responses.

Note that the top end of the merit scale includes not only requirements for the magnitude and volume of positive comments but also some limitation on the magnitude and volume of negative comments. Including guidelines for both positive and negative elements is often essential for creating a good merit determination rubric.

The rubrics in Tables 9.2 and 9.3 allowed us to convert the three types of data into explicit determinations of merit. The next step was to take these three determinations of merit and combine them to draw an overall conclusion about the nutrition component (and then other components) of the school health program (Exhibit 9.3).

To combine these three sources of data (now all converted to determinations of merit), we used a step-by-step process that began with the strongest source of data (usefulness ratings) to provide us with what we called a "working grade." We then blended in the data from the open-ended comments that would influence the working grade by up to half a grade depending on how incongruous they were with the quantitative usefulness ratings. For example, the usefulness ratings for the mental health component were low (C), but the open-ended comments showed a strong positive impact. Therefore, the working grade for the mental health component was adjusted up to B/C. Finally, the satisfaction ratings were taken into account. These could influence the working grade by up to a third of a grade. The full working for this step-by-step merit determination process is shown in Table 9.4.

Table 9.3 Rubric Used for Converting Data From the Open-Ended Responses Into Merit Ratings for the Nutrition Component of the Health Program

Merit Rating	*Evidence*
Excellent	*Evidence of a strong positive impact:* very positive comments, with a substantial number that indicated a very strong impact; few if any neutral or negative comments
Good	*Evidence of a noticeable positive impact:* a good number of positive comments (few neutral or negative), clearly showing that the program had made a noticeable positive effect on students
Satisfactory	*Evidence of some positive impact:* a mix of positive and negative comments, skewed somewhat toward the positive; evidence pointing in the right direction but not to a very noticeable impact
Marginal	*Little or no impact either way:* a real mix of comments; no clear skew in either the positive or negative direction
Poor[a]	*Evidence of some negative impact:* a mix of positive and negative comments, skewed somewhat toward the negative; not enough evidence to call this a really noticeable negative impact

NOTE:

a. Categories lower than "poor" (e.g., "completely unacceptable") were not defined because none of the program components was performing that poorly. (There was no point in doing unnecessary work.)

Exhibit 9.3 Synthesizing Three Data Sources to Draw Evaluative Conclusions About the Merit of the Nutrition Component of the School-Based Health Program

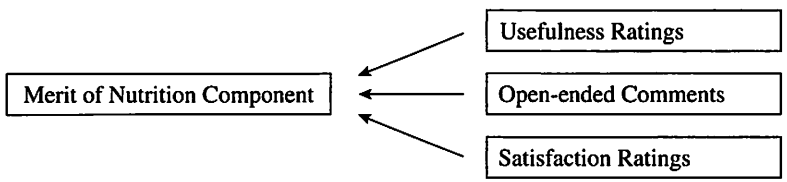


Table 9.4 Step-by-Step Determination of Program Component Merit for a School-Based Nutrition Program

Program Component	*Working Grade*		*Type of Effects From Responses to Open-Ended Questions*[a]		*Adjusted Working Grade*	*Satisfaction Level*		*Final Grade*	*Overall Component Rating*
Transportation	A	→	Some negative (need more of service) and noticeable positive	→	A–	Low to moderate	→	A/B[b]	Extremely good
Legal	A	→	None	→	A	Low to moderate	→	A/A–	Close to excellent
Case management for pregnant or parenting teens	A/B	→	Little to some negative[c]	→	B	High	→	B+	Very good
Anatomy	B/B+	→	No open-ended question asked	→	B/B+	Moderate to high	→	B/B+	Good to very good
Safer sex	B	→	Noticeable positive	→	B+	Moderate to high	→	B+	Very good
Nutrition	B/B–	→	Noticeable positive	→	B+	High	→	A–	Extremely good to excellent

Program Component	*Working Grade*		*Type of Effects From Responses to Open-Ended Questions*[a]		*Adjusted Working Grade*	*Satisfaction Level*		*Final Grade*	*Overall Component Rating*
Health educator or clinic nurse	B/B–	→	Some negative as well as noticeable positive	→	B–	Moderate	→	B–	Good to adequate
Siblings pregnancy prevention	B–	→	Little to some negative[c]	→	B/C	High	→	B/B–	Close to good
Mental health	C	→	Very noticeable positive	→	B/C	Low	→	C+	Adequate to good

NOTES:

a. Comments from students and staff.

b. The negative comments conveyed a need for more of the service since it was discontinued.

c. The negative comments conveyed that high turnover had led to a lack of continuity in these services.

The process described here for synthesizing multiple determinations of merit to rate program components might seem like overkill for a small program like this. That may well be true given the miniscule size of the evaluation budget. The main reasons for doing this were (a) clarification and improvement of the evaluator's logic in synthesizing a mix of data, (b) making that logic clear and transparent for the client and other interested parties, and (c) as an opportunity for learning and methodology development that could be applied to other projects. Hopefully, it will now be useful for others to apply or develop.

Qualitative (nonnumerical) Weighting Example 2 (with "hurdles")

The third synthesis example also uses a qualitative (nonnumerical) weighting system. It also uses a variation on bars called "soft hurdles" and "hard hurdles." Soft and hard hurdles were developed specifically for an evaluation of an organization's learning capacity (Davidson, 2001), which is the example used here.

A **bar** is a minimum level of performance on a *specific dimension,* performance below which cannot be compensated for by much better performance on other dimensions, for example, a rating of 1 (totally unacceptable performance) in the small accounting firm's performance appraisal system described earlier.

A **hard hurdle** (Davidson, 2001) is an *overall passing requirement* for an evaluand as a whole, for example, no more than one rating of 2 (mediocre [substandard] performance) in the small accounting firm's performance appraisal system described earlier. If the evaluand or evaluee fails to meet the requirement, he, she, or it fails overall. Hard hurdles are referred to elsewhere as "global bars" (Scriven, 1991).

A **soft hurdle** is an overall requirement for entry into a high rating category (Davidson, 2001). Unlike a bar, it does not automatically classify an evaluand as "failed" (i.e., it is nonfatal); rather, it places a *limit* on the maximum rating that can be achieved if the evaluand does not clear a particular soft hurdle (e.g., to get an overall A for a course, none of the assignments completed during the semester can be lower than a B–).

The specific example to be used here was an evaluation of the learning capacity of a small biotechnology start-up company in the United States referred to here as "Biosleep." Biosleep's performance was rated on 27 subdimensions of organizational learning capacity (Table 9.5). This performance profile was derived from survey and interview data, a merit determination rubric similar to Table 8.2 in the previous chapter, and importance determination Strategy 6 (using program theory and evidence of causal linkages) from Chapter 7. The comparative element, in this case with other organizations, was built directly into the merit determination rubric.

The synthesis in this case required two steps. First, the performance ratings on the subdimensions needed to be packed together to make determinations of merit on the eight main dimensions. From there, the final step was to combine the performances on the eight dimensions to draw an overall conclusion about the organization's learning capacity.

Subdimensions → Dimensions

Using the hurdle principle, an algorithm was created to guide the way in which conclusions at the dimensional level could be derived from different combinations of subdimension ratings. A synthesis algorithm that included soft hurdles was used (Table 9.6). Because there was no compelling evidence to suggest that any of the subdimensions should be given greater consideration (or weight) than any of the others, each subdimension was treated equally in the synthesis.

For each organizational learning dimension, the median rating on the relevant subdimensions (of which there were between two and five) was the initial criterion used to determine the probable overall rating. The reason for this was twofold. First, it ensured that extreme ratings on one subdimension did not have a disproportionate effect on results. Second, it avoided making the erroneous assumption that the rating categories represented an interval scale (as would have been required if a "mean" rating had been computed).

Based on the algorithm in Table 9.6, a more condensed organizational learning capacity profile for Biosleep that summarized its performance on the eight main dimensions was generated (Exhibit 9.4).

Dimensions → Overall Evaluation

Having profiled Biosleep on the eight learning culture dimensions, the final task was to synthesize this information one step further to draw overall

Table 9.5 Subdimensional Learning Culture Profile for Biosleep

Dimension	*Subdimensions*	*Performance Rating*				
		P	S	G	VG	Ex
Experimentation (extremely important)	Support for risk taking					
	Diversity of practice or methods					
	Marketplace for ideas					
	Continuous improvement					
Practicing excellent evaluation (extremely important)	Tapping true value in personnel evaluation					
	Flexible use of goals					
	Multiple evaluative perspectives					
	Customer needs focus					
	Benchmarking and comparisons					
Mental models (very important)	Valuing diversity of thought					
	No "sacred cows"					
	Open communication and trust					
Shared vision (very important)	Shared vision and purpose					
	Shared sense of identity					
	Using own good judgment					
External or future scanning (very important)	External or market scanning					
	Future and scenario scanning					
	Openness to change					
Personal mastery (moderately important)	Striving for excellence					
	Seeking out criticism					
	Seeing the performance gap					
Team learning (desirable)	Team synergy or intelligence					
	Dialogue and debate					
	Cross-project communication					
Systems thinking (desirable)	Understanding interdependence					
	Seeing systemic causes					

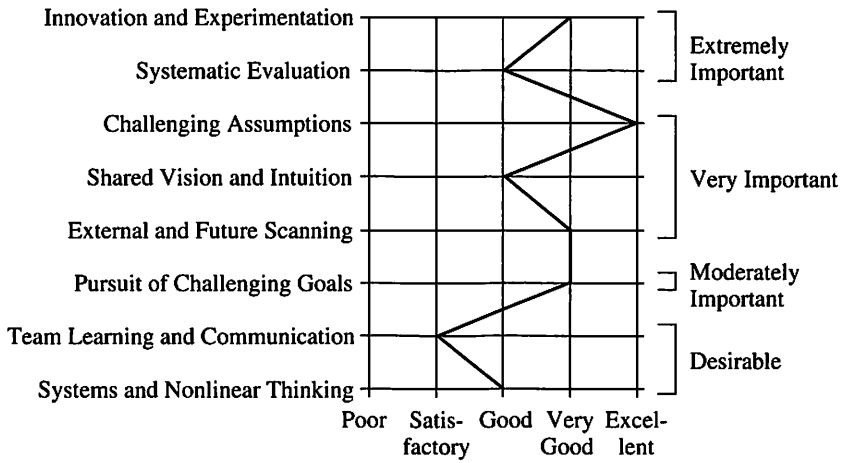
NOTE: P = poor; S = satisfactory; G = good; VG = very good; Ex = excellent.

Table 9.6 Guidelines for Synthesizing Subdimensions Into Dimensions

Dimensional Rating	*Median Subdimension Rating*	*Subdimensions Below "Good"*	*Subdimensions Below "Satisfactory"*
Excellent	Excellent	0	0
Very good	Very good (or better)	0	0
Good	Good (or better)	< 35%	0
Satisfactory	Satisfactory (or better)	(No restrictions)	< 35%
Poor	(No restrictions)	(No restrictions)	(No restrictions)

NOTE: Conditions in all three columns must be met to receive the corresponding rating.

Exhibit 9.4 "Dimension" Learning Culture Profile for Biosleep



conclusions about the organization's learning capacity. Using similar logic to the dimensional synthesis step, an algorithm was constructed to combine performance on both the outcome dimensions and the eight dimensions of the learning culture to yield an overall evaluative conclusion about each organization's learning capacity (Table 9.7). Using these guidelines, Biosleep was rated as having a *high* organizational learning capacity.

Table 9.7 Guidelines for Determining Overall Organizational Learning Capacity

Organizational Learning Capacity Rating	*Conditions for Learning Culture Dimensions*[a]		
	All Learning Culture Dimensions	*Extremely Important and Very Important Dimensions*	*Moderately Important and Desirable Dimensions*
Extremely high learning capacity	• No ratings below good • At least six excellent ratings	• All extremely or very important dimensions rated very good or higher	• Maximum one dimension rated good; all others rated very good or excellent
Very high learning capacity	• No poor ratings • At least six ratings in the very good to excellent range	• Maximum one dimension rated good; all others rated very good or excellent	• Maximum two dimensions rated satisfactory; all others rated good or higher
Biosleep High learning capacity	• At least six ratings of good or better	• Maximum one dimension rated satisfactory; all others rated good or higher	• No more than one poor rating
Moderate learning capacity	• No more than two poor ratings	• No poor ratings	• (No restrictions)
Learning impaired	• One or more poor ratings on extremely or very important dimensions *or* more than two poor ratings overall		

NOTE:

a. Conditions in all three columns must be met to receive the corresponding rating except where noted otherwise.

Note that the logic here is more complex than simple averaging. To receive a high overall rating for its organizational learning capacity, an organization must have several dimensions of organizational learning capacity in the highest performance categories, with very few dimensions in the lowest performance categories. In line with the logic underlying hurdles, stricter limitations were placed on those dimensions classified as extremely important or very important.

The criteria in each cell (e.g., requiring at least six dimensional ratings of excellent or higher for an extremely high overall learning capacity rating) were based on the author's own judgment and a careful read of the literature on organizational learning. The rationale was that to receive the top rating, an organization should have a clear majority of its learning culture dimensions rated as excellent. Given that there were eight dimensions altogether, six was seen as a "clear majority."

As with any attempt to make such cutoff criteria very explicit, there might well be room for debate and revision of these cutoffs. For example, if future research finds that organizations with only five learning culture dimensions rated excellent consistently outperform virtually all competitors in terms of adaptiveness, survival, and financial performance, this would provide a solid rationale for revising the cutoff down to five. In the absence of such a body of evidence, the intent here was to propose a viable working methodology that would be subject to constant refinement and improvement as the relationship between organizational learning and effectiveness or survival is better understood.

SYNTHESIZING FOR "RANKING"

Chapter 2 explained why, for some evaluations, the evaluation team needs to determine the quality or value of something relative to one or more other evaluands. Examples might include published rankings of graduate programs, comparative evaluation of several different pilot programs or interventions, evaluation of job candidates for hire or promotion, and rankings of consumer products (e.g., in *Consumer Reports,* in computer magazines, in car magazines).

When the primary evaluative task is ranking, there is a need for a much more explicit treatment of the Comparisons checkpoint in the KEC. Rather than simply using information about "critical competitors" to place performance data in context for evaluative interpretation, a ranking evaluation requires much more detail about the critical competitors and much more in-depth comparison.

Once again, there is a choice to be made regarding the use of numerical synthesis methods or nonnumerical synthesis methods: numerical weight and sum (NWS) versus qualitative weight and sum (QWS).

Numerical Weight and Sum

Numerical weight and sum is a quantitative synthesis methodology (i.e., one that uses numerical importance weights and merit scores) for summing evaluand performance across multiple criteria.

NWS involves ascribing numerical importance weights and a numerical performance score to each evaluative dimension, multiplying weights by performance scores, and then summing these products. The resulting sum represents the overall merit of the evaluand.

NWS works adequately for ranking provided that (a) there are only a small number of criteria, (b) there is some other mechanism for taking bars into account, and (c) there is a defensible needs-based strategy for ascribing weights. Perhaps the most important thing to understand about NWS is some of the problems that can occur when the first two of these three conditions are not met. These are illustrated in the following hypothetical example.

Suppose that you had been asked to conduct a comparative evaluation of three different interventions for training managers: (a) a mountain retreat featuring interactive sessions with multiple world-class management gurus, (b) an in-house training and mentoring program run by human resources, and (c) a set of videos and the latest book on management from management guru Peter Drucker.

We start with a brainstorming session to identify the main dimensions of merit (under the headings of Process, Outcomes, and Cost) and follow this up with a needs assessment to make sure that we understand the nature and extent of the needs and the relative importance of the various dimensions of merit. This leads us to generate a list of the main dimensions to examine when comparing the three interventions (Table 9.8).

The next step is to collect relevant data about each of the three evaluands on these dimensions and to rate their performance using any one of five ratings: excellent, very good, good, fair, or poor (for hypothetical ratings, see Table 9.9).

Table 9.8 List of Dimensions for the Evaluation of Management Training Interventions

Dimension of Merit		*Importance*
Process dimensions	Content matches needs of participants	Important
	State-of-the-art management content	Desirable
	Tailored to organization's strategic needs	Desirable
	Professionally presented	Desirable
	Interesting for participants	Desirable
	Useful materials for later referral	Desirable
	Built-in follow-up for transfer of training	Important
	Networking opportunities	Desirable
	Quality of facilities provided	Desirable
	Quality of refreshments	Desirable
Cost dimensions	Time costs for participants (away from job)	Extremely important
	Financial cost of training (to organization)	Important
Outcome dimensions	Impact on effective management of people	Extremely important
	Impact on communication and persuasion	Important
	Impact on organizational skills	Desirable

The next step is to convert the importance weights and performance ratings into numerical form so that we can complete the NWS. In Table 9.10, the weight "extremely important" was given a 3, "important" was given a 2, and "desirable" was given a 1.[2] For the performance ratings, numerical scores of 5, 4, 3, 2, and 1 were used to represent the ratings of excellent (5) through poor (1). The dimensions have also been ordered by importance to make it easier to comprehend the main differences among the evaluands.

The final step is the final synthesis step. Here we take each performance rating and multiply it by the importance weight. Then all of these products are summed and the evaluands are compared (Table 9.11).

Table 9.9 Evaluation of Management Training Interventions on All Dimensions

Dimension of Merit	*Ratings for:*		
	Retreat	*In-House*	*Video*
Content matches needs of participants	Good	Very good	Fair
State-of-the-art management content	Excellent	Fair	Very good
Tailored to organization's strategic needs	Excellent	Good	Fair
Professionally presented	Excellent	Fair	Very good
Interesting for participants	Excellent	Fair	Fair
Useful materials for later referral	Excellent	Fair	Good
Built-in follow-up for transfer of training	Poor	Very good	Fair
Networking opportunities	Excellent	Good	Poor
Quality of facilities provided	Excellent	Fair	Poor
Quality of refreshments	Excellent	Fair	Poor
Time costs for participants (away from job)	Poor	Good	Excellent
Financial cost of training (to organization)	Poor	Good	Excellent
Impact on effective management of people	Good	Very good	Good
Impact on communication and persuasion	Very good	Fair	Poor
Impact on organizational skills	Poor	Fair	Good

Two important points should be apparent from Table 9.11. First, the mountain retreat option won largely because it did extremely well on nearly all of the "trivia" (aspects of the evaluand that are desirable if good but not important),

Table 9.10 Numerical Importance Weights and Performance Ratings for All Three Management Training Interventions

Dimension of Merit	*Importance Weight*	*Performance (unweighted)*		
		Retreat	*In-House*	*Video*
Impact on effective management of people	3	3	4	3
Time costs for participants (away from job)	3	1	3	5
Built-in follow-up for transfer of training	2	1	4	2
Content matches needs of participants	2	3	4	2
Financial cost of training (to organization)	2	1	3	5
Impact on communication and persuasion	2	4	2	1
State-of-the-art management content	1	5	2	4
Tailored to organization's strategic needs	1	5	3	3
Professionally presented	1	5	1	4
Interesting for participants	1	5	2	2
Useful materials for later referral	1	5	2	3
Networking opportunities	1	5	3	1
Quality of facilities provided	1	5	2	1
Quality of refreshments	1	5	2	1
Impact on organizational skills	1	2	2	3

even though it did poorly on many of the important dimensions. In more complex evaluations where the number of criteria is often extremely large, it is common for minor considerations to "swamp" the major ones in this way. Increasing the numerical size of the importance weightings can correct this problem, but this also drives "tie-breaker" considerations so far into the background that they can hardly influence the conclusion at all.

Table 9.11 Numerical Weight and Sum Synthesis for Three Management Training Interventions

Dimension of Merit	*Importance Weight*	*Importance × Performance*		
		Retreat	*In-House*	*Video*
Impact on effective management of people	3	9	12	9
Time costs for participants (away from job)	3	3	9	15
Built-in follow-up for transfer of training	2	2	8	4
Content matches needs of participants	2	6	8	4
Financial cost of training (to organization)	2	2	6	10
Impact on communication and persuasion	2	8	4	2
State-of-the-art management content	1	5	2	4
Tailored to organization's strategic needs	1	5	3	3
Professionally presented	1	5	1	4
Interesting for participants	1	5	2	2
Useful materials for later referral	1	5	2	3
Networking opportunities	1	5	3	1
Quality of facilities provided	1	5	2	1
Quality of refreshments	1	5	2	1
Impact on organizational skills	1	2	2	3
Sum of Importance × Performance		72	66	66

The second important point is that the mountain retreat was by far the most expensive option in terms of both time and money costs, being rated as poor on both. In many organizational contexts, it is likely that one or both of these costs could simply be too high. This means that the retreat should not have been rated as the top intervention because it was not cost-feasible. In other words, there needed to be a bar on one or both cost criteria. This may also have been true for some of the other dimensions besides cost.

The reality is that although NWS seems simple and intuitive, it can often leave the evaluation team looking at a conclusion that does not seem quite right. The temptation at that point is often to fiddle with the numbers to see whether the right answer can be coaxed out of the data. An alternative is to work with a synthesis strategy that incorporates the key elements of how the human brain naturally weights considerations, making them explicit so that they can be applied to larger numbers of dimensions. Qualitative Weight and Sum is such a strategy.

Qualitative Weight and Sum

Qualitative weight and sum (QWS) is a non-numerical synthesis methodology devised by Scriven (1991) for summing the performances of an evaluand on multiple criteria to determine overall merit or worth. QWS is a ranking methodology for determining the relative merit of two or more evaluands or evaluees. Typical applications include **personnel selection,** comparisons of experimental or pilot programs to decide which to roll out elsewhere, and selection among competing products, services, or proposals. QWS is not suitable for grading (i.e., determining absolute merit).

Step 1: Determine Importance in Terms of Maximum Possible Value

The first step is to assign each criterion a qualitative importance rating by determining whether the maximum possible value of excellent performance on that criterion should be considered extremely valuable, valuable, or marginally valuable. Importance ratings should be determined using the appropriate combination of strategies from Chapter 7.

To each importance label, assign the appropriate symbol from the list provided in Table 9.12.

Table 9.12 Maximum Possible Value of Excellent Performance on Each Dimension of Merit for Management Training Interventions

Dimension of Merit	*Maximum Possible Value*	*Symbol*
Content matches needs of participants	Valuable	▲
State-of-the-art management content	Marginally valuable	✚
Tailored to organization's strategic needs	Marginally valuable	✚
Professionally presented	Marginally valuable	✚
Interesting for participants	Marginally valuable	✚
Useful materials for later referral	Marginally valuable	✚
Built-in follow-up for transfer of training	Valuable	▲
Networking opportunities	Marginally valuable	✚
Quality of facilities provided	Marginally valuable	✚
Quality of refreshments	Marginally valuable	✚
Time costs for participants (away from job)	Extremely valuable	★
Financial cost of training (to organization)	Valuable	▲
Impact on effective management of people	Extremely valuable	★
Impact on communication and persuasion	Valuable	▲
Impact on organizational skills	Marginally valuable	✚

Step 2: Set Bars

For each criterion, regardless of importance, determine whether there is any completely unacceptable level of performance, that is, performance so poor that even excellence on all other criteria would not compensate (e.g., a price so high that consumers or funders simply could not afford it even if

the project is highly meritorious in other respects). This cut point between acceptable and unacceptable is called the bar. For each criterion that has a bar, describe what "completely unacceptable" performance would look like.

For example, if the participants in the training must spend 2 weeks or more away from their jobs, that might be considered an unacceptable time cost. Similarly, financial costs of more than $5,000 per participant might be considered too high for a particular organization. Based on the needs assessment, the evaluation team might also specify some minimum level of match between the training content and the individuals' needs.

Step 3: Create Value Determination Rubrics

This step is basically equivalent to creating a merit determination rubric (see Chapter 8) except that with the focus on value levels, it is more accurately referred to as a value determination rubric. It is best to start with the simplest rubrics and work your way up.

Start with the dimensions you have weighted as "marginally valuable." Define or describe what performance on each dimension would look like at two levels: marginally valuable and no noticeable value. (If the dimension also has a bar, you will have definitions for three possible levels of performance: marginally valuable, no value, and completely unacceptable.) Each definition should describe in evaluative terms (preferably referring to a mix of required qualitative and quantitative evidence) what performance would look like at each level. A simple example for rating the training in terms of networking opportunities for participants is shown in Table 9.13.

Table 9.13 Rubric for Rating Management Training Interventions on Networking Opportunities

Value Level	*Symbol*	*Description*
Marginally valuable	+	Sufficient opportunity for networking so that participants could develop business relationships
No noticeable value	(Blank)	Little or no opportunity for networking with other participants

Now take any dimensions you have weighted as valuable. Define or describe what performance on each dimension would look like at three levels: valuable, marginally valuable, and no noticeable value. Again, there will be a fourth level of performance if you have defined a bar, as in the example shown in Table 9.14, a rubric for rating the financial cost of training.

Table 9.14 Rubric for Rating Management Training Interventions on Financial Cost of Training

Value Level	*Symbol*	*Description*
Valuable	▲	Extremely cheap (money cost less than $400 per participant)
Marginally valuable	✚	Moderately priced (money cost between $400 and $1,500 per participant)
No noticeable value	(Blank)	Quite expensive (money cost between $1,500 and $5,000 per participant)
Unacceptable	✘	Excessively expensive (money cost in excess of $5,000 per participant)

Repeat the step for any criteria weighted as extremely valuable, this time defining four possible levels of performance (from extremely valuable to no value) as well as defining below the bar if there is a bar. Table 9.15 gives an example rubric for one of the most important dimensions, that is, impact on effective management of people.

Step 4: Check Equivalence of Value Levels Across Dimensions

The validity of the QWS method is highly dependent on ensuring that the value levels defined for each dimension are roughly equivalent. To check this, look across your criteria at the evaluative definitions you have created for each level of value and consider what they convey about equivalent value or trade-offs. One way in which to do this (especially when you are just starting with QWS) is to put the information into a matrix such as the one in Table 9.16.

Table 9.15 Rubric for Rating Management Training Interventions on Impact on Effective Management of People

Value Level	*Symbol*	*Description*
Extremely valuable	★	Very substantial impact[a] on participants' effective use of people management strategies on the job
Valuable	▲	Significant (but not substantial) impact on effective use of people management strategies on the job
Marginally valuable	✚	Just noticeable (but not significant) impact on effective use of people management strategies on the job
No noticeable value	(Blank)	No noticeable impact on effective use of people management strategies on the job
Unacceptable	✘	Noticeable detrimental impact on people management strategies used on the job

NOTE:

a. Evidence used included (a) interviews with participants 3 months after completing the program, (b) "360-degree" feedback (i.e., performance ratings gathered from direct reports, peers, and senior managers), and (c) employee turnover rate in the business unit.

Looking across the rows in Table 9.16, the evaluation team should consider whether each defined level of performance has a sufficiently similar value or whether it is more similar in value to entries in the rows above or below. One way in which to think about this is in terms of trade-offs by taking diagonal pairs. For example, according to the matrix in this table, a very cheap training program (< $400 per participant) that has a just noticeable impact on people management performance should be about as valuable to the organization as a moderately priced one (say, $1,000 per participant) that has a significant (but not substantial) impact on the same outcome. If one of the two also offered good networking opportunities, that would be the "tie-breaker" that would lead to the choice of that program.

The other thing to check is where the levels of maximum possible value have been set. Note that the impact on people management has a highest

Table 9.16 Comparison Matrix for Assessing the Equivalence of Value Levels Across Dimensions

Maximum Value	*Impact on People Management*	*Financial Cost per Participant*	*Networking Opportunities*
★	Very substantial impact on participants' effective use of people management strategies on the job		
▲	Significant (but not substantial) impact on effective use of people management strategies on the job	Extremely cheap (money cost less than $400 per participant)	
✚	Just noticeable (but not significant) impact on effective use of people management strategies on the job	Moderately priced (money cost between $400 and $1,500 per participant)	Sufficient opportunity for networking so that participants could develop business relationships
(Blank)	No noticeable impact on effective use of people management strategies on the job	Quite expensive (money cost between $1,500 and $5,000 per participant)	Little or no opportunity for networking with other participants
✘	Noticeable detrimental impact on people management strategies on the job	Excessively expensive (money cost in excess of $5,000 per participant)	

possible rating of extremely valuable, whereas the other two dimensions have lower maximum values. This means that if one training program produced a very substantial impact on people management performance and was moderately priced (say, $1,000 per participant), an alternative program that produced

a significant (but not substantial) impact could not beat it *no matter how low the cost* (assuming that ratings on all other dimensions were equal).

These trade-off comparisons, which may be made in collaboration with stakeholders, are essential for testing the validity of the value determination rubric. This phase of QWS often requires some adjustment of the bars and/or rubric categories.

Step 5: Rate Value of Actual Performance on Each Dimension

For each evaluand (there must be more than one evaluand because this is a ranking exercise), use the value determination rubric to ascribe a value rating (e.g., valuable, marginally valuable) to its performance on each dimension (Table 9.17).[3] Note that the rating on any criterion cannot be higher than the maximum possible value weighting you have assigned to that criterion.

Step 6: Tally the Number of Ratings at Each Level and Look for a Clear Winner (if evident)

Sum the number of ratings of each type separately (extremely valuable, very valuable, valuable, marginally valuable, no value, and completely unacceptable) for each evaluand (see the bottom rows of Table 9.17). Throw out any evaluands with unacceptable ratings (✘). Then look to see whether there is a clear winner.

For the two training programs still in the running, the difference is one ▲ versus three ✚. Because there is no fixed formula for how many ✚ are equivalent to one ▲, there is not yet a clear winner between the two. This is a key difference between QWS and NWS. In NWS the numbers make up your mind for you, whereas in QWS you are forced to stop and think explicitly about the trade-offs.

Step 7: Refocus

In the refocus step (Table 9.18), we drop the columns of evaluands that did not make the first cut (i.e., the mountain retreat). We also delete the rows on which the remaining evaluands score the same (i.e., the extent to which the training was tailored to the organization's strategic needs).

Table 9.17 Initial Value Ratings on Each Dimension for All Three Management Training Interventions (QWS)

Dimension of Merit	*Maximum Value*	*Actual Value*		
		Retreat	*In-House*	*Video*
Impact on effective management of people	★	▲	★	▲
Time costs for participants (away from job)	★	✘	▲	★
Built-in follow-up for transfer of training	▲		▲	✚
Content matches needs of participants	▲	✚	▲	✚
Financial cost of training (to organization)	▲	✘	▲	▲
Impact on communication and persuasion	▲	▲	✚	
State-of-the-art management content	✚	✚		✚
Tailored to organization's strategic needs	✚	✚	✚	✚
Professionally presented	✚	✚		✚
Interesting for participants	✚	✚		
Useful materials for later referral	✚	✚		✚
Networking opportunities	✚	✚	✚	
Quality of facilities provided	✚	✚		
Quality of refreshments	✚	✚		
Impact on organizational skills	✚			✚
Totals		—	—	—
★			1	1
▲		2	3	2
✚		9	4	7
✘		2		

Table 9.18 Refocus Step for Remaining Two Management Training Interventions (QWS)

Dimension of Merit	*Maximum Value*	*Actual Value*	
		In-House	*Video*
Impact on effective management of people	★	★	▲
Time costs for participants (away from job)	★	▲	★
Built-in follow-up for transfer of training	▲	▲	✚
Content matches needs of participants	▲	▲	✚
Financial cost of training (to organization)	▲	✚	▲
Impact on communication and persuasion	▲	✚	
State-of-the-art management content	✚		✚
Professionally presented	✚		✚
Interesting for participants	✚		
Useful materials for later referral	✚		✚
Networking opportunities	✚	✚	
Quality of facilities provided	✚		
Quality of refreshments	✚		
Impact on organizational skills	✚		✚
Totals		—	—
★		1	1
▲		3	2
✚		3	6

At this point, the evaluation team members need to examine exactly where the differences lie so that they can work out whether the three additional

marginally valuable elements offered by the video instructional program outweigh the one valuable element from the in-house training. The team may also consider reweighting the dimensions in light of the smaller range of values between the two remaining evaluands. For example, if the video training costs $350 per participant and the in-house training costs $800, it might be true for this organization that financial cost is now a relatively minor matter and should be weighted as a tie-breaker (✚) instead of a more important dimension (▲).

NOTES

1. Unfortunately, no information was available regarding how many students had actually attended the nutrition education program. Therefore, all we knew about the representativeness of the data was that it was higher for the quantitative items.

2. Clearly, there is a need to put considerably more thought into what the exact numerical weightings should be if NWS is to be used effectively. These values were chosen to maximize the illustrative value of the example.

3. In this case, I have derived these from the original performance ratings in Table 9.10. This would not usually be necessary, however, because when using QWS, the evaluands should be rated directly on each dimension using the value determination rubrics outlined earlier instead of being rated first on an excellent-to-poor scale and then converted.

ADDITIONAL READINGS

Entries in Scriven's (1991) *Evaluation Thesaurus:*

- Apples and oranges
- Linear combination approach
- Meta-analysis
- Numerical weight and sum
- Qualitative weight and sum
- Synthesis (in evaluation)
- Synthesis (of research studies)
- Unconsummated evaluation

Scriven, M. (1994). The final synthesis. *Evaluation Practice, 15,* 367–382.

EXERCISES

1. Design a quantitative synthesis algorithm to draw an evaluative conclusion about the overall grade for an evaluand of your choice. Make sure that

you explain what the evaluand is and clearly lay out what the dimensions of merit are. If you can find actual data for the evaluand, use them to try out the algorithm.

2. Find a copy of *Consumer Reports* and identify a product you might consider buying in the future and that the magazine has rated on several dimensions. Identify three or four products to compare. Supplement this information with additional information (e.g., from the Internet) if necessary. Do a QWS to determine which you should buy. The QWS should involve at least one refocus step.