

The book cover features a vibrant, geometric design with four distinct color zones: a red top section with ruler patterns, a green middle section with technical drawings, a blue bottom-middle section with grid and dollar signs, and a yellow bottom section with a fine grid. The title is centered across these zones.

*Using*  
**MULTIPLE  
MEASURES  
to REDEFINE  
SUCCESS**

*Six Steps to Making It Happen*

**W. James Popham**  
University of California, Los Angeles



*Using*  
**MULTIPLE  
MEASURES  
to REDEFINE  
SUCCESS**

*Six Steps to Making It Happen*

**W. James Popham**  
University of California, Los Angeles



**Deborah S. Delisle**, Executive Director and CEO

**Ronn Nozoe**, Associate Executive Director

**David Griffith**, Sr. Director, Government Relations

**Melissa Mellor**, Outreach Manager, Government Relations

**Andrea Hoffman**, Sr. Production Specialist

**Katie Freeman**, Sr. Associate Editor

**Melissa Johnston**, Sr. Graphic Designer

---

Founded in 1943, ASCD is the global leader in developing and delivering innovative programs, products, and services that empower educators to support the success of each learner. Comprising more than 125,000 members—superintendents, principals, teachers, professors, and advocates from more than 138 countries—the ASCD community also includes 54 affiliate organizations. The nonprofit’s diverse, nonpartisan membership is its greatest strength, projecting a powerful, unified voice to decision makers around the world. The association provides expert and innovative solutions in professional development, capacity building, and educational leadership essential to the way educators learn, teach, and lead.

---

**[www.ascd.org](http://www.ascd.org)**

1703 N. Beauregard Street

Alexandria, VA 22311

1-800-933-2723

1-703-578-9600

*Using*  
**MULTIPLE  
MEASURES  
to REDEFINE  
SUCCESS**

*Six Steps to Making It Happen*

Foreword	vii
A Concept's Permutations	2
How Does Multimetric Educational Evaluation Work?	9
A Six-Step Evaluation Framework	10
A Brief Look Back	27
Resources	29
About the Author	30



# Foreword



The launch of ASCD's Whole Child Initiative in 2007 heralded a bold new vision for education encompassing what we expect of educators, what we need from parents, and—most importantly—what we want for students. Working together, our mutual goal is to ensure that each child in our care is healthy, safe, engaged, supported, and challenged.

It is at once an entirely commonsensical yet indisputably audacious concept. At its essence, ASCD's whole child efforts are about redefining student success. Families intuitively understand that fully preparing students for college, career, and citizenship requires rich learning experiences and meaningful measures of achievement. Similarly, educators recognize that comprehensive support and services are essential for helping students reach this more ambitious goal.

Frustration has mounted over the past decade about the No Child Left Behind Act's test-based accountability system. The extreme emphasis on standardized test scores in just a few subjects as the defining measure of student achievement and school quality has too often impeded efforts to broaden the curriculum; incorporate additional subjects; or promote other important skills and abilities, such as social and emotional learning.

In our years of promoting whole child education, we have concluded that, for better or worse, the accountability system drives priority-setting and decision-making processes in education. Thus, new evaluation models are needed to allow a whole child approach to flourish. Fortunately, the enactment of the Every

Student Succeeds Act (ESSA) offers just such an opportunity. To the redound of Whole Child advocates around the country, the law gives states the authority to develop their own accountability systems that require a range of student performance measures for school quality and must include non-academic indicators.

We must take full advantage of this opportunity to put in place an array of indicators that fully reflect a more comprehensive definition of student success, accurately measure student learning, and more systematically track educators' efforts to engage and support learners.

I am so pleased that ASCD is publishing this resource by the renowned assessment expert W. James Popham for policymakers, educators, and the public. We believe this will be a useful guide to state and local decision makers as new, next-generation accountability systems are considered and developed across the nation. Of course, each state and local community must determine for itself the ideal outcomes for its students. That is why this guide offers a step-by-step process for identifying and evaluating the criteria by which progress toward these goals can be measured.


It is our strong belief that the more widespread adoption of multiple measures has great potential to support a whole child education—for the benefit of students and educators alike. Together, we can make it happen. We hope you will join us in this cause.



**Deborah S. Delisle**

Executive Director & CEO





Certain procedures are patently praiseworthy. Judging books by their contents rather than their covers or using all the relevant evidence to determine a defendant's guilt or innocence are examples of inherently laudable processes. Such procedures make so much sense because, well, they make so much sense.

In the education field, one such instantly applauded notion is that when we set out to determine the quality of an educational endeavor, we ought to rely on more than one source of evidence when arriving at our evaluative conclusion. In recent years, educational organizations and educators have often urged policymakers to evaluate the caliber of an educational program only after reviewing a number of different factors indicative of that program's quality.

In an attempt to increase educators' reliance on diverse evaluative factors when they appraise an educational program's success, ASCD has recently undertaken a major initiative designed to promote greater use of evaluation strategies that, instead of relying on only a single source of evaluative evidence, focus on multiple kinds of evaluative evidence. The following analysis describes the chief features of how a truly defensible multifocused evaluative strategy might function.

## A Concept's Permutations

This guide describes what *multimetric educational evaluation* is and supplies one continuing, illustrative example of how such an evaluative approach could be implemented. ASCD is providing this analysis not as an endorsement of a particular incarnation of multimetric evaluation, nor of how such evaluations must be carried out. Rather, what is described should be regarded as an illustration of *one way* that such an evaluative strategy might be implemented in the real world.

For the purposes of this guide, here is the definition of the multimetric educational evaluation to be employed in the following pages:

Multimetric educational evaluation consists of an attempt to appraise the quality of a specific educational intervention according to two or more evaluative criteria regarded as indicative of such an intervention's success.

The key features of this concept's definition are the *interventions* being evaluated, the *evaluative criteria* used in such evaluations, and the *evaluation contexts* in which such an evaluative strategy is appropriate. As noted above, a continuing example will be supplied to illustrate how a multimetric evaluation might be undertaken.

### The Intervention

Multimetric educational evaluation—as is true with most versions of educational evaluation—is focused on determining the quality, or the worth, of a particular educational intervention. Such interventions include a wide array of products and procedures educators use to enhance students' learning. For example, if classroom teachers experiment with specific instructional tactics for more

intensive peer collaboration, then the instructional tactics would be the intervention to be evaluated. Similarly, if a teacher attempts to encourage students to discover—on their own—not only what needs to be learned, but also how to learn it, then “discovery learning” would be the intervention to be evaluated.

Another frequently evaluated category of educational interventions includes what goes on in a school for an entire year—the whole works. We see such evaluations of regular school programs when schools and districts are evaluated each year—often by emphasizing students’ performances on a state’s annual accountability tests. These high-stakes exams are typically characterized as accountability tests, but at their heart they are simply being used to evaluate an educational intervention—in this instance, an entire year’s worth of schooling.

### **The Evaluative Criteria**

An evaluative criterion, as indicated in the above definition, is a factor that’s employed when evaluators try to determine the quality of an educational intervention. As the definition asserts, a multimetrix evaluation approach requires the use of *two or more* evaluative criteria. More often than not, the evaluative criteria employed to appraise the worth of educational interventions—typically instructional ones—center on the amount of student learning that has occurred as a consequence of whatever intervention is being evaluated. Moreover, this amount of student learning is almost always signified by students’ performances on achievement tests.

Regrettably, despite numerous calls from educators and parents for educational evaluations to be based on multiple evaluative criteria, far too many appraisals of educational interventions still hinge exclusively on students’ test scores. Even worse, rather than relying on measures of students’ learning provided by diverse kinds of achievement tests, in most instances we see evaluators relying on students’ performances on only a single achievement

test—with scores on that test functioning as a solitary evaluative criterion. Multimetric educational evaluation represents a clear attempt to combat the narrowmindedness associated with single-criterion educational evaluations.

Other than students' scores on educational achievement tests, what evaluative criteria might play a role in evaluating an educational intervention's caliber? For starters, we can certainly consider students' performances on *different* achievement tests—tests designed to measure the same construct or one close to it. For example, suppose an English teacher sets out to measure students' composition prowess and decides to use two types of tests to do so. In one test, the student must choose from three options a topic about which to compose “from scratch” a 500- to 800-word expository essay. In a second test, the teacher uses 30 multiple-choice items to measure students' knowledge of routine conventions dealing with sentence structure and punctuation. Both tests measure aspects of a student's composition prowess, but each test gauges an aspect of composition skill in a different way. In general, if doing so is not prohibitively expensive, using different kinds of achievement tests measuring the same sorts of evaluative criteria constitutes one reasonable way of getting a better fix on how much students have learned. But would reliance on multiple ways of measuring the same, solo evaluation criterion actually constitute multimetric educational evaluation? Here's an instance when reasonable educators can sensibly disagree.

You see, according to the definition of multimetric evaluation being used in this analysis, if only one evaluative criterion were to be employed when arriving at a quality judgment about an intervention, this would not represent *bona fide* multimetric educational evaluation. The evaluative approach being advocated here requires the use of two or more *different* evaluative criteria. Nonetheless, if an educational evaluation were to be carried out using several forms of evidence for the same *solo* evaluative criterion,

would such a diverse-evidence evaluation not be superior—other things being equal—to an evaluation in which a single evaluative criterion were represented by only a single sort of evidence? Of course, it would.

As you can see, then, an educator might reasonably quarrel with the definition of multimetric educational evaluation given here, a conception calling for the use of *multiple* evaluative criteria, preferring instead to rely on multiple sorts of evidence related to a single evaluative criterion. The evaluation strategies stemming from either of those approaches will surely be preferable to evaluations in which only one evaluative criterion is represented by only one type of evidence. The definition being used in this analysis is not sacrosanct; hence, it might reasonably be modified by educators who wish an educational evaluation to be more blatantly circumspect.

As indicated, we can also determine educational interventions' quality by employing evaluative criteria other than students' scores on achievement tests. We will assume, for purposes of this consideration, that the evaluation to be undertaken is focused chiefly on students' mastering a dominantly cognitive curricular aim that might be sought in a typical U.S. high school's American Government course. A state-required curricular outcome calls for students to understand how a bill becomes a law. Let's assume that the intervention to be evaluated is the entire course itself—because state curricular authorities recently revised the mandatory components of such a course. What evaluative criteria might be considered for inclusion if the effectiveness of such an American Government course were to be appraised using a multimetric evaluation approach?

In such an instance, we should determine what kinds of achievement tests were routinely given to the students who typically enroll in such a class. Let's assume that the state's education department has recently developed, in collaboration with a six-state assessment consortium, both a selected-response exam and

a constructed-response exam. Each exam is administered near the close of the school year and, although a substantial portion of the two exams is quite similar, some meaningful differences exist between the content treated in the pair of exams.

In addition, social studies teachers in one of the state's largest school districts have generated and field-tested an "Attitudes toward Government" self-report inventory that students complete anonymously. High scores on this affective inventory indicate that students possess a positive attitude toward the process by which state and federal laws come into existence. This inventory could be designated as Affective Inventory A. A second affective inventory, but one focused on students' confidence in being able to explain key social-studies concepts to other people, is also available (having been purchased from a commercial test-development group). We can describe this confidence inventory as Affective Inventory B and students' scores on it as "social studies confidence." Because both of these self-report inventories are to be completed anonymously, it is impossible for teachers to arrive at *student*-focused affective inferences based on either inventory, but it is definitely possible to make valid *group*-focused inferences about social studies attitudes and confidence levels for a classroom full of students.

In the examples presented here, then, we can see that a multimetric evaluation might be designed in which four evaluative criteria could play a role, namely, students' scores on: (1) the state-developed selected-response achievement exam, (2) the state-developed constructed-response achievement exam, (3) the district-developed Attitudes toward Government Inventory, and (4) the district-developed "social studies confidence" inventory.

Assuming that educators decide to proceed with a multimetric evaluation incorporating these four evaluative criteria, care must be taken to collect students' performances on these measures in a way that allows warranted inferences to be made regarding the intervention being evaluated (in this instance, the newly structured

American Government course). To illustrate, we might employ the use of pre-instruction versus post-instruction assessments, reliance on comparisons with untreated control students, or similar types of data-gathering design tactics.

Nor must the evaluative criteria employed in multimetric evaluations consist exclusively of students' responses to tests or to affective inventories. What's most important when considering a possible evaluative criterion for multimetric evaluation is a conviction that the evidence collected is, in fact, indicative of the success of the intervention being evaluated. So, for example, when evaluating many traditional educational endeavors, we might explore the suitability of such evaluative criteria as students' (1) subsequent volitional enrollment in similar-content courses, (2) course grades, (3) attendance, (4) tardiness, or (5) extracurricular activity patterns

### **Evaluation Context**

Multimetric educational evaluation is not restricted to only one genre of evaluation, such as either *summative* or *formative* evaluation. The distinction between summative and formative evaluation was first drawn by Scriven (1967) shortly after the passage of the 1965 Elementary and Secondary Education Act, in which state and local recipients of considerable federal dollars were obliged to evaluate "this year's" federally supported educational programs in order to receive "next year's" federal support. In an effort to clarify the nature of what had thus become statutorily prescribed evaluation, Scriven characterized "formative evaluation" as an attempt to ascertain the merit of a still-malleable educational program so that this under-evaluation program could be improved. On the other hand, "summative evaluation" called for evaluating a mature, already finalized educational program in order to arrive at a final continue-or-terminate decision.

Although many of the recent calls for using multiple criteria to evaluate educational interventions have arisen in summative settings—because summative evaluations tend to attract more public attention than their formative counterparts do—multimetric educational evaluation is *not* limited to use in summative contexts, which are often regarded as “accountability-focused” evaluations. When they are employed for accountability purposes, multimetric evaluations typically lead to more defensible decisions than single-criterion evaluations do, and the virtues of multimetric evaluations are equally present in formative-evaluation contexts. A teacher who is deciding whether to make any adjustments in an ongoing unit of instruction based on evidence of her students’ current learning will be able to make those decisions better if they’re based on multiple sources of evidence, rather than only one.

Realistically, given the heightened significance of decisions that result from many summative evaluations, additional funds are often needed to expand the number of evaluative criteria for multimetric evaluations. For example, if a school’s staff decides to use several assessment devices rather than only one, there’s a need to acquire, administer, score, and make sense of those assessments. Even on a smaller scale there can be informal, though not unsubstantial, costs. Consider a teacher who wants other teachers to evaluate—in person—the quality of her students’ oral presentations. There’s the hassle of recruiting those colleagues, orienting them and, perhaps, supplying their morning coffee, etc. Clearly, employing a multiple criteria evaluative approach usually is more involved and expensive than opting for a solo criterion.

In the case of formative evaluations, even though the merits of a multimetric evaluation strategy are demonstrably applicable, it may be that the implementation costs cannot be justified. Conceptually, however, the decision-influencing dividends of a multimetric strategy are equally applicable to both summative and formative educational evaluations.



To sum up, the single overriding dividend of a multimetric approach to educational evaluation is that its reliance on more than one evaluative criterion when appraising an educational intervention increases the odds, often dramatically, that a more defensible judgment of quality will be reached. Clearly, no guarantees come with this evaluative strategy; real-world mistakes will be made in identifying evaluative criteria or in determining their significance. Moreover, greater evaluative costs are almost always involved when a multimetric evaluation strategy is adopted. The increased defensibility of a multimetric evaluation, therefore, must always be contrasted with the increased costs of such an evaluation. Nonetheless, whenever it is affordable, *multimetric educational evaluation improves the likelihood of a defensible evaluation-based decision.*

## How Does Multimetric Educational Evaluation Work?

Having considered the reasons underlying the growing advocacy for multimetric educational evaluation, we now turn to a consideration of how such an evaluative approach might function. Let us assume that an evaluator has accepted the premise that reliance on more than a single evaluative criterion is likely to improve an evaluation study's conclusions. How, then, might this fictitious evaluator proceed?

### Disavowal Time

The analysis was undertaken with the clear understanding that the approach to multimetric educational evaluation being described in the following pages is an *illustration, not a to-be-followed template.* In short, this analysis highlights *one way* in which multimetric educational evaluation might be implemented. Other implementation

procedures—procedures that still benefit from the advantages of relying on diverse evaluative criteria—should definitely be considered.

In addition, because multimetric educational evaluation is, at its core, dependent on identifying and using situation-specific evaluative criteria, it would make little sense for any national entities to describe an implementation strategy that dictated the use of predetermined evaluative criteria. Indeed, a fundamental proposition of the multimetric procedure to be described in the remainder of this analysis is that the really significant choices to be made—namely, the choice of the evaluative criteria to be used and the weighting of their significance—are definitively a local-option enterprise.

## A Six-Step Evaluation Framework

What now follows is a multistep framework for implementing a multimetric evaluation strategy—an implementation framework applicable to both summative and formative evaluation. One of the desirable attributes of frameworks is that, in most instances, they are considerations, not constraints. As you review the suggested six-step process, please recognize that you can select those steps you regard as sound, yet reject or modify other suggested steps. And, of course, even if you were to accept verbatim what is about to be suggested, the key decisions about which evaluative criteria to select and how much to weight their import should always be done locally. The six-step framework is just that, no more and no less—a *framework*.

The intensity with which the six-step framework should be implemented is, as already noted, definitely a local decision. The higher the stakes associated with a particular evaluation, the more fully the following framework should be employed. In less portentous situations, more relaxed use of the framework will usually be sufficient. If, for example, a teacher wants to evaluate the success of a new mathematics course that she has never previously

taught and wants to rely on at least a loose implementation of multimetric evaluation, the teacher might design a more suitable evaluation simply by thinking through the steps, rather than actually implementing each step. On the other hand, if the decision at issue is an inordinately important one, then all of the framework's steps should be implemented with far more specificity than what will be described in the following analysis.

We will now explore each of the implementation framework's six steps—along with a continuing illustrative example. For an overview of what will be considered, it might be useful to first examine the graphic example provided in Figure 1. This figure supplies a schematic illustration of the entire six-step process.

## STEP 1

### **Identify Potential Purpose-Consonant Evaluative Criteria.**

A multimetric evaluation strategy starts by isolating what is going to be evaluated—usually a rather straightforward endeavor. However, the messy part of Step 1 is to identify the possible evaluative criteria that might be employed in an upcoming multimetric evaluation, and to do so in accord with the *purpose* of the educational intervention being evaluated. This can be a substantially more difficult task.

Earlier in this analysis, several examples of potentially evaluable educational interventions were presented. These interventions ranged from what a teacher might undertake in a classroom (such as discovery learning) or an entire school year's worth of instruction (such as the academic-year progress of students in a particular school). To help evaluators arrive at an uncontaminated determination of what's to be evaluated, it is particularly useful to identify the *decisions* that directly depend on the upcoming evaluation. There are few better ways to help get an accurate fix on what

---

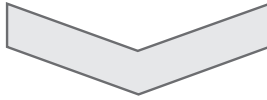
Figure 1. Overview: Six-Step Evaluation Framework

---



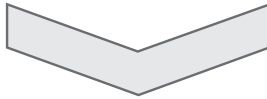
STEP ONE

**DETERMINE**  
*the Evaluative Criteria*



STEP TWO

**WEIGHT** *the Importance  
of the Evaluative Criteria*



STEP THREE

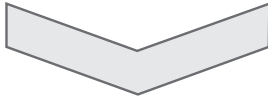
**IDENTIFY** *Evidence for  
Each Evaluative Criterion*





**STEP FOUR**

**RATE** *the Adequacy  
of the Evidence*



**STEP FIVE**

**ADJUST AND FINALIZE**  
*the Evaluative Criteria's  
Importance Weights*



**STEP SIX**

**COMMUNICATE**  
*the Evaluation Process  
to Decisionmakers  
& Stakeholders*

is to be evaluated than isolating the specific decision or decisions that must be made after stakeholders have considered the evaluation's results. Thus, the evaluator should immediately explicate the decision at issue—formally and in writing—because it is to make this decision more defensibly that constitutes the reason for doing any evaluation at all.

With this decision-focused purpose clearly in mind, then the major task of Step 1 is to identify potential evaluative criteria that might be used to judge the quality of the educational intervention being evaluated. At this Step 1 stage, however, we need a *general description* of each evaluative criterion—not an isolation of the specific evidence that will be used to operationalize, or measure, each evaluative criterion. (The isolation of the specific evidence to be used is part of Step 3.)

Although we are looking for general terms in this step, the evaluative criteria should not be so utopian that there is little chance, in one form or another, evidence reflecting them can be realistically collected. What is typically done at this point in a multimetric evaluation is to consider expansively, similar to a brainstorming approach, the indicators that might sensibly be used to determine the quality of whatever intervention is being evaluated.

Suppose the educational intervention to be evaluated is a fairly straightforward, semester-long instructional approach in which students are required every 20 to 30 minutes to *mentally* summarize (in two minutes or less) what they have just learned in mathematics, then relay this mental summary to a teacher-designated student partner. The two students then exchange roles as summarizers. The math teacher characterizes this instructional tactic as S&S (Summarize and Share). She and two of her math-teacher colleagues are trying to use a multimetric evaluation model when evaluating this S&S approach. After considerable deliberations about potential evaluation criteria to employ in the evaluation, they ended up with the criteria presented in Figure 2.

---

**Figure 2. Evaluative Criteria for Summarize and Share**

---

1. Improved Student Mathematics Achievement
2. Positive Attitude toward the Topics Studied
3. Increased Confidence in My Summarization Skill
4. Expanded Number of Close Classroom Friends

The challenge in carrying out this initial step is *not* to merely come up with a lengthy list of possible evaluative criteria. On the contrary, the evaluator's mission is to try to isolate those criteria that represent accurate reflections of the evaluation's success. Thus, it would be better to conclude Step 1 by identifying only the truly indicative representations of an educational intervention's effectiveness, rather than identifying a dozen such indicators, most of which may be only peripherally indicative of an intervention's success.

Step 1, therefore, typically consists of open deliberations among those guiding a multimetric evaluation regarding which evaluative factors are worthy of inclusion. It is particularly important during such deliberations to make sure those involved possess identical understandings of the evaluative criteria being discussed. Plenty of questions and proffered examples are typically required during such discussions in an effort to get all discussants to view the criteria in the same way.

Much of the Step 1 deliberations, of course, will deal directly with the reason(s) that a particular evaluative criterion should be incorporated in the multimetric evaluation being designed. The underlying question about each contending evaluative criterion

runs something like this: “Why should this evaluative criterion be included when determining the success of the educational intervention to be evaluated?” In the pro/con discussions of each evaluative criterion under consideration, multimetric evaluators typically gain a sufficiently clear idea about whether they have isolated a satisfactory array of factors to be used when determining the to-be-evaluated intervention’s success.

Once we have identified a tentative collection of suitable evaluative criteria, we can turn to Step 2. But first, please take note of the adjective “tentative” in the preceding sentence. Our six-step evaluation framework is just that, a framework, and there is nothing sacrosanct and unchangeable about earlier steps in the process. Often, as evaluators get more deeply into the specifics of a particular evaluation plan, they will realize that earlier decisions need revision, and this is perfectly acceptable. Evaluation frameworks exist to help guide evaluators, not hamstringing them.

## **STEP 2**

### **Determine the Importance of the Chosen Evaluative Criteria.**

Remembering that the result of Step 1’s actions are a set of generally labeled evaluative criteria, the Step 2 task is to rate their importance in determining the success of the intervention being appraised. This step, as is true of all six steps in this framework, can be carried out in collaboration with your colleagues or on your own. The question to consider is simple: How much should each evaluative criterion count, if at all, in determining the success of whatever is being evaluated?

One of the most understandable ways to complete Step 2 is to present those involved (if more than one person is taking part) with a list of Step 1’s evaluative criteria, then ask each evaluator to



allot a total of 100 points among all of the criteria identified. Any evaluative criterion can be weighted from zero to 100 points, so long as the total points across the full slate of criteria equals exactly 100. Obviously, the criteria deemed more important will be given greater numerical values relative to the other criteria that are included. If a Step 1 evaluative criterion is assigned an importance percentage of zero in Step 2, of course, then it should be deleted from the set of applicable evaluative criteria. If one of the evaluative criteria receives a 100-point importance rating, then the evaluators must take two actions: first, they must remove all of the previously identified criteria from consideration; and second, the evaluators must determine whether there are two or more forms of evidence for the 100-point criterion that can be considered and employed in Step 3. If not, then this is fine—but they are no longer engaged in a multimetric educational evaluation as defined in this analysis because, as discussed earlier, the “two or more forms of evaluative evidence” requirement has not been satisfied.

So, perhaps preceded by some discussion of the significance of each of the contending evaluative criteria, those involved in the process are asked to distribute 100 points of evaluative importance among the potential evaluative criteria. Results of such an importance rating can be seen in Figure 3 for our illustrative evaluation of the previously described math teacher’s S&S instructional tactic.

In these importance ratings, we see that the evaluators involved have assigned the bulk of their ratings to the two evaluative criteria representing more traditional academic achievement (70 points going to improved academic achievement) and students’ confidence in their ability to summarize (15 points). It is important to make sure that all those involved in providing Step 2 ratings distribute their entire 100-point allocations among the evaluative criteria—to assure the equalization in significance of each raters’ judgments. And, as will often be the case during the conduct of a multimetric evaluation, those involved in planning and

---

**Figure 3. Importance Weights for Summarize and Share**

---

1. Improved Student Mathematics Achievement  
**(Importance: 70 pts.)**
2. Positive Attitude toward the Topics Studied  
**(Importance: 10 pts.)**
3. Increased Confidence in My Summarization Skill  
**(Importance: 15 pts.)**
4. Expanded Number of Close Classroom Friends  
**(Importance: 5 pts.)**

implementing the evaluation should be reminded of the evaluation's purpose in relation to the decision at issue.

Yet, recalling that the Step 1 evaluative criteria were only identified in a general manner, and that each of those criteria must be operationalized by using one or more evidence-gathering procedures, it is now time to make sure that—given the actual evidence-collection techniques to be used—these Step 2 importance ratings hold up. In the next stage in the framework, we will designate evidence-elicitation procedures.

### **STEP 3**

#### **Select Evidence-Elicitation Procedures for Each Evaluative Criterion.**

In this step of a multimetric evaluation, we must isolate the ways in which evidence is to be collected regarding each of the remaining evaluative criteria. In other words, now we need to tie down the

specific assessment devices or other data-collection techniques that will be used to represent each of the evaluative criteria around which the multimetric evaluation has been organized. And this is when multimetric evaluators need to be inventive rather than passive by meekly accepting whatever tests happen to be at hand.

Only when we know how the evidence representing a given evaluative criterion will be collected do we truly understand what that criterion contributes to the evaluation. This issue often obliges multimetric evaluators to be particularly inventive in coming up with defensible evidence to support evaluative inferences about students' status regarding certain evaluative criteria.

Students' learning is the most common and often the most important of the evaluative criteria involved in multimetric evaluation. Such learning can sometimes be measured by two or more substantially different types of educational tests.

Ideally, the tests being used to show evidence of changes in students' learning will have been previously shown capable of doing so accurately. To illustrate, the tests to be used will be accompanied by evidence indicating each test is *instructionally sensitive*; that is, capable of distinguishing between well taught and poorly taught students.

Thus, for each evaluative criterion still on the table, we must identify one or more evidence-collecting procedure as the method by which each criterion's evidence is to be gathered. Considering everything we know about educational assessment, we want to employ the most cost-effective ways of garnering evidence that will allow us to invoke a particular evaluative criterion when judging the worth of an educational intervention. A good many textbooks dealing with the fundamentals of educational assessment provide a number of sound principles for creating first-rate educational tests (see several such assessment textbooks listed in the references for this analysis), and a mid-2014 publication of the influential *Standards for Educational and Psychological Testing* (AERA, 2014)

provides an excellent set of profession-approved guidelines for constructing and appraising such tests.

Often, the only two options available to those who wish to complete Step 3 in this approach will be to use an existing educational test or to build the needed assessment afresh. And this is why, for Step 3, those who are familiar with available assessment instruments will have an advantage. Beyond that knowledge, it will often be necessary for multimetric evaluators to actually construct an assessment instrument that meshes satisfactorily with an evaluative criterion being used in a given evaluation.

Continuing with our illustrative example of evaluating a teacher's S&S instructional intervention, please consider the Figure 4 presentation of proposed evidence-collecting procedures for each of the four evaluative criteria being used. Among those

---

**Figure 4. Evidence Collection Procedures for Summarize and Share**

---

1. Improved Mathematics Student Achievement
  - District-Developed Basic Math Test
  - Commercial Interim Math Test
2. Positive Attitude toward the Topics Studied
  - Existing Teacher-Developed Self-Report Inventory
3. Increased Confidence in My Summarization Skill
  - Inventory Recently Built by District Math Teachers
4. Expanded Number of Close Classroom Friends
  - To-Be-Built Self-Report Inventory

evidence options, we see several already available and at least one assessment (the to-be-developed self-report inventory) that must be constructed. This illustrative collection of diverse kinds of evidence is fairly typical of what takes place in a multimetric evaluation: evaluators consider evidence-collection options already at hand, then see whether any to-be-built assessments are needed.

This concludes Step 3—the choice of which evidence-elicitation procedures to use as an indicator of each already importance-rated evaluative criterion. The next step in our process calls for us to weight the *adequacy* of each of these evidence-elicitation procedures.

#### STEP 4

### **Rate the Adequacy of Each Evidence-Elicitation Procedure.**

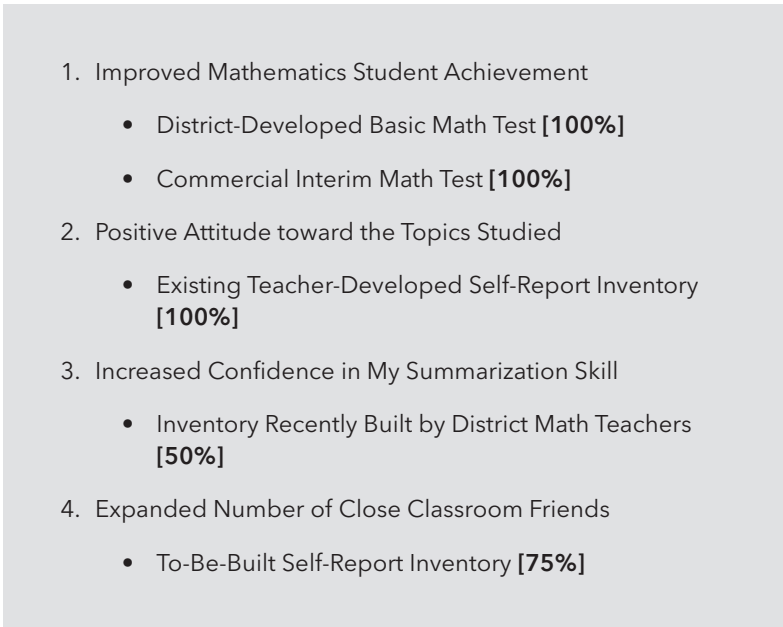
Not all evidence is equally persuasive. In Step 4, our task is to rate the appropriateness of the evidence we intend to consider when arriving at our final evaluative judgment. This is an important phase in the evaluation process because, without it, powerful evaluative evidence might be undervalued because of an abundance of other, less compelling evidence. Step 4's attention to the quality of evidence being employed reduces the likelihood of such a procedural error.

Please consider the illustrative adequacy ratings presented in Figure 5 to see how the four evidence-elicitation procedures chosen in our illustrative multimetric evaluation have been assigned weights. As you will see, the first three evidence-elicitation procedures have been assigned 100-percent adequacy ratings. This signifies that whatever importance weightings had been assigned to these evaluative criteria in Step 2 should be accepted without any adjustments. As you can see in Figure 5, the first two sources of evidence are a district-developed basic math test and a commercial interim math test. Ample technical information regarding the

---

**Figure 5. Adequacy Ratings for Summarize and Share**

---



quality of such assessments is typically available; in this example, we'll assume that such information has persuaded the evaluators that evidence from both tests should receive full credit in our multimetric evaluation. Similarly, the existing teacher-developed self-report attitude inventory has been used for several years with no complaints and, therefore, also receives a 100-percent adequacy weight.

The fourth evidence-elicitation procedure, a recently developed inventory built by district mathematics teachers, gets only a 50 percent adequacy rating in this example. We could explain this new instrument's lower rating by noting that it probably has only an abbreviated usage history, if any, and may have been negatively reviewed by the evaluators.

The fifth and final evidence option, a to-be-built inventory, receives a 75 percent adequacy weight even though it has not yet been constructed. Such a positive, yet less than 100 percent weight, might have been assigned because a committee of highly able teachers has agreed to develop the new inventory. Moreover, because the particular evaluative criterion to which this evidence is linked received an importance rating of only 5 points in Step 2 (see Figure 3), it will have only a minor influence on the overall evaluation.

One interesting situation arises when we see that two different tests have been selected to supply evidence bearing on the same evaluative criterion--here, improved mathematics student achievement. Recalling (from Figure 3) that this evaluative criterion was assigned a 70-point importance rating, the evaluators will then need to decide how the 70-point importance rating will be divided between students' scores on the two tests. Assuming that both tests are equally good, it would be a simple matter to split the evaluative contributions of scores from each test to 35 points each. If the evaluators feel that one test is far more indicative of the intervention's success, then it may be weighted more heavily than the other test.

These adequacy ratings for the evidence-elicitation procedures must now be combined with the previously designated importance weights for each evaluative criterion. We will do this in Step 5.

## **STEP 5**

### **Adjust Evaluative Criteria Importance Weights by Evidence Adequacy.**

This step calls for us to make any necessary adjustments in the importance weightings of each evaluative criterion according to our judgments regarding the adequacy of the evidence-collection methods we are using to operationalize each such criterion. In

certain instances, of course, when all evidence-options receive a 100 percent adequacy rating, then no adjustments are required. However, in the illustration being used here (see Figure 6), two adjustments will be necessary.

Because evaluators regarded the final two evidence-elicitation procedures with some concern, we adjusted the original importance weights for those two evaluative criteria by multiplying the original importance weights by the adequacy percentage. To illustrate, we multiplied the original importance weight of 15 points for “Increased Confidence in One’s Summarization Skill” by the adequacy rating of 50 percent to yield an adjusted importance weight of 7.5 points.

---

**Figure 6. Adjusted Importance Weights for Summarize and Share**

---

1. Improved Mathematics Student Achievement
  - The District-Developed Basic Math Test **[35 points]**
  - A Commercial Interim Math Test **[35 points]**
2. Positive Attitude toward the Topics Studied
  - Existing Teacher-Developed Self-Report Inventory **[10 points]**
3. Increased Confidence in My Summarization Skill
  - Inventory Recently Built by District Math Teachers **[7.5 points]**
4. Expanded Number of Close Classroom Friends
  - A To-Be-Built Self-Report Inventory **[3.75 points]**



Because the adequacy ratings assigned to different evidence options can dramatically influence the final effect of key evaluative criteria in a multimetric approach, arriving at defensible adjustments regarding the adequacy of each evidence-collection procedure is crucial. Looking back, of course, it can be useful to remind ourselves that the original weighting of different evaluative criteria's importance and the subsequent rating of how well various evidence-gathering procedures measured those criteria are made *judgmentally*. Such judgments are made either by the evaluators themselves or by groups of stakeholders assembled for this purpose. To maximize an evaluation's transparency, the evaluation staff should document the most prominent factors influencing pivotal along-the-way decisions, such as those about the quality of an evaluation study's evidence-collection procedures.

Now, however, with evidence to inform us regarding the array of evaluative criteria being used, and any requisite adjustments having been made in the importance of the evaluative criteria, we are ready to describe the wrap-up step in this illustrative application of multimetric educational evaluation.

## STEP 6

### **Provide a Synthesized Evaluation Report to Decision Makers.**

As a multimetric educational evaluation winds down, what's left for the evaluators to do is to supply all relevant decision makers with an evaluation report that synthesizes the entire evaluation process so that decision makers can easily see where pivotal evaluative decisions have been made, and on what basis. This would be a wonderful time for evaluators to supply decision makers with brief rationales for the weighting of the study's evaluative criteria and the ratings of a study's evidence-collection methods. Those deci-

sion makers can then determine the degree to which their ultimate decisions will be informed by results of the evaluation.

Clearly, if evaluators used a six-step approach to multimetric evaluation akin to what's been described here, then they should concisely explicate the nature and consequences of each of those steps. If evaluators used a different multimetric approach, then they would need to describe its nature for the readers of an evaluation report. Given the inherent complexity of most multimetric evaluations, for it sometimes requires an evaluator to juggle several balls at the same time, the final report of a multimetric evaluation should not be so elaborate as to render the entire multimetric approach off-putting. Rarely, perhaps never, will an educator's report of a multimetric evaluation constitute a suitable submission for a doctoral dissertation or a potential Nobel Prize. Accordingly, keep the report clear, yet concise.

It is often helpful to ask a colleague who knows naught about what's being evaluated to review a draft evaluation report. Frequently, those carrying out an educational evaluation—be it multimetric or unimetric—will become so familiar with their evaluative machinations that they inaccurately assume others will automatically comprehend what's going on. Reactions from an uninvolved colleague can often identify some key points in a draft report that require editorial amelioration.

Although it is invariably the case that an evaluation report will be prepared chiefly for those individuals who will be making a decision—summative or formative—based on the report, sometimes multiple audiences should be kept in mind by those who prepare the evaluation report. For instance, in certain cases it makes sense to prepare a decision-focused report for the actual decision-makers as well as a more abbreviated report for the public at large. As with all writing, to isolate with certainty one's audience can markedly improve the likelihood that a first-rate report will be created. Remember, the educational intervention being eval-

uated via a multimetric evaluation will have an explicit purpose, and the function of the evaluation is to provide useful insights to those individuals who must render a decision about this intervention. In this final Step 6 of the evaluation, those in charge of the evaluation must make certain that the evaluation's activities have been described with sufficient lucidity. In general, of course, multimetric evaluations will be more complicated to describe than a single-criterion evaluation study. The skilled multimetric evaluator must bring simplified clarity to the report of such evaluations.

## A Brief Look Back

A virtue of multimetric evaluation is that it obliges decision makers to attend to more than one evaluative criterion when determining the worth of an educational intervention. A broadened consideration of relevant factors almost always leads to more appropriate evaluation-based decisions. Of course, poorly implemented multimetric evaluations will usually result in worse decisions than those that might have ensued from a properly implemented sole-criterion evaluation. However, the basic procedural operations in a multimetric educational evaluation are really quite straightforward.

The six-step procedure described herein, or one like it, is necessary for successful implementation of a multimetric evaluation system. Once a decision is made to rely on multiple factors to evaluate anything (in our case, an educational intervention), then we must identify those evaluative factors. Then, we must determine the importance of each evaluative criterion, and after that, choose suitable sources of evidence to represent each criterion. At that point, we must make any necessary adjustments in the importance of an evaluative criterion to mitigate shortcomings in the evidence-elicitation procedures to be used. Finally, we must provide a synthesized report of the evaluation to the decision makers involved.

Although the use of multimetric evaluations have been, and should be, applauded as a more defensible way to appraise interventions than relying on solo-criterion evaluation, a multimetric approach does not eliminate the need for human judgment. Mistakes will be made when judging the quality of an intervention even if using a multimetric evaluative strategy. Yet, such judgmental slipups will be less likely when we attend to several evaluative criteria rather than only one.

Better evaluation-based decisions are apt to follow from appropriately implemented multimetric evaluations, and thus students are apt to be more effectively educated because of those decisions. As such, a multimetric evaluation approach is the best option for promoting success in education, and it is imperative that today's educators and policymakers implement it with thoughtfulness, clarity, and care.

# Resources

American Educational Research Association. (2014). *Standards for educational and psychological testing*. Washington, DC: Author.

Nitko, A. J., and Brookhart, S. M. (2014). *Educational assessment of students* (7th ed.). Upper Saddle River, NJ: Prentice-Hall/Merrill Education.

Popham, W. J. (In press). *Classroom assessment: What teachers need to know*. Boston: Pearson.

Scriven, M. (1967). The methodology of evaluation. In R. W. Tyler, R. M. Gagne, and M. Scriven (Eds.), *Perspectives of curriculum evaluation*, Volume I (pp. 39–83). Chicago: Rand McNally.

Stiggins, R., and Chappuis, J. (2012) *An introduction to student-involved assessment for learning* (6th ed.) Boston: Pearson.

# About the Author



**W. James Popham** is Emeritus Professor in the UCLA Graduate School of Education and Information Studies. He has spent most of career as a teacher, largely at UCLA, where for nearly 30 years he taught courses in instructional methods for prospective teachers and graduate-level courses in evaluation and measurement. At UCLA he won several distinguished teaching awards, and in January 2000, he was recognized by *UCLA Today* as one of UCLA's top 20 professors of the 20th century.

In 1968, Popham established IOX Assessment Associates, a research and development group that created statewide student achievement tests for a dozen states. In 2002, the National Council on Measurement in Education presented him with its Award for Career Contributions to Educational Measurement. He is a former president of the American Educational Research Association (AERA) and the founding editor of *Educational Evaluation and Policy Analysis*, an AERA quarterly journal. In 2006, he was awarded a Certificate of Recognition by the National Association of Test Directors. In October 2009, he was appointed by Secretary of Education Arne Duncan to the National Assessment Governing Board, the policy-setting group for the National Assessment of Educational Progress.

Popham is the author of more than 30 books, 200 journal articles, 50 research reports, and nearly 200 papers presented before research societies. His most recent books are *Classroom Assessment: What Teachers Need to Know*, 7th ed. (2014), *Mastering*

*Assessment* (2011), and *Assessment for Educational Leaders* (2006), Pearson; *Evaluating America's Teachers: Mission Possible?* (2013), *Everything School Leaders Need to Know about Assessment* (2010), and *The ABCs of Educational Testing: Demystifying the Tools That Shape Our Schools*, *in press*, Corwin; *The Truth About Testing* (2001), *Test Better, Teach Better* (2003), *Transformative Assessment* (2008), *Instruction that Measures Up* (2009), and *Transformative Assessment in Action* (2011), ASCD; *America's "Failing" Schools* (2005), Routledge; *Unlearned Lessons* (2009), Harvard Education Press.

