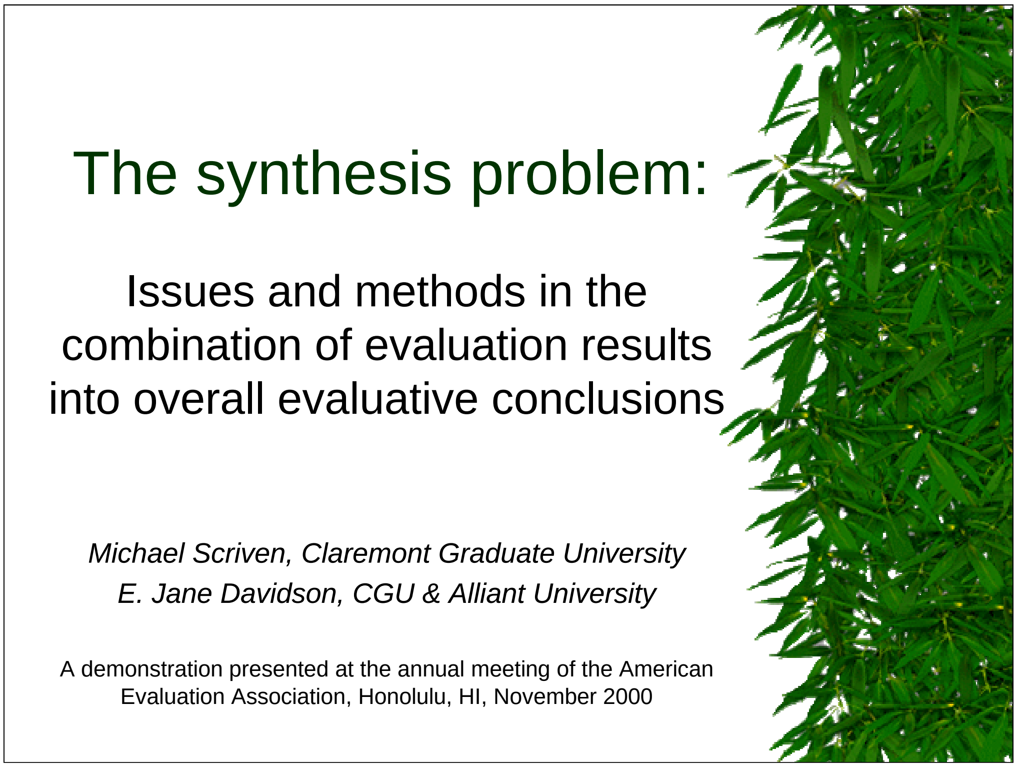
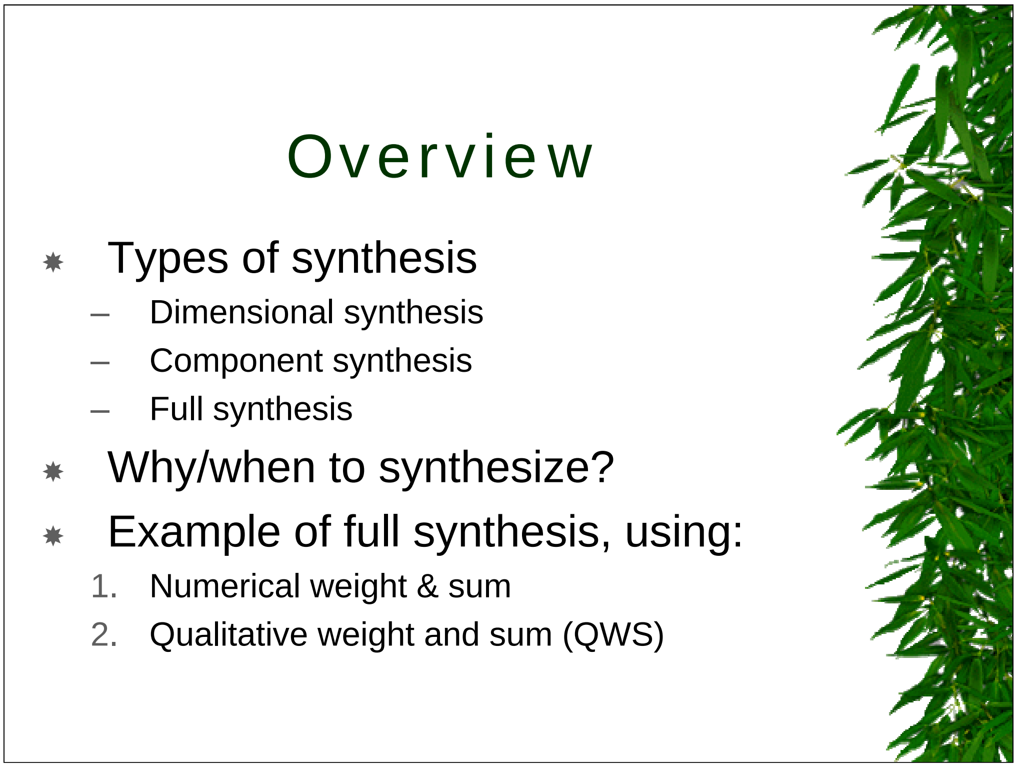


The synthesis problem:

Issues and methods in the combination of evaluation results into overall evaluative conclusions

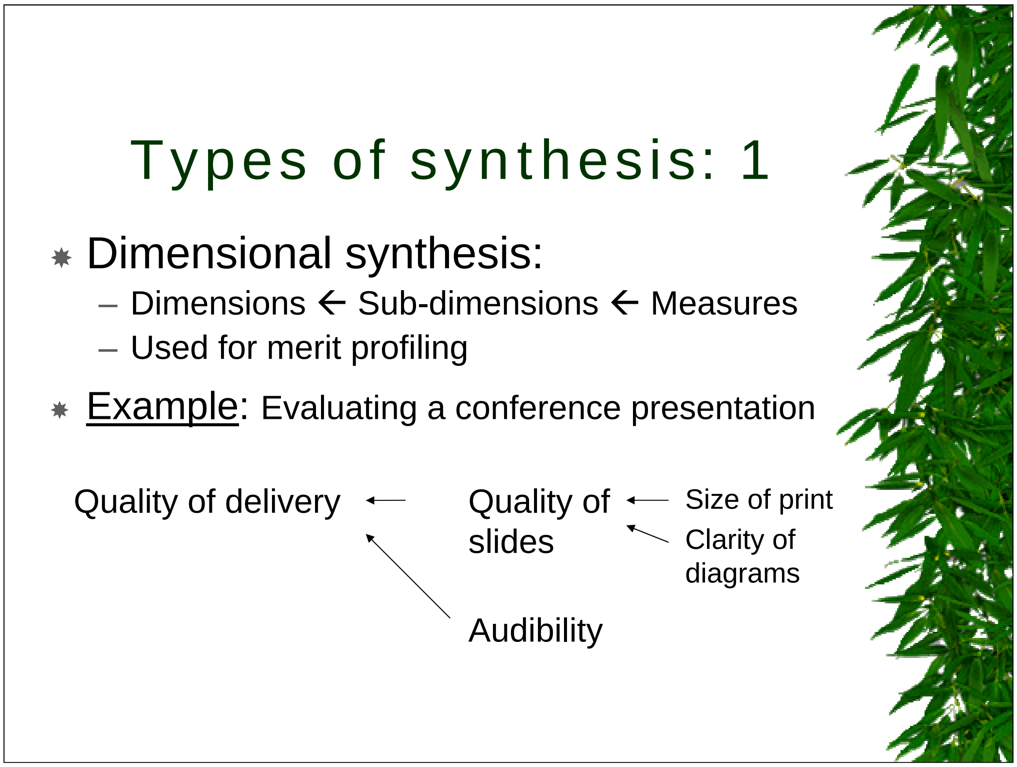
Michael Scriven, Claremont Graduate University
E. Jane Davidson, CGU & Alliant University

A demonstration presented at the annual meeting of the American Evaluation Association, Honolulu, HI, November 2000



Overview

- Types of synthesis
  - Dimensional synthesis
  - Component synthesis
  - Full synthesis
- Why/when to synthesize?
- Example of full synthesis, using:
  1. Numerical weight & sum
  2. Qualitative weight and sum (QWS)



Types of synthesis: 1

- Dimensional synthesis:
  - Dimensions ← Sub-dimensions ← Measures
  - Used for merit profiling
- Example: Evaluating a conference presentation

Quality of delivery ← Quality of slides ← Size of print

Quality of slides ← Clarity of diagrams

Quality of delivery ← Audibility

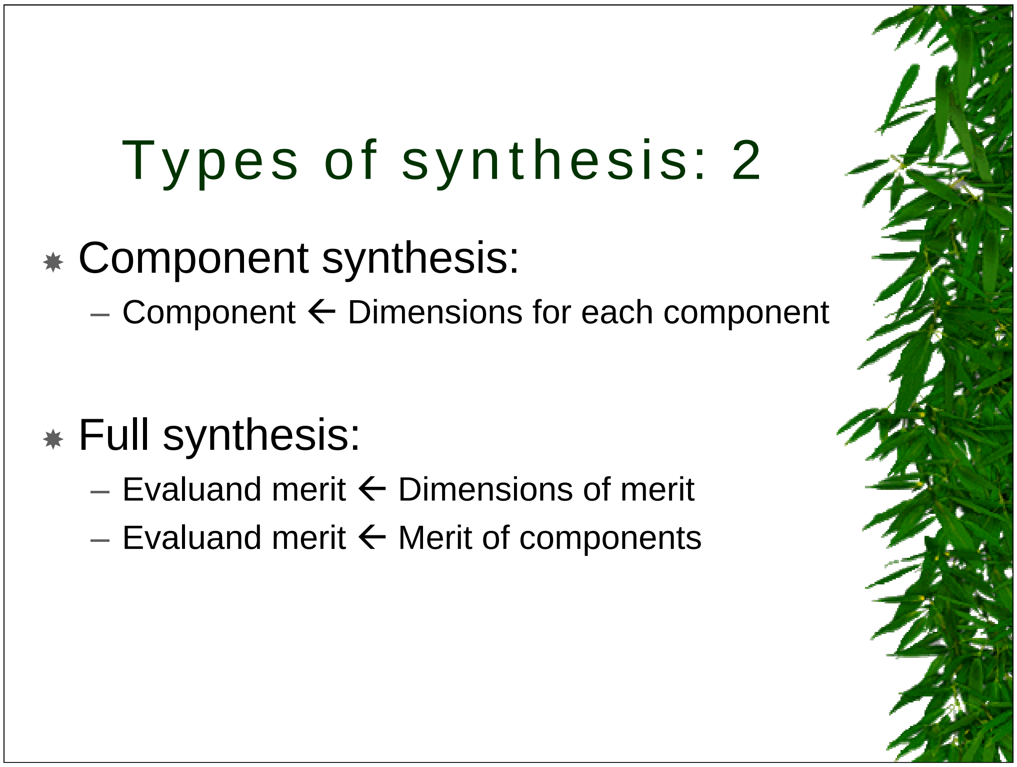
For virtually every single evaluation, at least some form of synthesis is necessary.

The simplest case is synthesis by dimensions. An evaluand (be it a program, a product, or something/someone else) is typically evaluated on a number of ‘dimensions’ (criteria of merit). What makes a good (i.e., well-functioning) evaluand of type X? The answer to this question constitutes the list of dimensions.

The initial list of dimensions is usually very broad, and needs to be broken down further into sub-dimensions, which are then broken down further to specific measures or other assessments of merit.

To give an example, what makes a good conference presentation?

Dimensional synthesis involves packing performance on the various performance measures back into the sub-dimensions, and then back again into the main dimensions of merit. This yields a dimensional profile of evaluand merit.



Types of synthesis: 2

- Component synthesis:
  - Component ← Dimensions for each component
- Full synthesis:
  - Evaluand merit ← Dimensions of merit
  - Evaluand merit ← Merit of components

A component synthesis involves evaluating <u>pieces</u> or <u>sections</u> of an evaluand, rather than *aspects* of it. As an example, suppose when evaluating the AEA conference, one might evaluate the following components:

•The professional development workshops

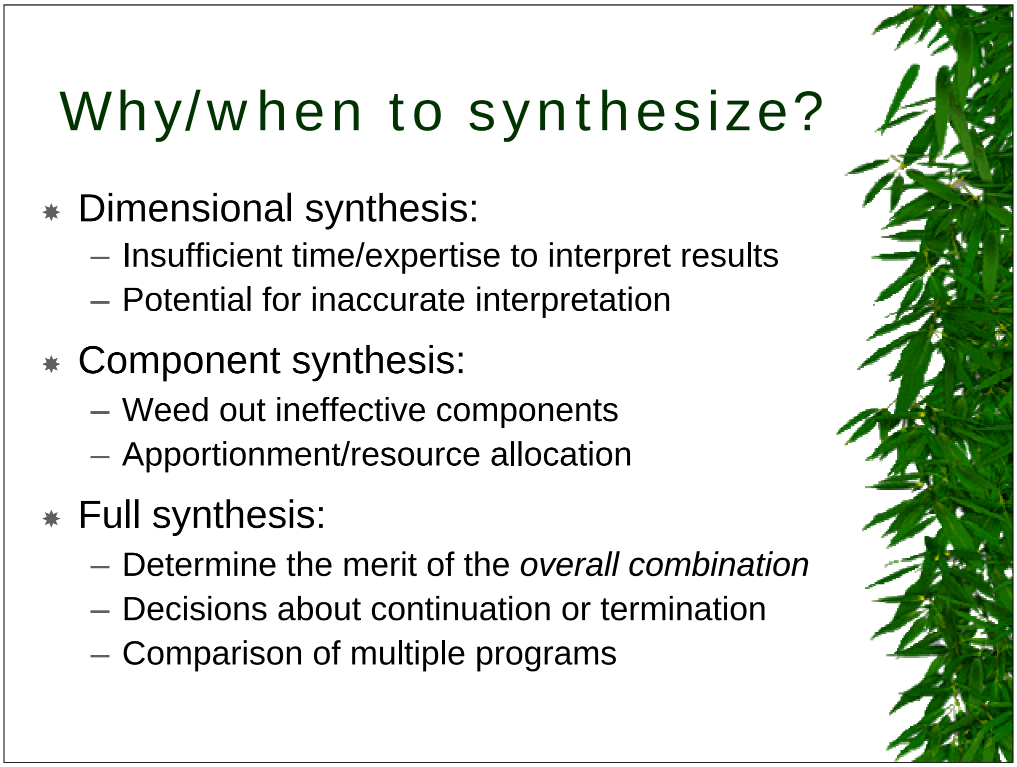
•The presidential strand

•Other sessions

•The receptions/social hours

Each component is evaluated on a number of dimensions. Component synthesis involves packing performance on those dimensions together in order to determine the merit or worth of each component.

Finally, full synthesis involves taking <u>either</u> a dimensional or a component evaluation and combining those sub-evaluations to draw an overall conclusion about the merit of the evaluand.



Why/when to synthesize?

- Dimensional synthesis:
  - Insufficient time/expertise to interpret results
  - Potential for inaccurate interpretation
- Component synthesis:
  - Weed out ineffective components
  - Apportionment/resource allocation
- Full synthesis:
  - Determine the merit of the *overall combination*
  - Decisions about continuation or termination
  - Comparison of multiple programs

Use dimensional profiling when:

(a) readers have insufficient time and/or expertise to interpret descriptive results and draw their own conclusions, *and/or*

(b) readers have significant ego-investment in the favorability of results and there is room for varied interpretation of results as either favorable or unfavorable.

Use component synthesis when:

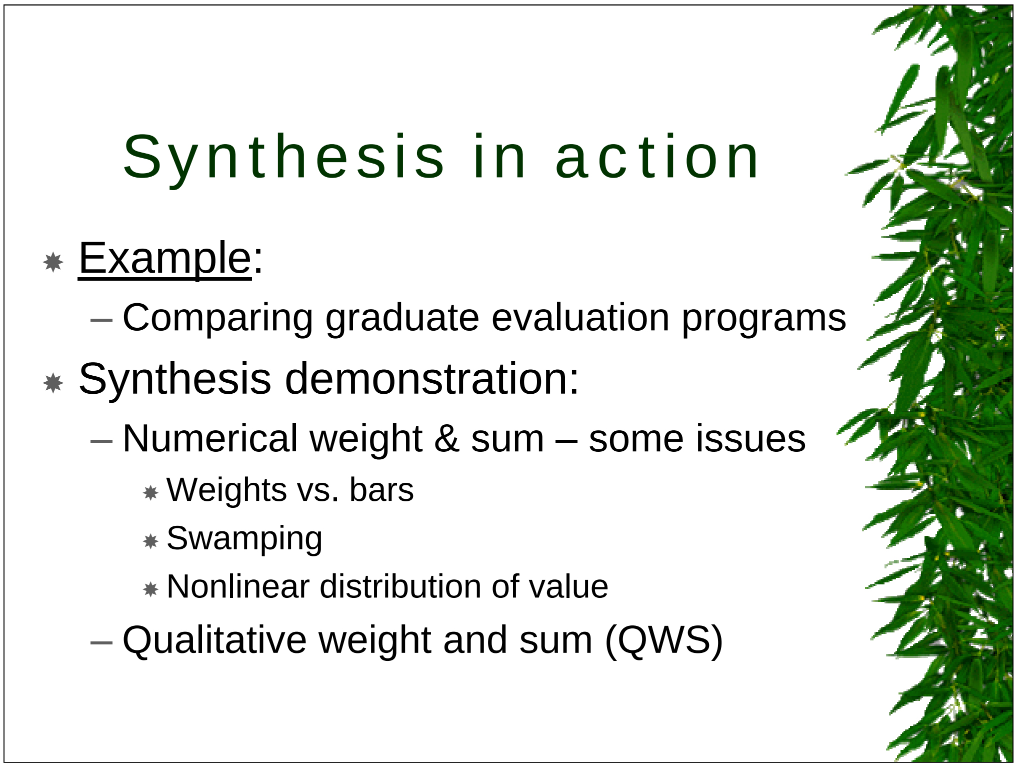
•the organization needs to decide which of a number of change initiatives to devote resources to and which to abandon altogether

Use full synthesis when the organization needs to know how its program is faring overall, relative to the other options, e.g.

•To learn how good the *overall package* (which is often greater or less than the sum of its parts) is as a combination.

•To make a decision about whether to continue the whole program or not

•In order to choose among multiple programs (e.g., several interventions being piloted, or are being looked at for possible implementation at a new site).



Synthesis in action

- Example:
  - – Comparing graduate evaluation programs
- Synthesis demonstration:
  - – Numerical weight & sum – some issues
    - Weights vs. bars
    - Swamping
    - Nonlinear distribution of value
  - – Qualitative weight and sum (QWS)

To illustrate some of the issues and methods available for full synthesis (in this case, packing *dimensions* into an overall determination of worth), we will use the example of evaluating three graduate programs from the perspective of a consumer trying to decide which one to go to.

This will be done using two methodologies

•The numerical weight and sum (NWS)

•The qualitative weight and sum (QWS)

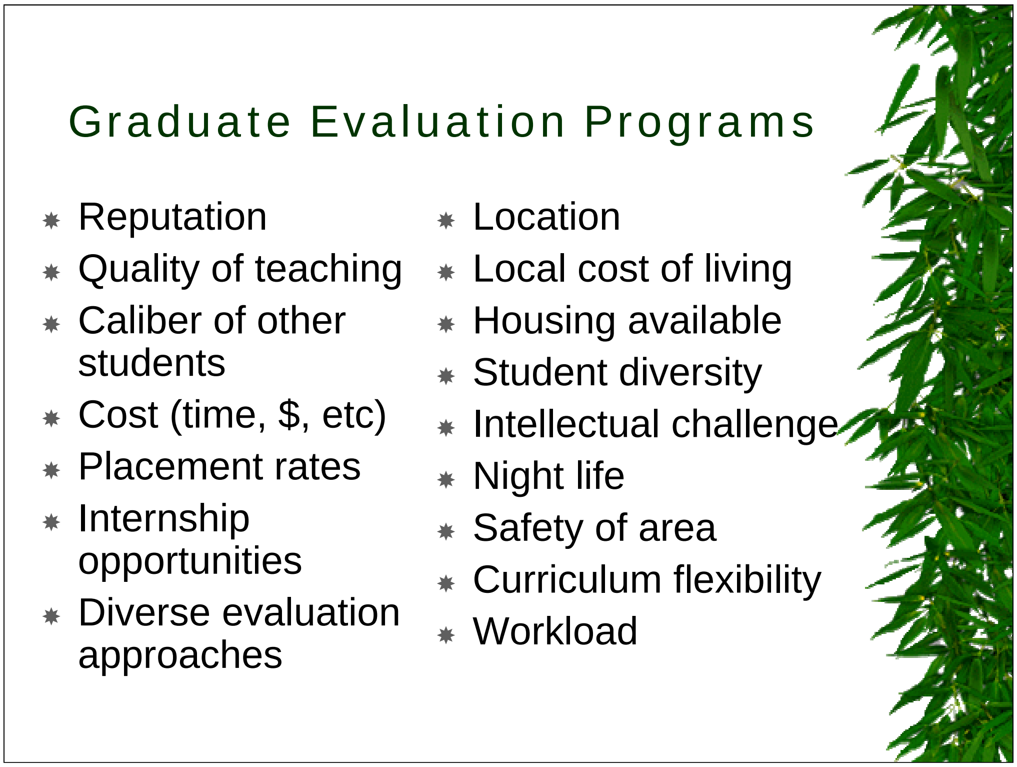
The following issues with NWS will be illustrated:

Weights vs. bars

Swamping

Nonlinear distribution of value

Next, it will be demonstrated how QWS avoids these serious problems.

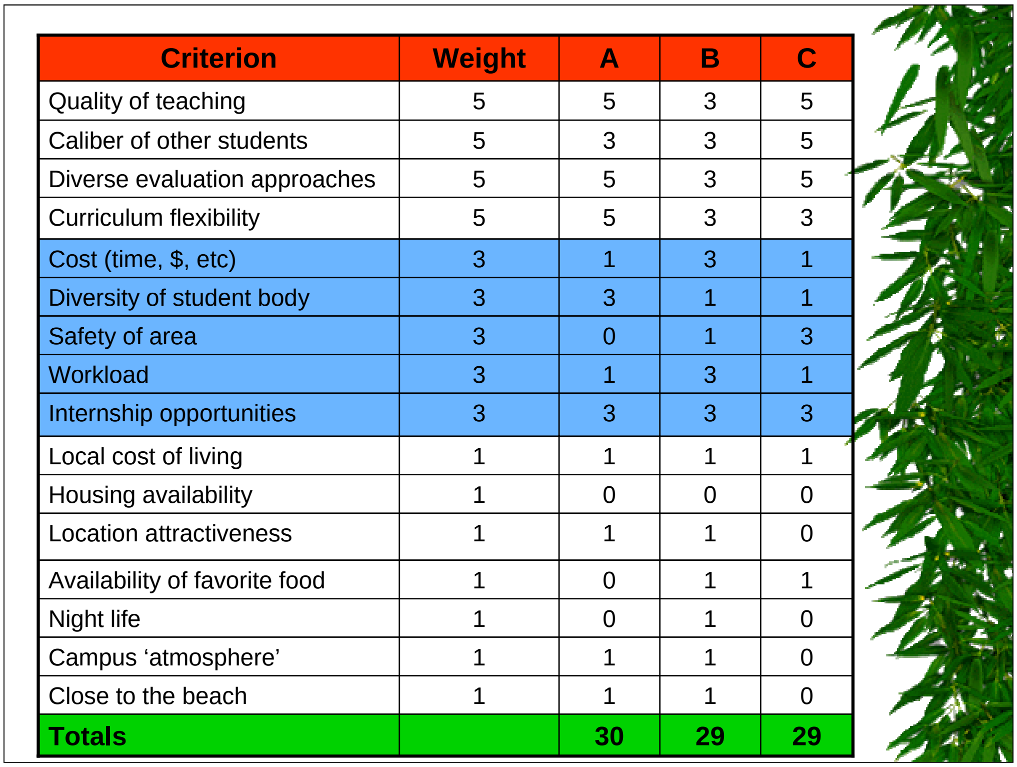


Graduate Evaluation Programs

- Reputation
- Quality of teaching
- Caliber of other students
- Cost (time, $, etc)
- Placement rates
- Internship opportunities
- Diverse evaluation approaches
- Location
- Local cost of living
- Housing available
- Student diversity
- Intellectual challenge
- Night life
- Safety of area
- Curriculum flexibility
- Workload

In order to illustrate some key points about synthesis, let us begin with an example. The above list represents a selection of some of the key criteria for what makes a good (i.e., well-functioning) graduate evaluation program.

Most evaluators would have no trouble finding measures or indicators of each of these criteria, but how would one pull together performance on all 16 to draw an overall conclusion about, say, which was the best out of three programs?



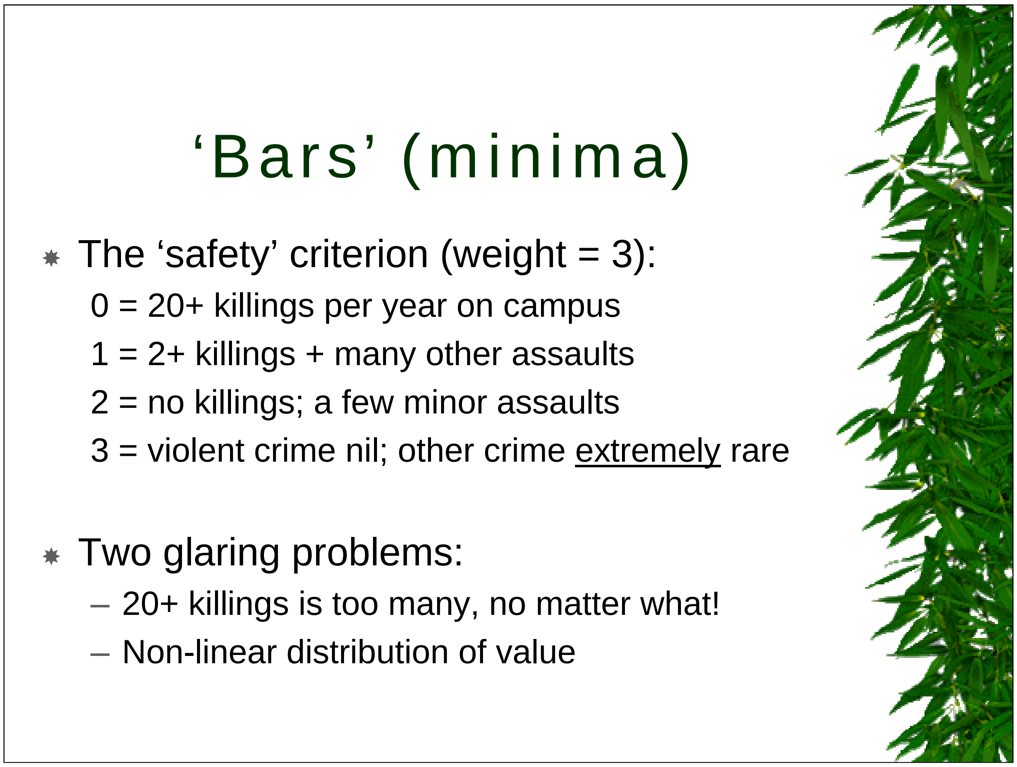
Criterion	Weight	A	B	C
Quality of teaching	5	5	3	5
Caliber of other students	5	3	3	5
Diverse evaluation approaches	5	5	3	5
Curriculum flexibility	5	5	3	3
Cost (time, $, etc)	3	1	3	1
Diversity of student body	3	3	1	1
Safety of area	3	0	1	3
Workload	3	1	3	1
Internship opportunities	3	3	3	3
Local cost of living	1	1	1	1
Housing availability	1	0	0	0
Location attractiveness	1	1	1	0
Availability of favorite food	1	0	1	1
Night life	1	0	1	0
Campus ‘atmosphere’	1	1	1	0
Close to the beach	1	1	1	0
Totals		**30**	**29**	**29**

Based on a quick poll of student opinions (of course, one would usually be more systematic – this was just for illustrative purposes), we categorized the criteria by importance and applied numerical weights of 1 (minimal importance), 3 (moderate importance), or 5 (maximum importance).

We then rated three hypothetical graduate programs on each criterion, using the weight as the maximum score, and then summed the scores. [This is numerically equivalent to scoring each criterion on (say) a five-point scale, multiplying by the weights, and then summing the weighted scores. The only difference is that the numbers would be larger (but proportionate).]

Three important results should be noted:

1. School A, which scores excellently on many criteria, scores extremely low on safety. Further investigation reveals that there are more than 20 killings per year on campus. Cause for concern?
2. The difference between Schools A and B on the safety criterion has half the value in the overall score as the difference between B and C. But are these differences of equal value, as the numerical weighting system assumes? [Note that even if you juggle them so they are, it is mighty tricky to get this kind of equivalence within a scale, let alone across criteria (another assumption necessary when the numbers are summed).]
3. School B and School C have come out equal. But School B has only mediocre scores on the most important criteria, but has aced almost all the criteria of ‘minor importance’; School C is the opposite. [This problem gets much worst with large lists of minor criteria.]



‘Bars’ (minima)

- The ‘safety’ criterion (weight = 3):

  0 = 20+ killings per year on campus

  1 = 2+ killings + many other assaults

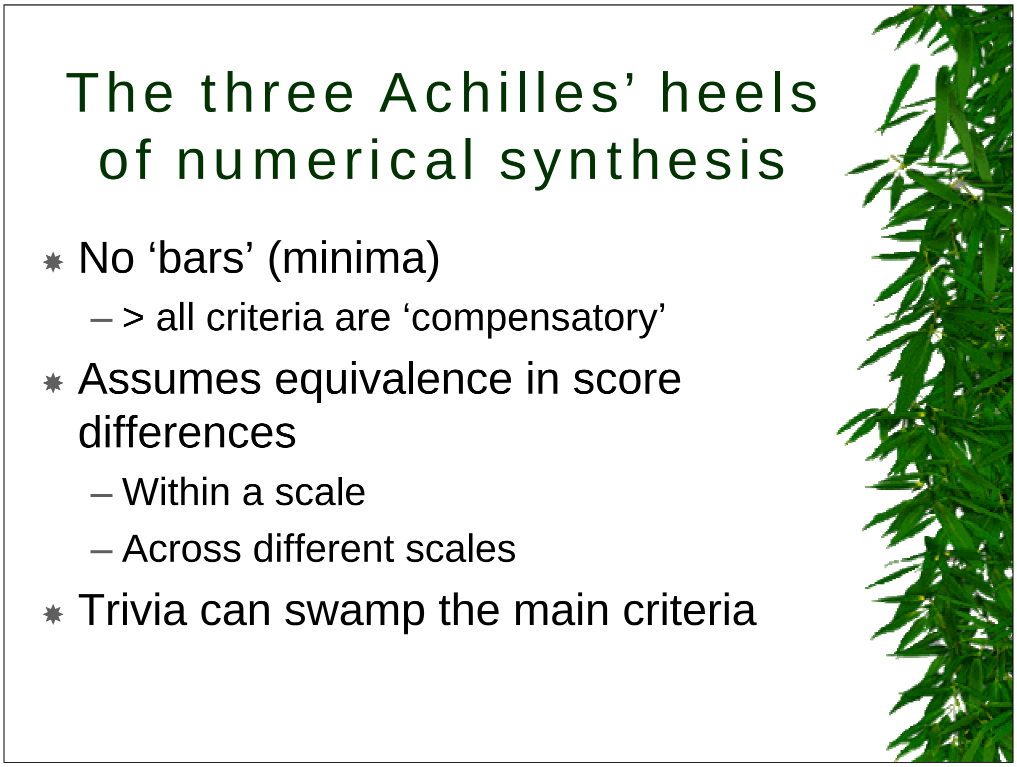
  2 = no killings; a few minor assaults

  3 = violent crime nil; other crime extremely rare

- Two glaring problems:
  - 20+ killings is too many, no matter what!
  - Non-linear distribution of value

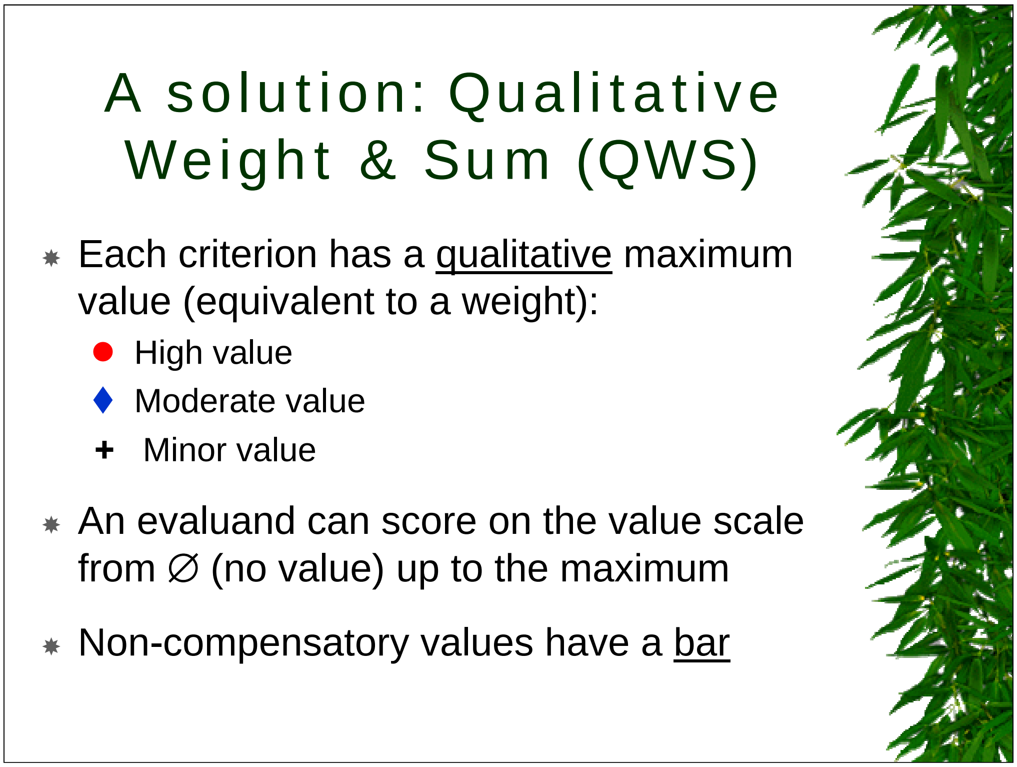
When we look more closely at the rating criteria for the (hypothetical) safety scale, we can see that:

1. More than 20 killings should be unacceptable, no matter how good the program is in other respects. In other words, a score like this should be non-compensatory.
2. The difference between a score of 0 and 1 is qualitatively very different from a difference between 2 and 3 – yet the numerical weighting system treats them as exactly equivalent! This is known as the non-linear distribution of value along the rating scale.



The three Achilles' heels of numerical synthesis

- No 'bars' (minima)
  - – > all criteria are 'compensatory'
- Assumes equivalence in score differences
  - – Within a scale
  - – Across different scales
- Trivia can swamp the main criteria



A solution: Qualitative Weight & Sum (QWS)

- Each criterion has a qualitative maximum value (equivalent to a weight):
  - ● High value
  - ♦ Moderate value
  - + Minor value
- An evaluand can score on the value scale from ∅ (no value) up to the maximum
- Non-compensatory values have a bar

In QWS:

Each criterion has a qualitative weight – this removes the temptation to combine different values in an invalid way.

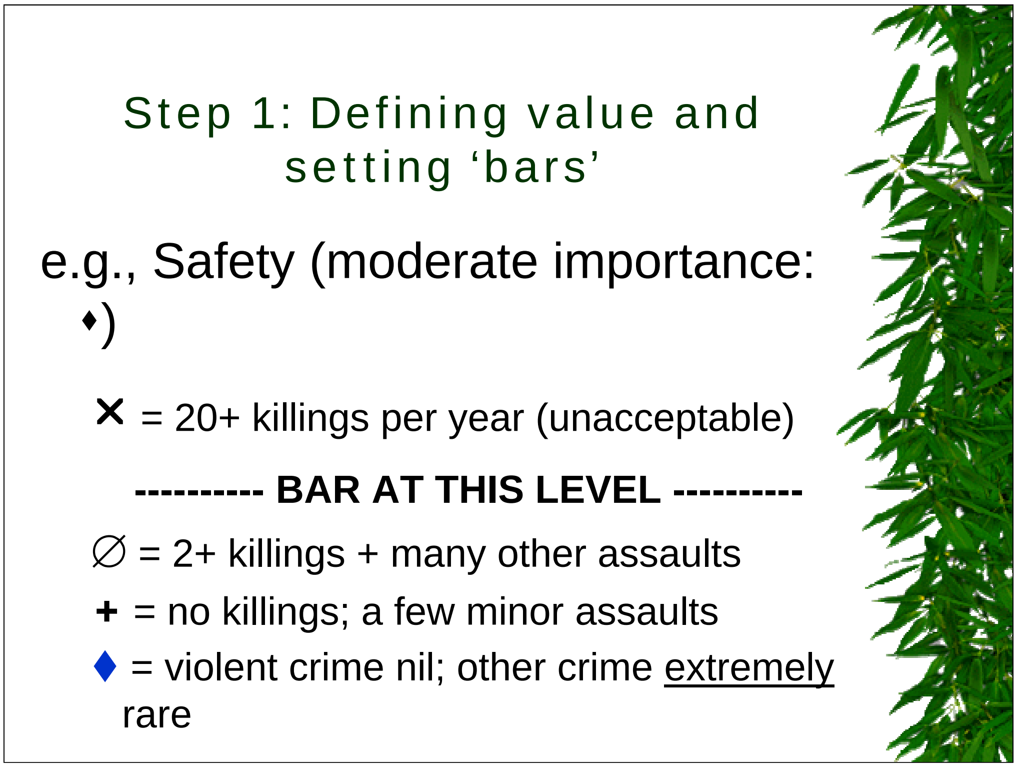
As with the quantitative system, each evaluand can score anything up to the ‘weight’ of that criterion. In other words:

● High value – an evaluand can score ∅ (no value), + (minor value), ♦ (moderate value), or ● (high value).

♦ Moderate value - an evaluand can score ∅ (no value), + (minor value), or ♦ (moderate value).

+ Minor value - an evaluand can score ∅ (no value), or + (minor value).

If an evaluand scores below the bar on any one criterion, it is deemed unsatisfactory (i.e., it ‘flunks’ overall). [This is an important consideration when deciding where to set the ‘bar’ – it is, after all, the ‘death card’ for an evaluand. There are often situations where one sets bars like this (e.g., in personnel evaluation, saying a candidate needs a minimum of X years experience, or fluency in a language), only to find that none of the available candidates fits the bill, and the bars have to be reset.]



Step 1: Defining value and setting ‘bars’

e.g., Safety (moderate importance: ♦)

✕ = 20+ killings per year (unacceptable)

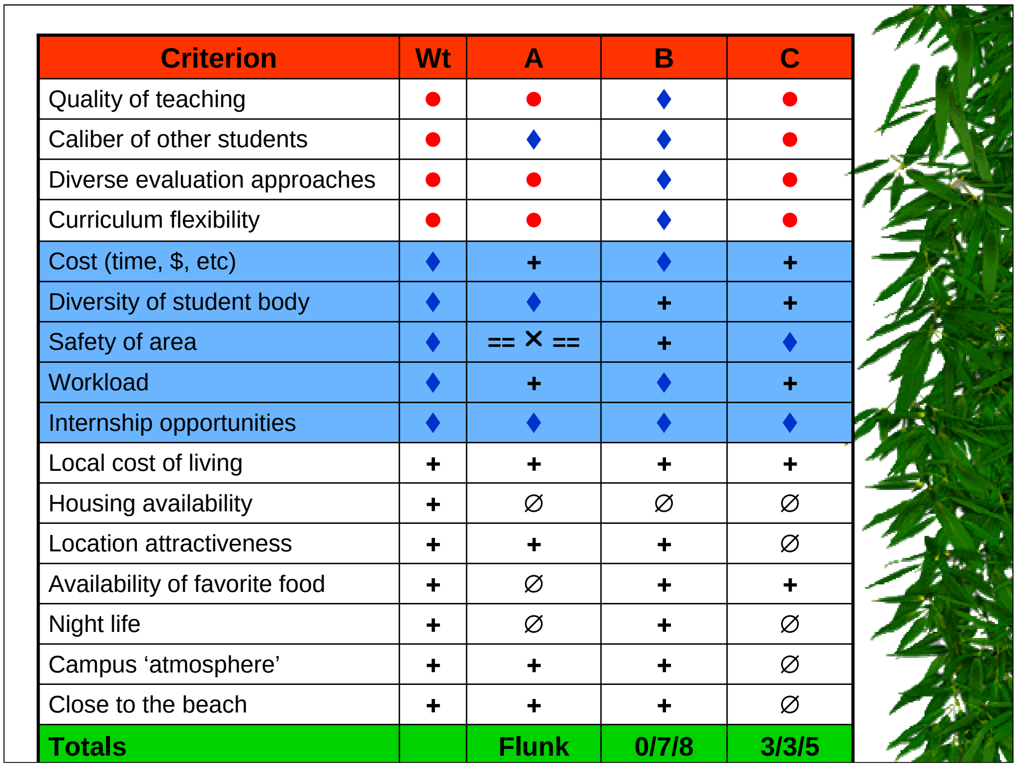
---------- BAR AT THIS LEVEL ----------

∅ = 2+ killings + many other assaults

+ = no killings; a few minor assaults

♦ = violent crime nil; other crime <u>extremely</u> rare

Here, 20+ killings on campus, and we are definitely prepared to say we would not consider a school with that many. [Strictly speaking, a school with 19 killings (which is more than 2 but fewer than 20) would be rated ∅; however, the reality is that the ‘bar’ is often a bit more of a ‘grey area’ than that.]

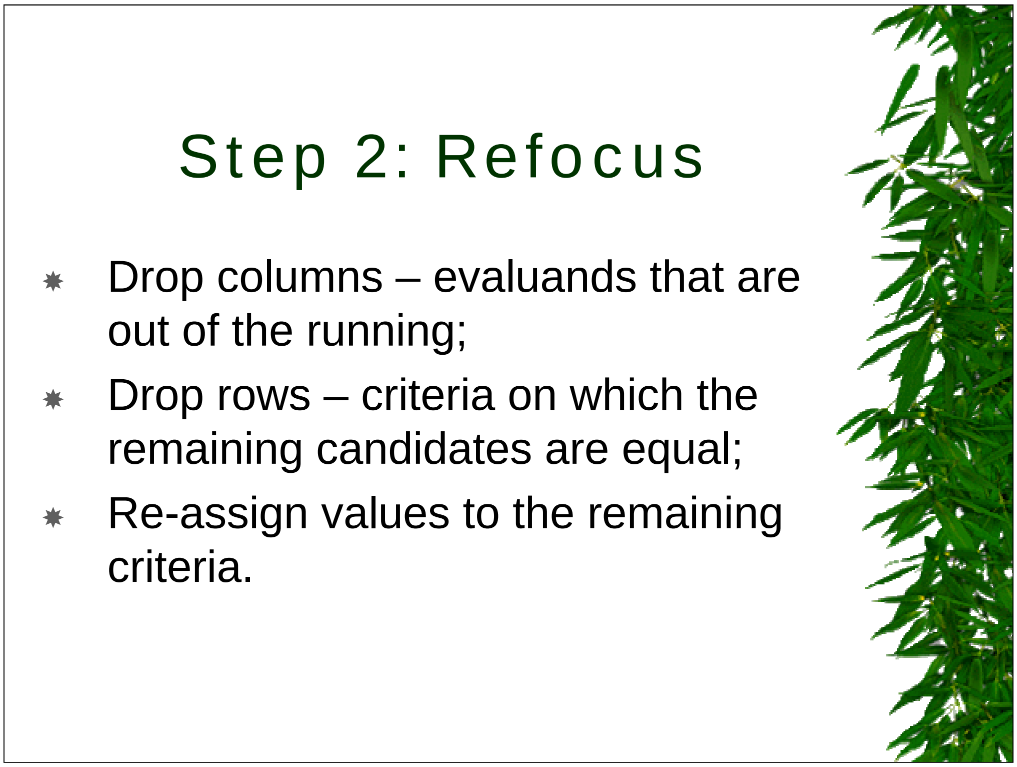


Criterion	Wt	A	B	C
Quality of teaching	●	●	♦	●
Caliber of other students	●	♦	♦	●
Diverse evaluation approaches	●	●	♦	●
Curriculum flexibility	●	●	♦	●
Cost (time, $, etc)	♦	+	♦	+
Diversity of student body	♦	♦	+	+
Safety of area	♦	== ✕ ==	+	♦
Workload	♦	+	♦	+
Internship opportunities	♦	♦	♦	♦
Local cost of living	+	+	+	+
Housing availability	+	∅	∅	∅
Location attractiveness	+	+	+	∅
Availability of favorite food	+	∅	+	+
Night life	+	∅	+	∅
Campus ‘atmosphere’	+	+	+	∅
Close to the beach	+	+	+	∅
Totals		**Flunk**	**0/7/8**	**3/3/5**

Here we do the same synthesis with schools A through C using the QWS. Note that:

1. School A flunks immediately for failing to clear the ‘bar’ on safety
2. Instead of one total, we now have three totals – one for each level of value. Most importantly, we can now tell at a glance that School B scores higher on the (relative) trivia, while School C aces three of the four most critical features.

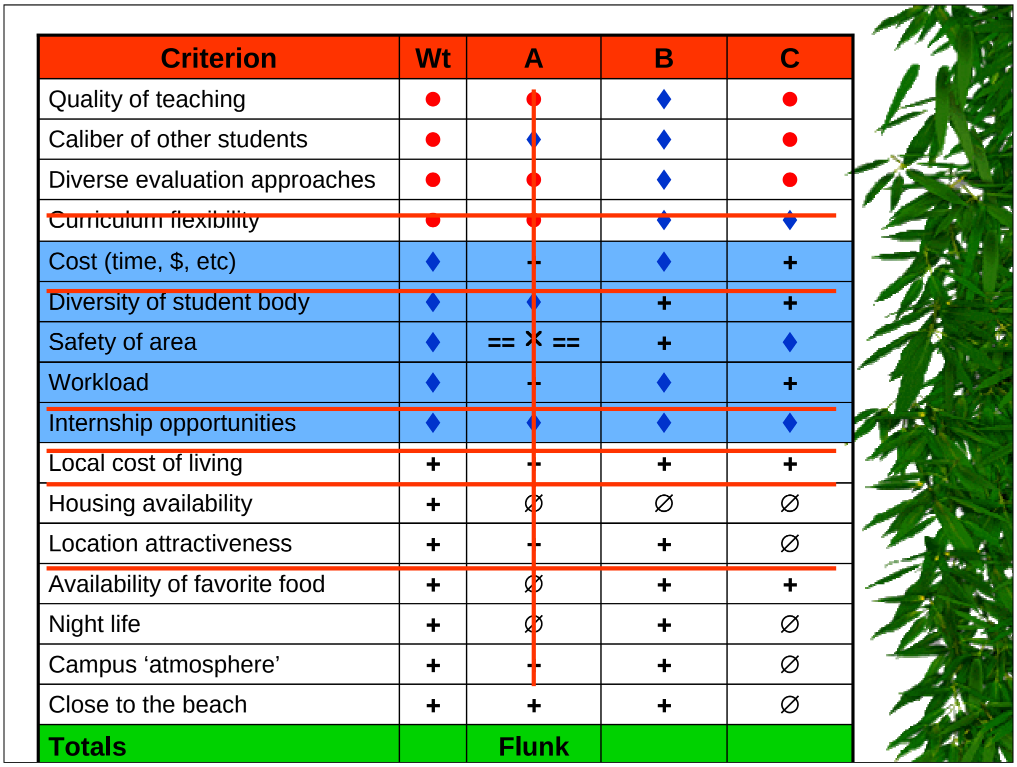
However, it’s still not 100% clear which should be the winner. So now we refocus in order to hone in on the real differences between the two schools on the ‘shortlist’.



Step 2: Refocus

- Drop columns – evaluands that are out of the running;
- Drop rows – criteria on which the remaining candidates are equal;
- Re-assign values to the remaining criteria.

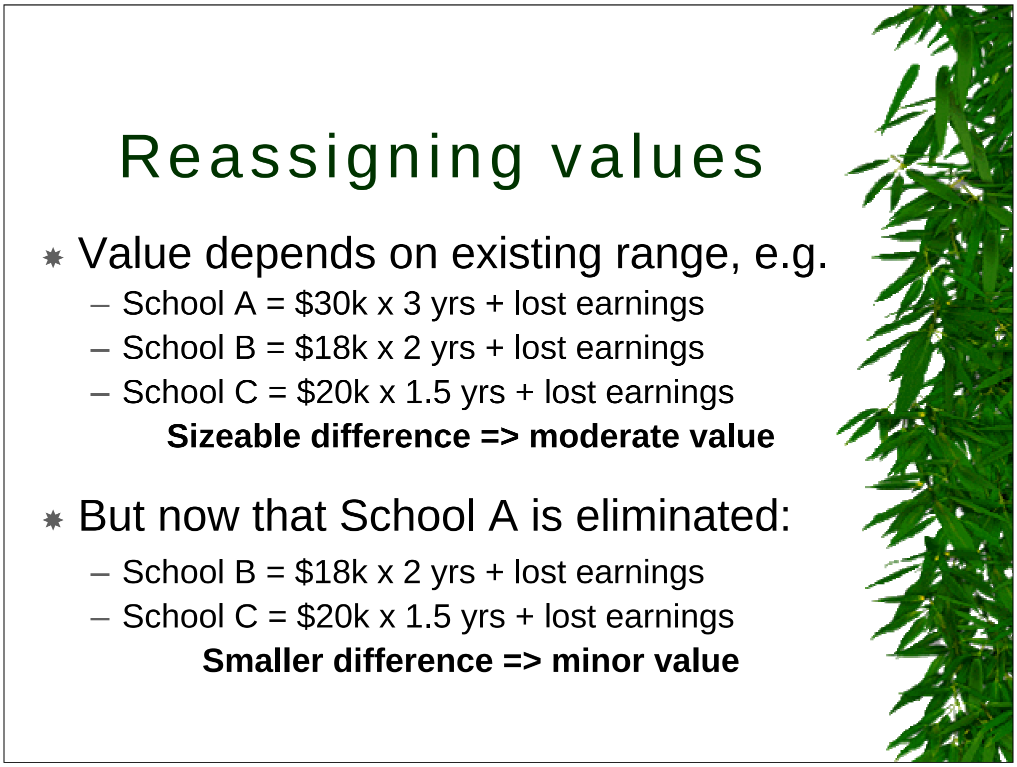
See next slide for an illustration of what we mean here ...



Criterion	Wt	A	B	C
Quality of teaching	●	●	♦	●
Caliber of other students	●	♦	♦	●
Diverse evaluation approaches	●	●	♦	●
~~Curriculum flexibility~~	●	●	♦	♦
Cost (time, $, etc)	♦	-	♦	+
~~Diversity of student body~~	♦	♦	+	+
Safety of area	♦	== X ==	+	♦
Workload	♦	-	♦	+
~~Internship opportunities~~	♦	♦	♦	♦
~~Local cost of living~~	+	-	+	+
~~Housing availability~~	+	∅	∅	∅
Location attractiveness	+	-	+	∅
~~Availability of favorite food~~	+	∅	+	+
Night life	+	∅	+	∅
Campus ‘atmosphere’	+	-	+	∅
Close to the beach	+	+	+	∅
Totals		**Flunk**		

In the refocus step, we:

- Drop columns – evaluands that have flunked, or are clearly out of the running;
- Drop rows – criteria on which the remaining candidates are equal;
- Re-assign values to the remaining criteria (see next slide for an example).



Reassigning values

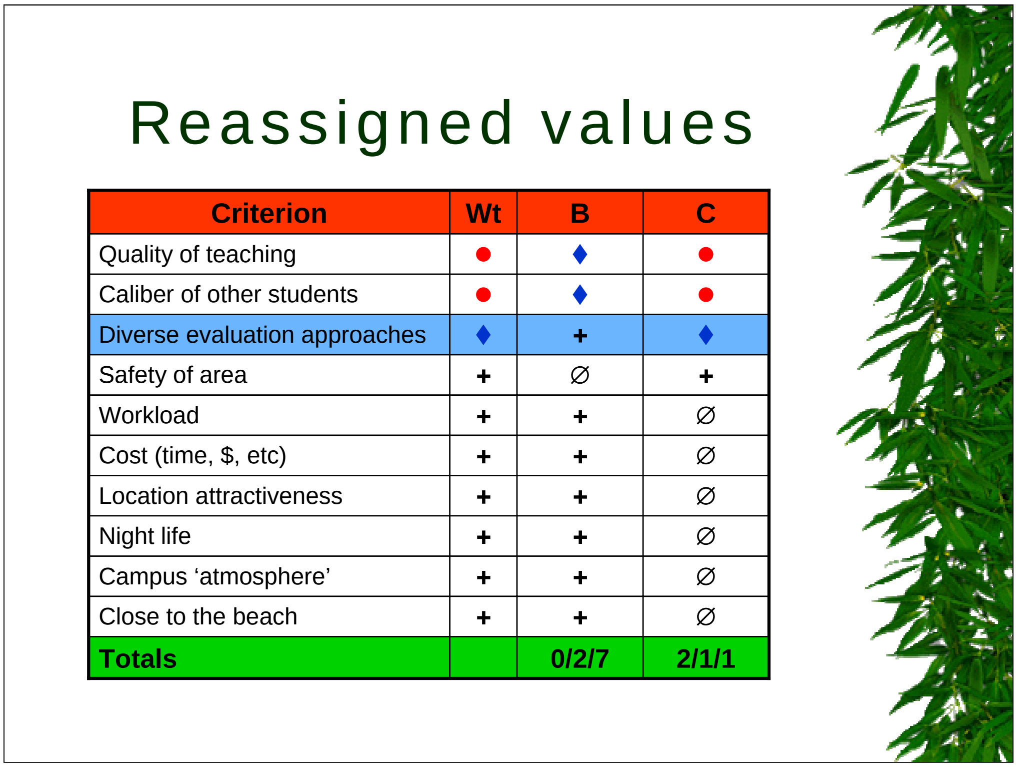
- Value depends on existing range, e.g.
  - School A = $30k x 3 yrs + lost earnings
  - School B = $18k x 2 yrs + lost earnings
  - School C = $20k x 1.5 yrs + lost earnings

  Sizeable difference => moderate value

- But now that School A is eliminated:
  - School B = $18k x 2 yrs + lost earnings
  - School C = $20k x 1.5 yrs + lost earnings

  Smaller difference => minor value

Note that the weight of a criterion (a.k.a. dimension) depends on the range involved. If school fees range from $18,000 to $30,000 (plus lost earnings), this is substantial, and gets an ‘important’ weight. When the difference is only about $2000 a year (plus some much smaller lost earnings), the differences mean much less and so the criterion is given a smaller weight.

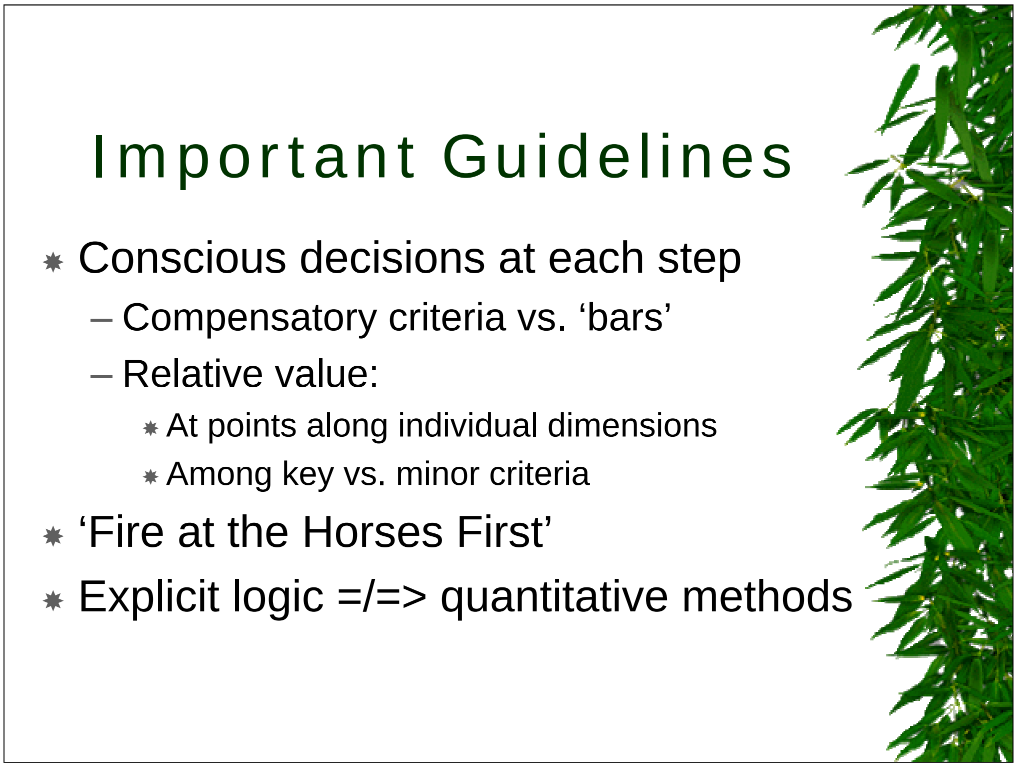


Reassigned values

Criterion	Wt	B	C
Quality of teaching	●	♦	●
Caliber of other students	●	♦	●
Diverse evaluation approaches	♦	+	♦
Safety of area	+	∅	+
Workload	+	+	∅
Cost (time, $, etc)	+	+	∅
Location attractiveness	+	+	∅
Night life	+	+	∅
Campus ‘atmosphere’	+	+	∅
Close to the beach	+	+	∅
Totals		**0/2/7**	**2/1/1**

Note that, after the refocus, it is now absolutely clear where the differences/trade-offs lie between the two candidates. Here, what we call ‘triad arithmetic’ kicks in – we have to make decisions about how many + scores should be considered equal to a ● score, and so on.

If the decision is still not clear, one may have to refocus again, trying for finer gradations. If even that does not produce a clear winner, it may be time to go on holistic impressions to make the final decision, e.g., what is the ‘feel’ of the school as you walk around, etc.



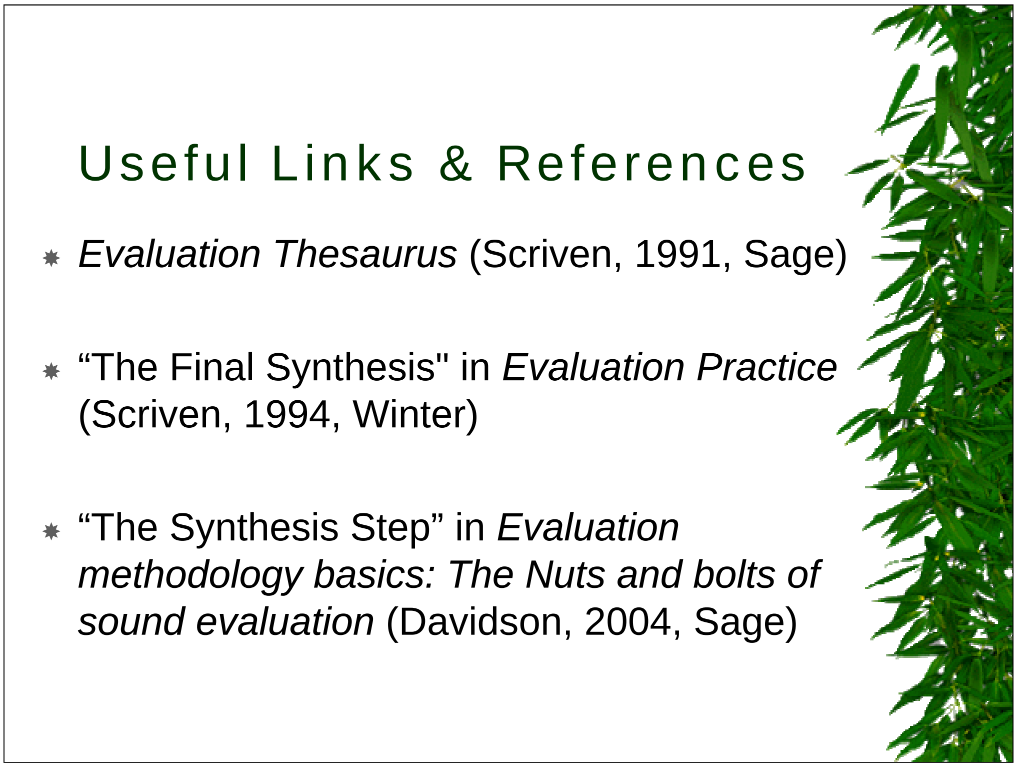
Important Guidelines

- Conscious decisions at each step
  - Compensatory criteria vs. ‘bars’
  - Relative value:
    - At points along individual dimensions
    - Among key vs. minor criteria
- ‘Fire at the Horses First’
- Explicit logic =/=> quantitative methods

The main advantage of the QWS is that, unlike the NWS (in which the numbers make the decisions for you), the decisions are made consciously all along the way.

‘Fire at the horses first’ means that you should first check that candidates clear the ‘bars’ to make sure you don’t waste resources evaluating them on the other dimensions.

Just because the evaluative method is qualitative doesn’t mean that we shouldn’t be explicit and transparent in how it’s laid out.



Useful Links & References

- *Evaluation Thesaurus* (Scriven, 1991, Sage)
- “The Final Synthesis" in *Evaluation Practice* (Scriven, 1994, Winter)
- “The Synthesis Step” in *Evaluation methodology basics: The Nuts and bolts of sound evaluation* (Davidson, 2004, Sage)