ISAHP ' 96, Vancouver, Canada, July 12-15, 1996

BENCHMARK MEASUREMENT: BETWEEN RELATIVE AND ABSOLUTE

William C. Wedley, Eng Ung Choo and Bertram Schoner
Faculty of Business Administration
Simon Fraser University,Burnaby, B. C. Canada, V5A 1S6
wedley@sfu.ca/ schoner@sfu.ca/ choo @sfu.ca

Abstract Benchmark measurement is a new procedure for gaining the advantages of absolute measurement with the benefits of relative measurement. Rather than comparing items to all relevant alternatives as in relative measurement or rating with known fixed standards as in absolute measurement, benchmark measurement makes comparisons to a set of predetermined benchmark alternatives. The main advantage of this new technique is the ability to evaluate many alternatives when absolute standards are not available.

Introduction

At the Third International Symposium on the Analytic Hierarchy Process, Matti Verkasalo (1994) presented a paper on repetitive use of the uppermost levels of an AHP hierarchy. Verkasalo was concerned that too much time was required on the part of executives at Nokia Telecommunications each time relative measurement is used for a major decision. Even though the various problems seemed to have common corporate objectives, each new application of AHP required a unique structure and numerous paired comparisons. In order to avoid continual restructuring, Verkasalo investigated the possibility of using the same upper portion of a hierarchy like a template for similar problems. He concluded that this would be possible if the decisions formed a coherent group and if certain precautions were made to avoid AHP problems such as rank reversal and structural adjustments.

As a consequence of that paper, discussions took place with Verkasalo on how a referenced or linking pin mode of aggregation can allow new alternatives to be added to the hierarchy without destroying the relative ratios of original alternatives (Schoner, Wedley and Choo, 1993). Further investigation lead to the realization that an indeterminate number of additional alternatives can be added to a template hierarchy provided the original hierarchy and its alternatives are treated benchmarks to which all new alternatives are compared. It was from this insight that lead to the concept of benchmark measurement.

Further inquiry lead to the realization that this process had traits of both relative and absolute measurement, but cound not be characterized as belonging to either. With relative measurement, direct pairwise comparisons are made of elements at all levels of the hierarchy, including bottom level alternatives. This procedure is used when the number of alternatives is small and when there are no objective standards on which to judge those alternatives. The other mode, absolute measurement, is used when the number of alternative is large and when well-known standards are available to judge them. Benchmark measurement, like absolute measurement, does not require all alternatives to be evaluated at one time relative to each other (as does relative measurement). But like relative measurement, it includes alternatives as part of the hierarchy and not as a separate rating procedure. In benchmark measurement, temporally separated alternatives can be added to the hierarchy in a manner that puts them on the same scale as the original benchmark items

This paper outlines this new measurement mode which falls between absolute and relative measurement. We call it benchmark measurement, because it relies upon comparisons to benchmark alternatives which are part of a base hierarchy. In order to develop the concepts of benchmark measurement, we first review the use of relative and absolute measurement. Next, we introduce the benchmark procedure and explain why it is a new mode of measurement. Then we give a mathematical formulation and an illustrative example. The discussion explains the significance of this new mode of measurement and contrasts it with current AHP practice.

The authors would like to thank the Natural Sciences and Engineering Research Council of Canada for financial support of this project.

Relative and Absolute Measurement

Relative measurement is the standard form for AHP applications. It requires structuring a problem into hierarchical levels whereby the lower levels are dependent upon or contribute to the immediate higher level. Pairwise comparisons are made amongst the items below each hierarchical element, and an eigenvector routine is used to derive ratio scaled priorities for the items (frequently called local weights). These local weights under any node sum to unity. Once priorities are derived for all nodes, the local weights are multiplied by the priorities of higher levels to yield globally weighted priorities (usually called global weights). Then, for each unique alternative at the bottom level, these global weights are synthesized (weighted and added) to yield composite weights. These composite weights, like the local and global weights, are supposed to be in ratio form.

Types of Relative Measurement

Two forms of relative measurement are commonly recognized: the distributive and ideal methods. Both derive the same local priorities, but they differ in how weights are allocated downward to form global priorities. For the distributive mode, Saaty (1993) has given the following equation for determining the composite weights (W_i) of the alternatives:

$$W_i = \sum_{j=1}^{m} W_{ij}\left(x_j \Big/ \sum_{k=1}^{n} W_{kj} \right) \qquad (1)$$

where $x_j,\ j = 1,\ldots, m,$ is the weight of the j^{th} criterion, $\sum_{j=1}^{m} x_j = 1$; and W_{ij} are ratio values (before normalization) of the i^{th} alternative for that criterion. Here, k is a counter for the number of alternatives and the values $1/\sum_{k=1}^{n} W_{kj}$ are scaling factors that make the local weights for the bottom level alternatives sum to one. Multiplication by the criterion weight x_j makes the sum of the global weights under a node equal to the criterion weight.

With the ideal method, alternative local weights are normalized with respect to a single criterion so that the best alternative has a value of one. This results in the global weighting process transferring 100 percent of a criterion weight to the best alternative, with less important alternatives receiving proportionally smaller amounts. (Belton and Gear, 1983).

Saaty (1993) expresses this process as:

$$W_i = \sum_{j=1}^{m} W_{ij}\left(x_j / \max_{1 \le k \le n} W_{kj} \right) \qquad (2)$$

where $1/\max_{i \le k \le n} W_{kj}$ converts all local weights of alternatives to a fraction of the best alternative (best alternative for each criterion has a value of 1).

According to Saaty (1994), the ideal mode is used when the alternatives are independent of each other and preservation of rank is desired. So long as the ideal alternative is not replaced when adding or deleting alternatives, the ideal mode will not result in reversals of rank amongst the original alternatives, although the composite ratios will change. The distributive mode, on the other hand, will allow reversals and should be used when the overall choice depends upon which alternatives are relevant. If more or fewer alternatives are present, then available resources will be distributed in a different manner and changes in rank may occur.

Two less well known methods of relative measurement are referenced AHP and linking pins AHP. Referenced AHP (Schoner and Wedley, 1989) is very similar to the distributive mode, except that upper level weights are established via reference to the alternatives in the choice set. If alternatives are added or deleted, the criteria weights must be re-evaluated. Linking pins (Schoner, Wedley and Choo, 1993) is similar to the ideal mode, except that the ideal alternative (or any other alternative) is used as a linking referent when establishing weights for upper levels of the hierarchy. Both these modified techniques anchor the criteria weights to the alternatives and avoid problems such as rank reversal and incorrect composite priorities. The distributive and ideal modes assume that the upper levels of the hierarchy are independent of lower levels whereas referenced and linking pin AHP require the upper levels to calibrate lower level scales to commensurate units. This concept of anchoring and linking the levels of the hierarchy is used in benchmark measurement.

Absolute Measurement

Absolute measurement, the other recognized mode for AHP analysis, is applied when there are well established standards for judging alternatives. Unlike the relative measurement techniques, absolute measurement has the alternatives evaluated at a later stage separate from the formal hierarchy. In the first stage when the formal hierarchy is used, intensities or indicators of the alternatives are placed below the lowest criterion level. Since there are known standards for these indicators, they can be compared amongst themselves to get relative priorities for indicators (intensities). As well, weights for the criteria are established within the hierarchical framework. Then in a second stage, the criteria and intensity weights (standards) are transferred outside the hierarchy (frequently to a spreadsheet) where the alternatives are rated for which intensity (indicator) of each criterion that they possess. Applying the criterion weight to each rated intensity and summing across criteria gives an overall score for each alternative. Since the alternatives are judged one at a time, they do not all have to be evaluated at the same time.

The overall absolute ratings (Saaty, 1994) can be determined by:

$$W_i = \sum_{j=1}^{m} W_{ij} x_j \qquad (3)$$

where x_j, j = 1, ... , m, is the weight of the jth criterion and W_{ij} are priorities of rated intensities for the ith alternative. In this rating procedure, each alternative is designated as having a specific intensity rating for each criterion. So long as there is no change in the number of criteria, the indicator intensities, or the ratings each alternative receives, there can be no change in relative rank. Thus, absolute measurement preserves rank.

Comparison of Absolute and Relative Measurement

The main disadvantage of relative measurement is that it can only handle a modest number of alternatives. Since it requires direct paired comparison of the alternatives, the process becomes cumbersome, tiring and unmanageable if the number of alternatives is large. Moreover, the requirement of comparisons relative to each other means that all alternatives have to be present and evaluated at the same time.

For absolute measurement, on the other hand, numerous alternatives can be analyzed at different times. This attribute is important because there are many situations like selection and admission decisions where we do not want to wait until all candidates are present to make a decision. In absolute mesurement, we have a method to evaluate independent alternatives at different times, but it requires rigid standards and does not have the precision of paired comparisons.

Relative measurement techniques offer the advantages of greater precision from direct comparison of alternatives, but they are cumbersome when it comes to analyzing numerous alternatives over a period of time. Moreover, the distributive and ideal methods of relative measurement give different composite answers to the same problem, leaving decision makers confused about which method to use. Although the ideal mode could be used with temporally separated alternatives without rank changes (if the ideal alternatives remain unchanged), it does not preserve original ratios. The linking pin and referenced methods preserve both rank and ratio relationships when new alternatives are added, but they too suffer from the number of alternatives that can be analyzed in a single hierarchy. What is needed is a technique which maintains the precision of paired comparisons while still allowing numerous alternatives to be analyzed, free from reversals or altered ratios.

The Benchmark Measurement Procedure

The solution that overcomes the main disadvantages of relative and absolute measurement is benchmark measurement. As will be shown, the benchmark procedure can handle large numbers of alternatives that can be presented at different times, yet without the need for fixed standards Benchmark measurement has many of the advantages of absolute measurement, yet it uses the precision of direct pairwise comparison of alternatives. It is a combination of the best attributes of relative measurement and absolute measurement.

Benchmark measurement, like absolute measurement, is a two phase process. First a template hierarchy is established with benchmark alternatives. These benchmark alternatives must be well known to the evaluator and provide good dispersion across the range of potential alternatives. Next, the template structure is evaluated in the usual AHP manner for relative measurement, except that the weights for the criteria and other upper levels are established in relation to the chosen benchmark alternatives. In this manner, the template becomes an integrated standard or anchor for comparing all other alternatives.

In the second phase, new alternatives are added to the bottom level clusters of the benchmark hierarchy, but they are not evaluated in relation to themselves (since they are independent of each other). Rather, they are compared only to the benchmark items, leaving part of the pairwise comparison matrix incomplete. Then, priorities based upon comparisons to and amongst benchmark items are determined using Harker's (1987) concepts for incomplete matrices. Finally, synthesis is carried out in a manner which normalizes overall values of the benchmark items to sum to one. The composite results of other alternatives, are expressed as ratios of the benchmarks.

If two or more new alternatives are added to the bottom of the incomplete matrix, it is possible for their interaction with the benchmarks to affect the final composite priorities. To remove the effect of other non-benchmark alternatives, each new alternative could be considered in a matrix which contains just that alternative and the benchmarks. Alternatively, or in addition, the sole effect of the new alternative has less impact if perfect consistency is assumed to exist between the benchmark alternatives. This could be achieved by using the ratios of the benchmark priorities as the benchmark comparisons in the incomplete matrix. This latter process would be treating the benchmark comparisons like absolute comparisons rather than relative comparisons.

If desired, the composite weights of all alternatives, benchmark and non-benchmark, could be renormalized to sum to one. This final step, however, would not alter the relative ranks or ratios. We should also note that there is no theoretical upper limit to the size of the incomplete comparison matrix. Since new alternatives are only compared to the benchmark items, the number of comparisons for each new alternative is kept to manageable size while gaining the advantage of handling large numbers of alternatives.

Mathematical Foundations

Let n_B be the number of benchmark alternatives. Without loss of generality, we may assume that the first n_B of A_1, ..., A_n are the benchmark alternatives. Composite benchmark priorities are calculated by:

$$W_i = \sum_{j=1}^{m} W_{ij} \left(b_j \Big/ \sum_{k=1}^{n_B} W_{kj} \right) \qquad (4)$$

where b_j, $j = 1, \ldots, m$, is the benchmark weight of the j^{th} criterion and W_{ij} are unnormalized ratio ratings of all alternatives, benchmark and non-benchmark. This formulation will cause $\sum_{i=1}^{n_B} W_i$ to sum to one and all other W_i, $n > n_B$, to be in ratio form to this benchmark set.

In matrix notation, the benchmark weights are:

$$W = A\ S_B\ b = A\ N(N^{-1} S_B)\ b \qquad (5)$$

where A is comprised of unnormalized W_{ij} values, S_B is the baseline diagonal matrix for normalizing local benchmark priorities to sum to one, N is a diagonal matrix which represents different ways of normalizing A, and b is the vector of benchmark criteria weights.

$$(S_B)_{jj} = 1 \Big/ \sum_{i=1}^{n_B} W_{ij} \qquad (6)$$

is a fixed structural criterion which assures that the unnormalized benchmark alternatives will always sum to one as A changes with the addition of new alternatives. If A is normalized in any particular manner, then as shown in (5), this may be interpreted as rescaling S_B or b. This differs from the ideal or distributive methods of relative measurement where the structural factor is replaced by the new normalization factor.

An Illustrative Example

To illustrate the benchmark measurement, we have selected a personnel evaluation problem posed by Forman (1994) and discussed by Saaty (1993) in the context of the rank reversal/preservation debate. In this example, three employees, Susan, John and Michelle, are evaluated according to seven criteria: dependability, education, experience, quality of work, quantity of work, attitude and leadership ability. Later, a fourth alternative, Bernard, is added to the choice set.

We will treat Susan, John and Michelle as the benchmark alternatives and we will show how Bernard can be added via benchmark comparisons. Then, we will modify the problem to show how three other individuals, Bill, Bert and Eng, can be added to the candidate list.

The baseline AHP priorities expressed in matrix notation are as follows:

$$\begin{array}{l} \\ \text{Susan} \\ \text{John} \\ \text{Michelle} \end{array} \begin{array}{c} \begin{array}{ccccccc} \text{C1} & \text{C2} & \text{C3} & \text{C4} & \text{C5} & \text{C6} & \text{C7} \end{array} \\ \begin{bmatrix} .429 & .412 & .263 & .421 & .391 & .409 & .389 \\ .333 & .294 & .474 & .316 & .304 & .318 & .278 \\ .238 & .294 & .263 & .263 & .304 & .273 & .333 \end{bmatrix} \end{array} = A\ S_B \tag{7}$$

where C1, ..., C7 represent the seven criteria. We are not given unnormalized ratio ratings of alternatives, A, for each criterion, but this is not important since they would be simply a proportional transformation of the columns of (7). Therefore, let us regard the matrix in (7) as A, which means that S_B is simply an identity matrix. Accordingly, the unit values of column sums in (7) becomes our benchmark normalization factor, S_B.

The benchmark priorities for the criteria normalized to sum to one are presented in (8).

$$\begin{array}{ll} \text{Dependability (C1)} & .172 \\ \text{Education (C2)} & .060 \\ \text{Experience (C3)} & .344 \\ \text{Qualifications (C4)} & .211 \\ \text{Quantity (C5)} & .130 \\ \text{Attitude (C6)} & .045 \\ \text{Leadership (C7)} & .038 \end{array} = b \tag{8}$$

We assume that these baseline criteria weights relate to the benchmark hierarchy and are derived in reference to Susan, John and Michelle, the benchmark alternatives (Schoner and Wedley, 1989). Thus, they can be applied to (7) via multiplication to transform the alternative local priorities to global weights that are in commensurate units.

$$W = (AS_B)b = \underbrace{\begin{bmatrix} .429 & .412 & .263 & .421 & .391 & .409 & .389 \\ .333 & .294 & .474 & .316 & .304 & .318 & .278 \\ .238 & .294 & .263 & .263 & .304 & .273 & .333 \end{bmatrix}}_{(AS_B)} \bullet \overbrace{\begin{bmatrix} .172 \\ .060 \\ .344 \\ .211 \\ .130 \\ .045 \\ .038 \end{bmatrix}}^{b} = \begin{array}{l} \begin{bmatrix} .362 \\ .369 \\ .269 \end{bmatrix} \begin{array}{l} \text{Susan} \\ \text{John} \\ \text{Michelle} \end{array} \end{array} \tag{9}$$

Notice that John has the highest composite priority (.369), followed by Susan (.362) and Michelle (269). John is .369/.362 = 1.02 times better than Susan and 1.37 times better than Michelle.

Next, a fourth employee, Bernard, is introduced to the evaluation procedure. With Bernard added, the normalized local priority weights reported by Forman (1994) are :

$$\begin{array}{l} \\ \text{Susan} \\ \text{John} \\ \text{Michelle} \\ \text{Bernard} \end{array} \begin{array}{c} \begin{array}{ccccccc} \text{C1} & \text{C2} & \text{C3} & \text{C4} & \text{C5} & \text{C6} & \text{C7} \end{array} \\ \begin{bmatrix} .347 & .368 & .192 & .381 & .360 & .375 & .350 \\ .269 & .263 & .346 & .286 & .280 & .292 & .250 \\ .192 & .263 & .192 & .238 & .280 & .250 & .300 \\ .192 & .105 & .269 & .095 & .080 & .083 & .100 \end{bmatrix} \end{array} = \hat{A}\hat{N} \tag{10}$$

where $\hat{A}$ is the unnormalized matrix of alternatives with Bernard added. $\hat{A}$ is the same as (7) except that an extra row has been added for Bernard. Since $\hat{A}$ does not sum to 1, a new normalizing matrix, $\hat{N}$, was required to make the columns of (10) to sum to 1

For benchmark measurement, the columns of (10) are derived by comparing Bernard to the three benchmark candidates. For example, Figure 1 shows the output from a computer program called AHP Tree for the first column of (10). AHP Tree allows either complete or incomplete comparisons but in this instance the comparisons are

complete, because Bernard, the only non-benchmark alternative, is compared to all the others. Later when we add more candidates, we will show the incomplete procedure.

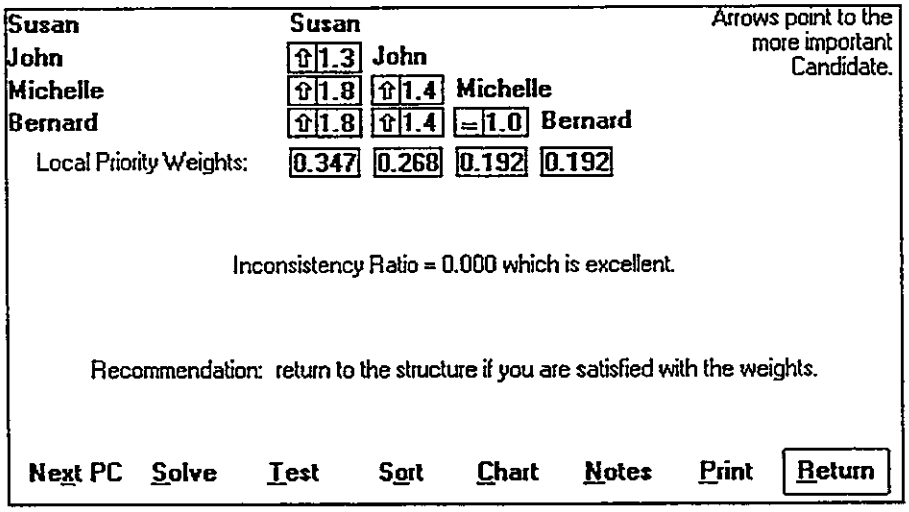

Figure 1 -- Benchmark Comparisons with Bernard Added

If we treat (10) as a proportional transformation of $\hat{A}$, then our rescaled $(S_B)_{jj}$ as defined in (6) is simply the inverse of the sum ofhe first three rows of (10) (i. e. our benchmark rows) Whereas formerly $(S_B)_{jj}$ was unity, now it is:

$$\begin{bmatrix} \frac{1}{.807} & & & & & & \\ & \frac{1}{.895} & & & & & \\ & & \frac{1}{.731} & & & & \\ & & & \frac{1}{.905} & & & \\ & & & & \frac{1}{.92} & & \\ & & & & & \frac{1}{.917} & \\ & & & & & & \frac{1}{.90} \end{bmatrix} = \hat{N}^{-1} S_B = \hat{S}_B \qquad (11)$$

With Bernard added to the benchmark hierarchy, the new composite weights are:

$W = (\hat{A}\hat{N})\ \hat{N}^{-1} S_B\ b$

$$= \underbrace{\begin{bmatrix} .347 & .368 & .192 & .381 & .360 & .375 & .350 \\ .269 & .263 & .346 & .286 & .280 & .292 & .250 \\ .192 & .263 & .192 & .238 & .280 & .250 & .300 \\ .192 & .105 & .269 & .095 & .080 & .083 & .100 \end{bmatrix}}_{(\hat{A}\hat{N})} \bullet \underbrace{\begin{bmatrix} 1.24 & & & & & & \\ & 1.12 & & & & & \\ & & 1.37 & & & & \\ & & & 1.10 & & & \\ & & & & 1.09 & & \\ & & & & & 1.09 & \\ & & & & & & 1.11 \end{bmatrix}}_{\hat{N}^{-1}S_B} \bullet \underbrace{\begin{bmatrix} .172 \\ .060 \\ .344 \\ .211 \\ .130 \\ .045 \\ .038 \end{bmatrix}}_{b} = \begin{bmatrix} .362 \\ .369 \\ .269 \\ .216 \end{bmatrix} \begin{matrix} \text{Susan} \\ \text{John} \\ \text{Michelle} \\ \text{Bernard} \end{matrix} \qquad (12)$$

Notice that the overall priorities for Susan, John and Michelle are the same as in (9). John is still 1.02 times better than Susan and 1.37 times better than Michelle. There is no rank reversal with the benchmark measurement and there is no change in relative composite ratios. Notice also that our benchmark priorities add up to one, and the non-benchmark alternative, Bernard, is in the same commensurate ratios as the benchmarks. If we wished, we could make a proportional transformation of the final results so that all priorities, (benchmarks and non-benchmarks) add up to one.

Next, to see the power of benchmark measurement to handle many alternatives, we add three more hypothetical candidates called Bill, Bert, and Eng. Their inclusion in the paired comparison process for the first criterion is depicted in Figure 2.

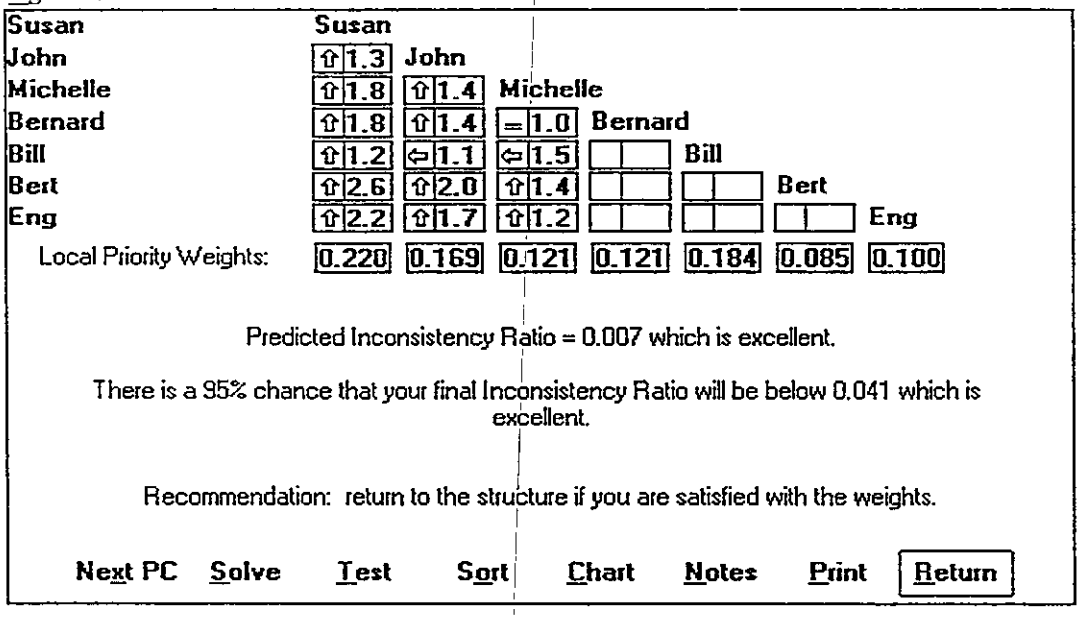

Figure 2 -- Benchmark Comparisons with Four Non-Benchmark Alternatives

The first thing to notice about Figure 2 is that all comparisons are in the first three columns. This is because we only make comparisons to and between the benchmark alternatives. This means that if more alternatives are added, the number of additional comparisons for each new alternative will equal the number of benchmarks. In our example, just three additional comparisons for each new alternative is not an arduous task. It is slightly more time consuming than absolute ratings, but it is probably more precise.

In the comparisons of Figure 2, we maintained the consistent comparisons from Figure 1, and we tried to be as consistent as possible on the remaining comparisons. This was done to show that the incomplete matrix procedure produces the correct ratios. Notice that except for rounding errors, the ratios between Susan and John, Susan and Michelle, and John and Michelle remain at 1.29, 1.8 and 1.4 in both Figures 1 and 2.

Had inconsistency been allowed for the non-benchmark comparisons, then it would have been possible for the benchmark priorities to be different from Figure 2. Accordingly, the results from an incomplete matrix that has some inconsistency can cause the base to change. As the number of non benchmarks increases in the incomplete matrix, the possibility of skewing the original benchmark ratios goes up. To avoid this influence, it may be best to compare new alternatives one at a time to the benchmarks (as in Figure 1) rather than as a group in a larger incomplete matrix (as in Figure 2). The results would then measure the individual association of each new alternative to the benchmark alternatives. This process would be used if we want to look at alternatives one at a time.

We would consider the enlarged incomplete matrix if the decision situation required us to look at the alternatives as a group. Inconsistency could then have a greater effect on deflecting priorities from the base case, but we would have the assurance that we are looking at only the joint effect of comparisons of all alternatives to the benchmarks. There would be no influence of non-benchmarks to each other. Provided the inconsistency is not severe, the departure from the base case would be minor. In any event, we should recalibrate benchmarks from time to time, but only as they are compared to each other or if we change the benchmark set.

Using AHP Tree with consistent comparisons for all of the criteria yields the following local weights.

$$
\begin{array}{c} \\ \textit{Susan} \\ \textit{John} \\ \textit{Michelle} \\ \textit{Bernard} \\ \textit{Bill} \\ \textit{Bert} \\ \textit{Eng} \end{array}
\begin{array}{c}
\begin{array}{ccccccc} \text{C1} & \text{C2} & \text{C3} & \text{C4} & \text{C5} & \text{C6} & \text{C7} \end{array} \\
\begin{bmatrix}
.220 & .190 & .122 & .202 & .202 & .205 & .167 \\
.169 & .135 & .216 & .154 & .155 & .160 & .119 \\
.121 & .135 & .122 & .126 & .155 & .135 & .142 \\
.121 & .054 & .169 & .051 & .045 & .045 & .048 \\
.184 & .161 & .123 & .075 & .155 & .160 & .261 \\
.085 & .189 & .149 & .171 & .103 & .135 & .154 \\
.100 & .135 & .100 & .221 & .185 & .160 & .109
\end{bmatrix}
\end{array}
= \hat{A}\hat{N} \qquad (13)
$$

Again, if we treat (13) as a proportional transformation of $\hat{A}$, then our rescaled S_{Bjj} as defined in (5) is the inverse of the first three rows of (13) (i. e. our benchmark rows) Now, S_{Bjj} becomes:

$$
\begin{bmatrix}
\frac{1}{.510} & & & & & & \\
& \frac{1}{.460} & & & & & \\
& & \frac{1}{.460} & & & & \\
& & & \frac{1}{.482} & & & \\
& & & & \frac{1}{.512} & & \\
& & & & & \frac{1}{.500} & \\
& & & & & & \frac{1}{.428}
\end{bmatrix}
= \hat{N}^{-1} S_B = \hat{S}_B \qquad (14)
$$

With Bernard and the three new candidates added to the baseline group, the new composite weights are:
$W = (\hat{A}\hat{N})\ \hat{N}^{-1} S_B\ b$

$$
\begin{array}{c}
(\hat{A}\hat{N}) \\
\begin{array}{ccccccc} \text{C1} & \text{C2} & \text{C3} & \text{C4} & \text{C5} & \text{C6} & \text{C7} \end{array}
\end{array}
\begin{bmatrix}
.220 & .190 & .122 & .202 & .202 & .205 & .167 \\
.169 & .135 & .216 & .154 & .155 & .160 & .119 \\
.121 & .135 & .122 & .126 & .155 & .135 & .142 \\
.121 & .054 & .169 & .051 & .045 & .045 & .048 \\
.184 & .161 & .123 & .075 & .155 & .160 & .261 \\
.085 & .189 & .149 & .171 & .103 & .135 & .154 \\
.100 & .135 & .100 & .221 & .185 & .160 & .109
\end{bmatrix}
\cdot
\begin{array}{c}
\hat{N}^{-1} S_B \\
\begin{bmatrix}
1.96 & & & & & & \\
& 2.17 & & & & & \\
& & 2.17 & & & & \\
& & & 2.07 & & & \\
& & & & 1.95 & & \\
& & & & & 2.00 & \\
& & & & & & 2.34
\end{bmatrix}
\end{array}
\cdot
\begin{array}{c}
b \\
\begin{bmatrix} .172 \\ .060 \\ .344 \\ .211 \\ .130 \\ .045 \\ .038 \end{bmatrix}
\end{array}
=
\begin{array}{cc}
\begin{bmatrix} .363 \\ .368 \\ .269 \\ .216 \\ .285 \\ .292 \\ .294 \end{bmatrix} &
\begin{array}{l} \textit{Susan} \\ \textit{John} \\ \textit{Michelle} \\ \textit{Bernard} \\ \textit{Bill} \\ \textit{Bertl} \\ \textit{Eng} \end{array}
\end{array}
\qquad (15)
$$

Except for rounding errors, Susan, John and Michelle maintain their benchmark weights, Bernard keeps his former weight, and Bill, Bert and Eng, like Bernard, receive weights proportional to the benchmarks.

Discussion

Table 1 is helpful for comparing between relative, benchmark and absolute measurement. As can be seen from the table, the benchmark mode is more similar to absolute measurement than relative measurement. The one thing it has in common with relative measurement is direct paired comparisons of the actual alternatives. There is no intermediary rating by intensities nor is there any need for firm standards as required in absolute measurement. Yet, like absolute measurement, the benchmark procedure can accommodate a very large number of alternatives and they do not all have to be present at one time. Rather than rate which intensity an alternative has, benchmark measurement requires paired comparisons to the benchmark items. Although this is likely to take a little longer than absolute ratings, it is quicker than the complete comparisons of relative measurement.

Table 1

Comparison Between Relative, Benchmark, and Absolute Measurement

	Relative Measurement	Benchmark Measurement	Absolute Measurement
Type of comparisons	relative to each other alternative	relative to benchmark alternatives	relative to absolute standards, then rating to those standards
Evaluation of alternatives	All relevant alternatives compared to each other at one time.	Two steps: (1) comparison of benchmark alternatives (2) comparing decision alternative to benchmark alternatives	Two stages: (1) comparison of absolute intensities, (2) rating alternatives for their intensities
Time perspective	now -- all alternatives analyzed at one time	future -- alternatives can be analyzed at different times	future -- alternatives can be analyzed at different times
Potential number of alternatives	small, usually <10	large	large
Effort to analyze numerous alternatives	Large	Moderate	Small
Nature of Criteria Priorities	treated independent of alternatives (distributive and ideal) or dependent on alternatives (linking pins or referenced)	dependent upon benchmark alternatives and modified by subsequent alternatives	independent of alternatives
Normalization Procedure	All alternatives sum to one	Only benchmark alternatives sum to one; other alternatives are in ratio form to the benchmarks	Indicators (intensities) sum to one.
Effect of Adding or Deleting Alternatives	Allows rank to reverse and composite ratios to change(distributive and ideal)	Maintains rank and composite ratios	Maintains rank and composite ratios
Items at bottom level of hierarchy	Relevant alternatives	Benchmark alternatives	Intensities (indicators) of the alternatives

By making criteria comparisons dependent upon the alternatives, benchmark measurement is akin to linking pin and referenced AHP modes of relative measurement, at least as far as the synthesis process is concerned. When the benchmark is a single alternative, benchmark synthesis is the same as linking pins which can use any alternative as a link between hierarchical levels. With multiple benchmark alternatives, the synthesis mode is the same as referenced AHP. Although we have displayed benchmarkmeasurement with referenced criteria, we could have done the same thing with linking pins (Schoner, Wedley and Choo, 1993).

There are, however, three important distinctions between benchmark measurement and linking pin and referenced AHP. The first relates to the generation of local priorities. Benchmark measurement, as its name implies, only makes comparisons to benchmark alternatives. The referenced and linking pin modes make comparisons to all alternatives. Thus, there is a limit to the number of alternatives that linking pins and referenced AHP can consider. No such constraint is enforced with benchmark measurement.

The second distinction relates to the way that linking pins and referenced AHP generate priorities for the upper levels of the hierarchy. Linking pins makes reference to a specific alternative when making criteria comparisons, referenced AHP refers to the average or total amount of the criterion possessed by the alternatives, and so does benchmark measurement. When an alternative is added or deleted in referenced AHP, we require the user to go back and modify criteria weights (since the average or total has changed). In benchmark measurement and linking pins, we do not attempt to rescale the criteria, but rather rescale the local priorities so that the criteria benchmarks remain unchanged.

The third distinction is that benchmark measurement is conducted in two stages. In the first stage, comparisons are done for a hierarchy which contains benchmark alternatives, benchmark criteria, and benchmark comparisons. The priorities derived in this first step become the benchmarks for the second stage when additional alternatives are

evaluated. Thus in Figures 1 and 2, the benchmark comparisons come from the first stage. They are not repeated each time a new alternative is added. Referenced and linking pins, like the distributed and ideal modes, are conducted in only one step. All relevant alternatives are evaluated at the same time.

We should note that although benchmark and absolute measurement are conducted in two stages, there is a subtle difference between how the stages are carried out. In absolute measurement, the rating procedure is done separate from the hierarchy which was used to generate intensities. In benchmark measurement, on the other hand, the second stage comparisons are still related to the baseline hierarchy.

One potential deficiency of benchmark measurement is the possibility that the original benchmark priorities deviate from their base values when there is inconsistency in second phase comparisons. One way to overcome this problem is to ignore any change in the original benchmark values while accepting new alternatives at their derived values. This would mean that the incomplete eigenvector solution is the value for just non-benchmark alternatives. Another approach is to compare new alternatives to the benchmarks without making use of the incomplete matrix procedure. Such a process would keep the original benchmark priorities invariant while still producing commensurate priorities for new alternatives.

Using the benchmark priorities for C1 and C2 in (7) and the comparisons for Bernard in Figure 1, we can illustrate this latter process.

	Benchmark Priority	Comparisons for Bernard	Imputed Priority	Average Priority
C1 Dependability				
Susan	.429	1/1.8	.238	
John	.333	1/1.4	.238	.238
Michelle.238	1/1	.238		
C2 Education				
Susan	.412	1/3.7	.111	
John	.294	1/2.4	.122	.117
Michelle.294	1/2.5	.118		

Here, we have three estimates (imputed priority) for each criterion. For the C1 criterion, the three estimates are equal (.238) because they are based upon consistent comparisons. For C2, the comparisons to benchmarks are inconsistent and the resulting imputed priorities are different. Since all three comparisons are of equal importance, we can take the arithmetic average (.117) as our estimate of the Bernard's priority in commensurate terms.

Where the imputed priority indicates a lack of consistency, it is possible to determine which of the non-benchmark comparisons is the most inconsistent. The quickest method is to take the ratios of the derived priorities to the benchmark priorities. This yields an imputed comparison which can be compared to the actual comparison. Another approach is to use the triad relationship $a_{ij} * a_{jk} = a_{ik}$ to determine the absolute deviation from perfect consistency If perfect consistency exists, then $(a_{ij} * a_{jk})/a_{ik} = 1$); values above or less than 1 represent departures from consistency.

In passing, we should note that these measures of inconsistency are different from the consistency ratio of conventional AHP. That measure is based upon the largest positive eigenvector of a positive reciprocal matrix. It is not appropriate for our purposes, because it is affected by benchmark and non-benchmarks alike. We are only interested in the inconsistency which may exist between the benchmark alternatives and a newly added alternative. Since new additions are not compared to each other and are independent of each other, they should not affect each other in a comparison matrix. Similarly, the within benchmark comparisons have an effect on the entire matrix and should be excluded re their effect on non-benchmark consistency. For benchmark comparison, it is more desirable to have a measure which identifies which comparison to benchmarks is most inconsistent and what the overall inconsistency is for a new alternative.

Finally, it should be noted that it is acceptable to use entirely different individuals and a different number of benchmark alternatives under each of the criteria so long as the criteria weights are calibrated to those referents. For C3, for example, it may have been more appropriate to use Ruth, Tom, Mickey, and Kathryn. Rather than using a common set of alternatives across criteria, it would be better to use well understood alternatives which enable us to capture the variability within a criterion. Just as the intensities across a criterion capture absolute values, so too must our benchmark alternatives capture the range of a criterion. Using near copies for our benchmarks would not make sense, because similar items would not encompass the potential variability.

Conclusion.

The benchmark procedure proposed in this paper is a new mode of measurement for the AHP. It maintains the precision of paired comparisons from relative measurement while providing the flexibility of easily handling large numbers of alternatives. It can be used with equal effectiveness in either a referenced or linking pin mode.

Benchmark measurement is like two-phased referenced AHP. In the first stage, comparisons amongst benchmarks allow us to make a ruler. In the second stage, we use that derived ruler to compare new alternatives. We have faith in the ruler, and because of that, employ it to measure lengths (attributes) of objects about whose lengths (attributes) we are uncertain. Similarly, the benchmarks set should consist of alternatives which have been calibrated and (periodically) recalibrated. We have faith in their relative measurements and will not readjust them on the basis of any inconsistencies in judgements made with respect to new alternatives about which we know relatively little. The resulting priorities for the new alternatives are in commensurate terms with the original benchmarks.

This new measurement mode has many advantages. Like absolute measurement, it will never have rank reversals, New alternatives can be evaluated at different times, in commensurate terms, and with greater precision. Moreover, the ability to use a template hierarchy overcomes many of the difficulties Verkasalo (1994) identified with having to derive new hierarchies for similar problems.

Although the concept of benchmark measurement is in its infancy, it holds out promise as a replacement for absolute measurement. The time to make comparisons to benchmarks may be slightly longer than the time to rate alternatives, but the greater precision is likely to be worth the effort. If speed rather than precision is desired, then it is possible to compare a new alternative to a subset of the benchmarks, even a subset as small as one.

Finally, benchmark measurement is more flexible. It can be used in cases where absolute measures are available, where a combination of absolute and relative measurement is available, and where no fixed standards are available. It is a more flexible and precise technique which should be added our list of decision making techniques.

References

Belton, V. and Gear, A. E. (1983) On a shortcoming of Saaty's method of analytic hierarchies," *Omega*, 11, 228-230.

Forman, Ernest H. (1994) "Multicriteria prioritization in open and closed systems", George Washington University, Washington, D. C. (forthcoming)

Harker, Patrick T. (1987). The Incomplete Pairwise Comparisons in the Analytic Hierarchy Process, *Mathematical Modelling*, 9, 837-848.

Saaty, T. L. (1993) "Rank Preservation and Reversal: the AHP Ideal and Distributive Modes," *Mathematical and Computer Modelling*, 17, 6, 1-16.

Saaty, T. L (1994) *Fundamentals of Decision Making and Priority Theory*, RWS Publication, Pittsburgh, PA.

Schoner, Bertram and Wedley, William C. (1989). Ambiguous Criteria Weights in AHP: Consequences and Solutions, *Decision Sciences*, 20, 3, 462-475.

Schoner, B., Wedley, W. C. and Choo E. U. (1993) "A unified approach to AHP with linking pins" *European Journal of Operations Research*, 64, 384-392.

Verkasalo, Matti (1994) "Repetitive Use of the AHP Hierarchy" *Proceedings of the 3rd International Symposium on the Analytic Hierarchy Process*, School of Business and Public Management, The George Washington University, Washington, D.C., 89-102