Ann. Data. Sci. (2014) 1(3–4):359–368
DOI 10.1007/s40745-014-0026-4

How Many Judges Should There Be in a Group ?

Thomas L. Saaty · Mujgan Sağır Özdemir

Received: 30 September 2014 / Revised: 1 November 2014 / Accepted: 10 December 2014 /
Published online: 11 January 2015

Abstract This paper briefly examines the question of how many judges are needed to obtain valid and consistent judgments when using the analytic hierarchy process. It turns out that if a judge is experienced and well versed in an area, he can be sufficient to provide the judgments instead of diluting his accuracy with the participation of others who may not be as good. How to discover such a person requires criteria used to judge his adequacy and that of others.

1 Introduction

We are often asked the question: What sample size of judges is best to provide the judgments when using the analytic hierarchy process (AHP)? Frequently students ask the question at universities where the advisor is a statistician. We thought it may be useful to attempt to answer this question in a short note with the hope that others who are more expert in gathering and interpreting public opinion may help deepen the query.

2 Observations

There are many criteria to consider in determining the size of a statistical sample. They can be found in the statistical literature and rely, among other thing, on the

T. L. Saaty (✉)
University of Pittsburgh, Pittsburgh, USA
e-mail: saaty@katz.pitt.edu

M. S. Özdemir
Eskisehir Osmangazi University, Eskisehir, Turkey
e-mail: mujgan.sagir@gmail.com

underlying probability distributions, the prescribed confidence levels and on the size of the population sampled.

It is very different when one collects judgments in the AHP. Generally, AHP applications are concerned with three different ways to frame the pairwise comparison questions. The first is to ask which of the pair of elements is more dominant or important with respect to an attribute or criterion, the second is to ask which is the more likely outcome as in the presidential elections, and the third is to ask which element is preferred with respect to the attribute, recognizing that preference is entirely subjective and depends on the whims, and likes or dislikes of the individual. We believe that the preference question can be answered by sampling as is done in statistics, and any judge can be free to express his or her preference. Validation with respect to what can happen out there is of no consequence in preference choices. Answering both importance and likelihood questions requires what is known as expert knowledge in the subject in which the decision is made.

Two factors affect the sample size of judges required to make the comparisons in the case of importance or likelihood questions. The first is the consistency of the judgments and the second their validity in practice. There are two further kinds of situations to consider. One is whether the judges must agree among themselves; for example a jury must unanimously declare a verdict of guilty or not guilty and this necessarily involves interdependence and agreement. The other is when the judgments involve judges who are independently providing their judgments and may be at large in the population and unable to carry out discussions with each other. There is a bit of a dilemma here as we shall show that the need for consistency limits the number to not more than 7 or 8 judges. What is particularly useful in the AHP is that the judges themselves can be assigned priorities that make the judgments of a high priority judge count more than those with lower priority. This is done by raising their individual judgments to the power or the respective judges' priorities, then taking the geometric mean, thus extending more weight to those judges that are believed to have more expert knowledge. It is done not according to sample size as in taking statistics about preference, but according to how much and how well they know the subject, based on some criteria such as education, years of experience, and level of attainment in society.

The usefulness and validity of judgments depend much on how richly or sparsely the problem is structured. It determines how valuable or lacking a judge may be in including the important criteria necessary to determine the outcome. If he is unable to structure the decision in sufficient detail to match what is in known in the law books about past similar cases, his knowledge may be lacking and potentially faulty.

The usefulness and validity of the pairwise comparison judgments in making a decision are greater the finer the subcriteria used to compare the alternatives are. That is because as the scope of the question narrows it is easier for people to make comparison judgments and ensure their accuracy. The following example, done many years ago with members of the department of the Interior in Washington, shows the structure a group arrived at that were trying to decide at what level: half full or full, a dam should be kept. The figures below show the gradual evolution of the model from a simple three level hierarchy shown in Fig. 1 to the four level hierarchy shown in Fig. 2 that includes decision makers, to the five level hierarchy shown in Fig. 3 which includes interests of the decision makers, to the final six level hierarchy shown

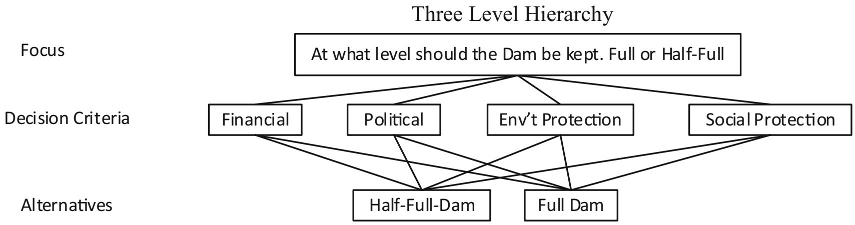

Fig. 1 Starting three level model

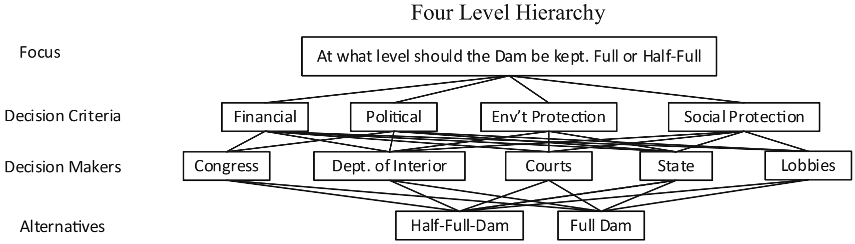

Fig. 2 Four level hierarchy including decision makers

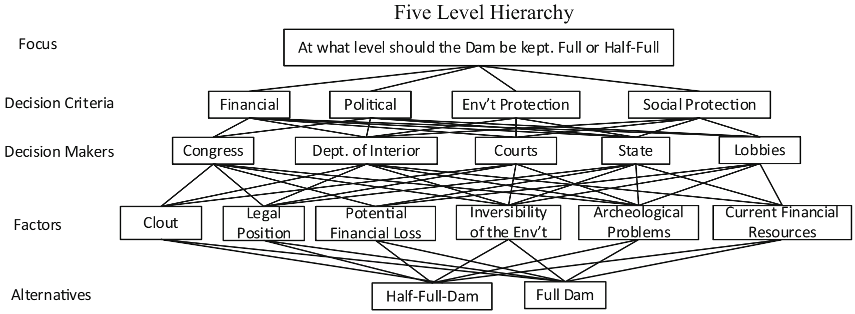

Fig. 3 Five level hierarchy including factors that distinguish the expertise of the decision makers

in Fig. 4 that includes people affected by the decision and the final figure is a structure that has been elaborated to frame the entire story.

This three level hierarchy involves comparing a practical third level of how full a dam should be with very general and abstract human concerns, a process that is very hard to do whether by assigning numbers directly or by making comparisons. Again, let us elaborate this hierarchy in a realistic way by attaching a level of decision makers.

This four level hierarchy involves comparing a practical fourth level of how full a dam should be with human institutions whose preferences may not touch at all at the practical side of the decision. Again, we elaborate with and additional level of influence factors of these institutions and find that they are too general to resolve the problem.

The five level hierarchy in Fig. 3 in turn shows the difficulty of comparing the alternatives with respect to the factors which represent the actors influence on the groups affected which fall in the next level.

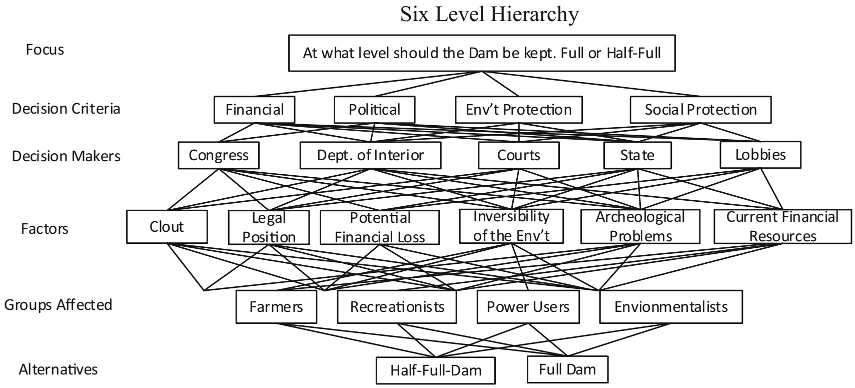

Fig. 4 Six level hierarchy that includes the people affected by the decision

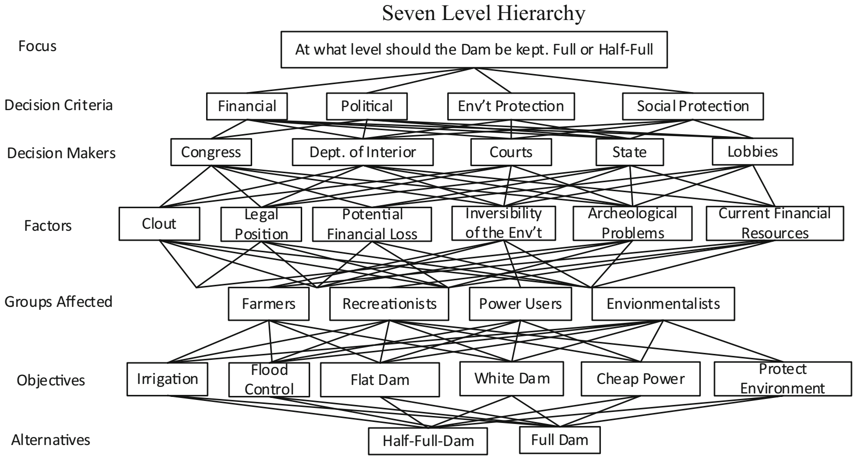

Fig. 5 Final hierarchy containing 7 levels; objectives of each of the affected groups have been added

In the six level hierarchy shown in Fig. 4 the names of the groups affected by the influences of the decision makers are included as factors in the in the fifth level but again though that hierarchy more accurately represents the situation and we are getting closer, we need to spell out what each member of this group is trying to achieve in advocating a certain level for the dam, so finally we add a level of factors in the sixth level as shown in Fig. 5 that represent the needs of the groups affected which makes it much easier to compare the alternatives and arrive at a synthesized best outcome that takes everyone's position into consideration.

At this point we interject a technical note. Figure 5 is a hierarchy, not a tree, and as such it is a compact structure that makes it possible to include all the issues in the problem in a minimum amount of space. It is quite difficult to turn a seven level hierarchy like this one into a tree, and if you did do that it would become unwieldy and hard to understand. All the data from the pairwise comparisons of the elements in the problem are assembled into a single data structure, the supermatrix, and raising it to powers captures all the interactions so that when it converges it nets out the overall priorities of the alternatives at the bottom of the structure, and incidentally, prioritizes all the other elements in the model at the same time.

3 Aggregating the Judgments of the Individuals in a Group

There are three papers that have been written about combining the numerical judgments of individuals into a representative group judgment. The first is by [2] which proves that the geometric mean is the way to combine individual judgments into a group judgment. The second is by [1] about how to combine the judgments of judges, who have different expertise, and thus whose judgments have different priority, and the third and more recent is by [5]. K. Arrow [3] proved the impossibility of deriving a representative group judgment if each individual only declares he prefers or does not prefer A to B. But in their paper Saaty and Vargas showed that by using the AHP fundamental scale, a scale of absolute cardinal numbers, and learning *how much* each individual prefers A to B, it is possible to carry out pairwise comparisons and derive a representative group judgment from the individual judgments that satisfies Arrow's four conditions. Below is a brief summary of their results. This material is included here to show the need for using an appropriate scale to make judgments in order to obtain a useful synthesis and it also indicates the kind of sharing needed in order to find a representative group judgment. This could have a bearing on the type of individuals to seek out when making decisions that involve several people's judgments.

3.1 Arrow's Four Conditions

Let the function $f(x_1, x_2, \ldots, x_n)$ for synthesizing the judgments given by n judges, satisfy the

(i) Separability condition (S): $f(x_1, x_2, \ldots, x_n) = g(x_1)g(x_2)\ldots g(x_n)$ for all $\mathrm{x}_1, \mathrm{x}_2, \ldots, \mathrm{x}_\mathrm{n}$ in an interval P of positive numbers, where g is a function mapping P onto a proper interval J and is a continuous, associative and cancellative operation [(S) means that the influences of the individual judgments can be separated as above].

(ii) *Unanimity condition* (U): $f(x, x, \ldots, x) = x$ for all x in a proper internal of positive numbers P [(U) means that if all individuals give the same judgment x, that judgment should also be the synthesized judgment].

(iii) *Homogeneity condition* (H): $f(ux_1, ux_2, \ldots, ux_n) = uf(x_1, x_2, \ldots, x_n)$ where $u > 0$ and x_k, ux_k $(k = 1, 2, \ldots, n)$ are all in P [For ratio judgments (H) means that if all individuals judge a ratio u times as large as another ratio, then the synthesized judgment should also be u times as large].

(iv) *Power conditions* (P_p): $f\left(x_1^p, x_2^p, \ldots, x_n^p\right) = f^p(x_1, x_2, \ldots, x_n)$. Special case $(R = P_{-1}) : f(1/x_1, 1/x_2, \ldots, 1/x_n) = 1/f(x_1, x_2, \ldots, x_n)$, the reciprocal property [(R) is of particular importance in making judgments with the AHP fundamental scale. It means that the synthesized value of the reciprocal of the individual judgments should be the reciprocal of the synthesized value of the original judgments].

Aczel and Saaty proved the following theorem:

Theorem *The general separable (S) synthesizing functions satisfying the unanimity (U) and homogeneity (H) conditions are the geometric mean and the root-mean-power. If, moreover, the reciprocal property (R) is assumed even for a single n-tuple $(x_1, x_2, \ldots, x_n)$ of the judgments of n individuals, where not all x_k are equal, then only the geometric mean satisfies all the above conditions.*

3.2 Prioritizing the Judges Themselves

In any rational consensus, those who know more should, accordingly, influence the outcome more strongly than those who are less knowledgeable. Some people are clearly wiser and more sensible in such matters than others, others may be more powerful and their opinions should be given appropriately greater weight. For such unequal importance of voters not all g's in (S) are the same function. In place of (S), the weighted separability property (WS) is now: $f(\mathrm{x}_1, \mathrm{x}_2, \ldots, \mathrm{x}_\mathrm{n}) = g_1(x_1)g_2(x_2)\ldots g_n(x_n)$ [(WS) implies that not all judging individuals have the same weight when the judgments are synthesized and the different influences are reflected in the different functions $(g_1, g_2, ..., g_n)$].

In this situation, Aczel and Alsina proved:

Theorem *The general weighted-separable (WS) synthesizing functions with the unanimity (U) and homogeneity (H) properties are the weighted geometric mean $f(x_1, x_2, \ldots, x_n) = x_1^{q1} x_2^{q2} \ldots x_n^{qn}$ and the weighted root-mean-powers $f(x_1, x_2, \ldots, x_n) = \sqrt[\gamma]{q_1 x_1^\gamma + q_2 x_2^\gamma \ldots + q_n x_n^\gamma}$, where $q_1 + q_2 + \ldots + q_n = 1, q_k > 0 (k = 1, 2, ..., n), \gamma > 0$, but otherwise $q_1, q_2, \ldots, q_n, \gamma$ are arbitrary constants.*

If f also has the reciprocal property (R) and for a single set of entries $(x_1, x_2, \ldots, x_n)$ of judgments of n individuals, where not all x_k are equal, then *only the weighted geometric mean* applies.

We give the following theorem which is an explicit statement of the synthesis problem that follows from the previous results:

Theorem *If $x_1^{(i)}, \ldots, x_n^{(i)}$ $i = 1, ..., m$ are rankings of n alternatives by m independent judges and if a_i is the importance of judge i developed from a hierarchy for evaluating the judges, and hence $\sum_{i=1}^{m} a_i = 1$, then*

$$\begin{pmatrix} m \\ \Pi \quad x_1^{a_1} \\ i = 1 \end{pmatrix}, \ldots, \begin{pmatrix} m \\ \Pi \quad x_n^{a_1} \\ i = 1 \end{pmatrix}$$

are the combined ranks of the alternatives for the m judges.

The power or priority of judge i is simply a replication of the judgment of that judge (as if there are as many other judges as indicated by his/her power a_i), which implies multiplying his/her ratio by itself a_i times, and the result follows.

3.3 Discussion Preliminary to the Theorem of Saaty and Vargas

Given a group of individuals, a set of alternatives (more than two), and individual ordinal preferences for the alternatives, Arrow's Impossibility Theorem proved that it is impossible to derive a rational group choice (that is, construct a social choice function that aggregates individual preferences) from ordinal preferences of the individuals that satisfy the following four conditions, i.e., at least one of them will be violated:

Decisiveness The aggregation procedure must generally produce a group order.

Unanimity If all individuals prefer alternative A to alternative B, then the aggregation procedure must produce a group order indicating that the group prefers A to B.
Independence of irrelevant alternatives Given two sets of alternatives which both include A and B, if all individuals prefer A to B in both sets, then the aggregation procedure must produce a group order indicating that the group, given any of the two sets of alternatives, prefers A to B.

No dictator No single individual preferences determine the group order.

Theorem *The geometric mean aggregation procedure $f : \mathfrak{P}^m \rightarrow \mathfrak{P}$ satisfies (i) unrestricted domain, (ii) pairwise unanimity, (iii) pairwise cardinal independence from irrelevant alternatives and (iv) pairwise cardinal non-dictatorship.*

Thus, the geometric mean gives rise to an aggregate (*social*) reciprocal pairwise relation that satisfies all four of Arrow's conditions. Here we assumed that all members of the group have the same importance. However, when the individuals have different importance, [1] extended the foregoing [2] result.

4 The Qualities Needed for a Good Judge

A country without a sea shore is unlikely to have anybody who knows about building ships no matter how large its population may be. On the other hand a country with a sea shore, no matter how small in population, is likely to have many people who know about ships and seal hunting. Thus the size of a population is not always a good determinant of how many people should vote on a certain matter.

Consider the case of a most popular person in the whole world who commits a murder. He is the only one who knows it and no one else can ever think that he could commit such an act. One and only one person knows about the crime and that is the murderer himself. There is no point in asking anyone else what their judgment is about who committed the murder. Here the number of judges does not matter. This is a counter example to any theorem that would prove that the number of judges is an important factor in collecting and synthesizing a group judgment. At least in this case, it is clear that number is not important.

Recall that the consistency of a homogenous group is not more than the inconsistency of the most inconsistent individuals of the group. In general, we have shown recently [6] that the number of people involved in a decision, should be no more than seven. If there are several areas of specialization in which experienced judgment is needed, then each aspect may be judged by a group of seven people who are specialists in that area, and work independently of the other groups on judgments involving their specialty. Other groups will be providing judgments in their area of specialization. In the end, their judgments lead to the derivation of priorities and the final outcome is obtained by a process of weighting the criteria themselves and adding the weighted priorities thus yielding an overall composite priority outcome using the geometric mean when needed to combine individual or different group judgments

Interestingly we discovered practical research that confirms our conjecture. In referring to a part of their research and statistical findings, [4] write that "the most important aspect is the point at which the weighted sum of errors is least. This point is reached at a jury size somewhere **between six and eight; the nearest whole number is seven.** The model therefore predicts that a jury of seven members will minimize errors in the fashion we assume would be optimum. Subject to the limitations on the coin-flipping model, we can refer to a seven-member jury as the optimum jury size for unanimous juries".

The function of a jury differs from that of a committee. The jury must agree on a judgment: guilty, not guilty, and undecided resulting in a hung jury. A committee tries to reach a consensus but need not agree unanimously. The members may be well-informed but may not agree on their conclusions. Thus if it is meaningful to compare jurors according to their degree of understanding, then their total number must be equal to about 7 in order that they be better able to identify the most inconsistent juror among them. They would then try to change his mind step by step to produce a more consistent comparison matrix with which the jurors are prioritized. Subsequently, they would then try to improve the inconsistency of the most inconsistent member with the revised judgments and so on.

Assume that there is an outside supreme judge that is able to compare the jurors on a jury according to their degree of understanding. Then according to what Saaty has found, if the number of jurors is around 7 it is easiest to identify the most inconsistent juror among them. (assume the outside supreme judge tells them who it is) and the other jurors can work to gradually change the mind of that juror to make him more consistent. Of course that does not necessarily mean he will then be in agreement or compatible with them, no matter how consistent he becomes, and the result would be a hung jury.

5 Core of the Matter

We will summarize here what we have learned about number. We may say that number is not necessarily a reliable measure of wisdom and sanity.

- Experience means that the judge has been successfully engaged for many years in the field that he is making the judgments about. Success is measured against commonly accepted standard of the constituency of the judge.
- Is each judge consistent enough with his own individual perceptions of the decision situation?
- Does the judge have a diversified awareness?
- How much opposition is there against that judge?
- One always has to balance the wisdom of crowds against the stupidity of mobs.

- Compute how compatible each individual member's final priority vector is when compared to that of the group. Number becomes important when it is necessary to have more than one judge, and the group must reach some kind of agreement. Agreement may be defined differently: unanimous, simple majority, 2/3 majority or some other fraction.

6 Conclusions

To engage judges to help with a decision should not be a random matter. One needs to know the area of expertise needed to make that decision and select a judge or judges that have both knowledge and practical experience with the matter. In this case *one expert judge* may suffice unless political expediency requires that several judges from different constituencies are necessary. In that case one might select several judges if they are available.

References

1. Aczél J, Alsina C (1987) Synthesizing judgements: a functional equations approach. Math Model 9:311–320
2. Aczél J, Saaty TL (1983) Procedures for synthesizing ratio judgements. J Math Psychol 27:93–102
3. Arrow KL (1963) Social choice and industrial values. Yale University Press, New Haven
4. Nagel SS, Neef M (1975) Deductive modeling to determine an optimum jury size and fraction required to convict. Wash Univ Law Rev 1975(4):933–978
5. Saaty TL, Vargas LG (2012) Possibility of group choice: pairwise comparisons and merging functions. Soc Choice Welf 38(3):481–496
6. Saaty TL (2014) The Natural Law of Structured Cooperation, to appear

Thomas L. Saaty Chair of Distinguished University Professor, University of Pittsburgh; member of National Academy of Engineering; 2012, Herbert Simon Award for best paper in 10 years; 2011 received the distinguished award Doctoris Honoris Causa in a ceremony at Poland's oldest and most prestigious Jagiellonian University, Krakow, Poland; 2008 awarded the Impact Prize by INFORMS for his work on the Analytic Hierarchy Process; 2000; awarded the gold medal from the International Society for Multicriteria Decision Making; 1977 award from the Institute of Management Sciences for one of the best applied studies of that year: The Sudan Transport Plan; 1972 the L.R. Ford prize in mathematics for a comprehensive paper "Thirteen Colorful Variations on Guthrie's Four Color Conjecture". Previously, professor, the Wharton School, University of Pennsylvania; Scientific Analyst at the Arms Control and Disarmament Agency in the Department of State, Washington: nuclear arms reduction negotiations with the Soviets in Geneva. He developed the Analytic Hierarchy Process (AHP) for decision-making and its generalization to feedback, the Analytic Network Process (ANP). Published 43 books and more than 300 papers; latest books *Group Decision Making: Drawing out and Reconciling Differences*: *Problem Solving & Decision Making; The Brain: Unraveling the Mystery of How It Works; Creative Thinking and Problem Solving; The Neural Network Process* (*NNP*).

Mujgan Sağır Özdemir holds her bachelors and masters degrees from the Anadolu University and the doctoral degree from Eskisehir Osmangazi University, in Industrial Engineering and Operations Research. Her research interest lies in timetabling problems (course scheduling), cutting and assortment problems, vehicle routing, quantitative and qualitative decision making. Her articles have appeared in journals such as the European Journal of Operational Research, Applied Mathematics and Computation, Mathematical and Computer Modeling, International Journal of Information Technology and Decision Making. She teaches courses on linear and nonlinear optimization, multiobjective decision making, creativity and mathematical programming.