USAGE OF SCALES WITH DIFFERENT NUMBER OF GRADES FOR PAIR COMPARISONS IN DECISION SUPPORT SYSTEMS

V.V.Tsyganok
E-mail: vitaliy.tsyganok@gmail.com

S.V.Kadenko*
E-mail: seriga2009@gmail.com

O.V.Andriichuk
E-mail: andreychuck@ukr.net

National Academy of Sciences of Ukraine
The Institute for Information Recording
Kyiv, Ukraine

ABSTRACT

In this paper we suggest an original approach to conducting individual pair comparisons during individual and group multi-criteria decision-making (including AHP/ANP-based decisions). With this approach every expert is given an opportunity to use the scale, in the degree of detail (number of points/grades) that most adequately reflects his/her competence in the issue under consideration for every single pair comparison. Before aggregation all separate expert estimates (judgments) are brought to a unified scale, and scales in which these judgments were built are assigned respective weights. A respective instrument for pair comparison conduction has been developed, and an experiment has been organized. The experiment statistically proves that as a result of suggested technology usage, there is an increase in the degree of correspondence between estimates, input by an expert, and his (her) own notions on examination objects.

Keywords: group decision making; decision support system; expert judgments; pairwise comparisons; different scales

1. Introduction

Multi-criteria decision making facilitates solutions for a broad spectrum of problems. A vast amount of research, both theoretical and practical is being conducted to facilitate multi-criteria decision-making in management, environmental protection, production (DeFelice & Petrillo, 2013a,b), logistics (Noorul Haq & Kannan, 2007), (Kannan, Noorul Haq & Sasikumar, 2008), project selection (Zandi & Tavana, 2010) and other areas

* Corresponding author

ISSN 1936-6744
http://dx.doi.org/10.13033/ijahp.v8i1.259

(Kalika & Rossinsky, 2003, Vaidogas & Zavadskas, 2007). Pair comparisons are widely used for multi-criteria decision-making in various weakly-structured domains (i.e., domains, where no benchmarks exist and expert data is the only credible information source). The AHP has a special place among pair comparison-based approaches and related methods, which are utilized in various spheres of human activity.

The practice of expert examination conduction (including AHP-based examinations) indicates that there are certain difficulties that arise when verbal scales are used for expert examination. The expert/decision-maker is often allowed to use only one scale for pair comparisons. In order to get thorough and undistorted data from an expert, (s)he must be offered the opportunity to input estimates in a scale, which most adequately corresponds to his/her competence (awareness) level of the issue under consideration. The suggested research resolves the issue of using verbal scales with a different degree of detail for each particular pair comparison, in order to ensure maximal credibility of knowledge obtained from an expert (expert information must be thorough and undistorted).

To ensure the information obtained from an expert is thorough, we suggest using verbal scales with a sufficient degree of detail: the more points the scale includes, the more information an expert can, potentially, input into a DSS using this particular scale. To avoid information distortion (if an expert is unsure of the degree of dominance between objects in a pair, i.e. (s)he is not competent enough), we suggest giving experts the opportunity to use scales with a low degree of detail, or even allowing them to refuse to estimate preferences in certain object pairs. In our research we also consider an important factor that influences the level of expert information distortion – quantitative equivalent, corresponding to a respective value from a verbal scale. Correspondence between preference value input by an expert and this expert's notions about the ratio of object weights on a pair is an issue of great importance, as it influences the credibility of expert data-based decision-making recommendations.

2. Literature Review

A key study in the described area is the recent research by Elliott (2010) addressing the influence of a chosen quantitative scale upon correspondence between estimation results and an expert's own notions. It was demonstrated that scale selection has considerable impact upon the resulting decision variant estimate. Three quantitative scale types were analyzed, whose point values were assigned to fundamental scale points of two kinds, i.e. scales with 5 and 9 grades (Saaty, 2006). Quantitative scales under consideration included integer, balanced and power scales (Salo & Hamalainen, 1997, Stevens, 1957). Besides that we should mention research done in the context of AHP by Ma & Zheng (1991) and Dodd, Donegan & McMaster (1995).

The most popular and probably the simplest scale is an integer scale where standard linguistic (verbal) values correspond to respective numeric equivalents (from 1 to 9) (see Table 1).

Table 1
Verbal expressions used by experts to determine the preference degrees in alternative pairs in an integer scale

Verbal expression	Numeric equivalent
Equal	1
Weakly or slightly preferred	2
Moderately preferred	3
Moderately plus preferred	4
Strongly preferred	5
“Strongly plus” preferred	6
Very strongly preferred	7
Very, very strongly preferred	8
Extremely preferred	9

When a scale with 5 grades (instead of 9) is used, only 5 verbal expressions corresponding to odd values (1, 3, 5, 7, 9) are utilized.

We should note that phrases presented in Table 1, introduced by Tom Saaty (1980), are often referred to as fundamental scale (or Saaty scale) values. Originally the scale consisted of only 5 grades, so even values (2, 4, 6, 8) and respective verbal expressions correspond to intermediary (transitional) preference degrees. Although it may seem that more elaborate verbal expressions should facilitate more laconic and exact descriptions of preference degrees, the trouble is that these expressions must provide a clear and exact description of the relationship between different preference degrees. For example, a person may clearly imagine a ‘weak’ dominance being weaker than ‘moderate’ and ‘strong’ dominance, but the relationship (ratio) between ‘extreme’ and ‘absolute’ dominance degrees is unclear.

Integer values, corresponding to linguistic phrases show “how many times” one alternative exceeds the other according to a given criterion, i.e. reflect multiplicative preferences. Some researchers noticed that usage of integer scales leads to uneven distribution of alternative weights calculated based on these scales. For instance, a change of preference from ‘weak’ (2) to ‘moderate’ (3) has a larger effect on respective alternative weight, than the change from ‘very very strong’ (8) to ‘extreme’ (9). In order to overcome this drawback, Salo and Hamalainen (1997) suggested a balanced scale where the change of weights remains constant when preferences change.

In the so-called balanced scale the alternative weights, which are calculated based on pair comparisons, are evenly distributed depending on initial pair comparison data. Numeric values, corresponding to verbal expressions are calculated according to the formula: $a=\frac{w}{1-w}$, where w is the weight of alternative, which dominates in the respective pair, as presented in Table 2.

Table 2
Numeric equivalents for balanced scale

Verbal expression	Numeric equivalent
No dominance (equal)	0.5/0.5 = 1
Weak or insignificant dominance	0.55/0.45 = 11/9
Moderate dominance	0.6/0.4 = 3/2
More than moderate dominance	0.65/0.35 = 13/7
Strong dominance	0.7/0.3 = 7/3
More than strong dominance	0.75/0.25 = 3
Very strong dominance	0.8/0.2 = 4
Very-very strong dominance	0.85/0.15 = 17/3
Extreme dominance	0.9/0.1 = 9

It should be stressed that the scale presented in Table 2 is fully balanced only in the case of 2 alternatives.

Another attempt to make a scale whose values more clearly represent the estimator's preferences is a power scale suggested, among other authors, by Stevens (1957) and Lootsmaa (1980, 1991). Numeric values, corresponding to linguistic phrases for power scale, are calculated based on the expression: $a=\sqrt[y-1]{9^{x-1}}$, where x is an integer value from Table 1, corresponding to the same verbal expression, while y is the number of scale grades. For 9 grades the numeric equivalents for power scale are presented in Table 3.

Table 3
Numeric equivalents for power scale

Verbal expression	Corresponding numeric value
No dominance (equal)	$\sqrt[8]{9^0}=1$
Weak or insignificant dominance	$\sqrt[8]{9^1}\approx 1.316$
Moderate dominance	$\sqrt[8]{9^2}\approx 1.732$
More than moderate dominance	$\sqrt[8]{9^3}\approx 2.280$
Strong dominance	$\sqrt[8]{9^4}=3$
More than strong dominance	$\sqrt[8]{9^5}\approx 3.948$
Very strong dominance	$\sqrt[8]{9^6}\approx 5.196$
Very-very strong dominance	$\sqrt[8]{9^7}\approx 6.839$
Extreme dominance	$\sqrt[8]{9^8}=9$

In contrast to a 'balanced' scale, weights obtained based on pair comparisons in a power scale are evenly distributed under any number of alternatives.

Speaking about ‘convenience’ of different scales, we should mention the results obtained by Elliot (2010), particularly an analysis of data on expert’s attitudes toward the proposed estimation scale obtained from 64 experts. The question asked the experts if they thought that the number of preference values to choose from was: a) too large; b) just fine; or c) too small. The expert’s opinions were distributed as shown in Table 4.

Table 4
Percentages of expert answers were as follows

	Too many grades	**Just fine**	**Number of grades is too small**
5 grades	43,8%	53,1%	3,1%
9 grades	84,4%	15,6%	0%

We tend to feel that the conclusion that was made based on data from Table 4 about the advantages of a 5-grade scale can only be relevant for the given group of experts and for a specific expert examination on a specific subject. Another set of conclusions to be made from the research is as follows: 1) the choice of an adequate number of grades in a scale to be used for expert estimation is a topical issue; 2) the fact that opinions of experts concerning the most “comfortable” number of grades in a scale, varies, indicates that offering every single expert a separate scale is better than selecting one scale for all experts to estimate alternatives; 3) for every aspect of examination the expert should be able to choose some scale, which is optimal for this particular issue in terms of the number of grades; 4) the optimal number of grades is not always 5 or 9.

A review and comparison of five scales is provided in Ji & Jiang (2003). Besides the already listed scales, the review also features the scale of Ma & Zheng (1991) and the scale of Donegan, Dodd & McMaster (1995). Numeric values, corresponding to verbal expressions in the scale of Ma & Zheng (1991) are calculated according to the expression: $a = \frac{y}{y+1-x}$, where x is a respective integer value from Table 1, while y is the number of grades in the scale. Numeric values are presented in Table 5.

Table 5
Numeric equivalents for the scale of Ma & Zheng (1991)

Verbal expression	Corresponding numeric value
No dominance (equal)	9/9 = 1
Weak or insignificant dominance	9/8
Moderate dominance	9/7
More than moderate dominance	9/6 = 3/2
Strong dominance	9/5
More than strong dominance	9/4
Very strong dominance	9/3 = 3
Very-very strong dominance	9/2
Extreme dominance	9/1 = 9

The scale, suggested by Donegan, Dodd & McMaster (1995) is a bit more difficult to understand. Numeric values, corresponding to verbal expressions in the scale, suggested are calculated according to the expression $a = \exp\left[\tanh^{-1}\left(\frac{x-1}{h-1}\right)\right]$, where x is a respective integer value from Table 1, while h is a parameter, calculated based on the concept of horizons (ranges). Calculation of an 8-based horizon ($h = 1 + 14/\sqrt{3}$) is based on an assumption that for alternatives *A*, *B* and *C* the following transitive relation holds: *A* dominates over *C* with the degree of dominance 9 ($a_{AC} = 9$), if $a_{AB} = a_{BC} = 8$, i.e. „8 • 8 = 9". Calculation of the 7-based horizon: $h = 1 + 6/\sqrt{2}$ is based on a similar assumption that $a_{AC} = 9$ if $a_{AB} = a_{BC} = 7$, i.e. „7 • 7 = 9". Numeric values for the scale of Dodd, Donegan & McMaster (1995) are set forth in Table 6.

Table 6
Numeric equivalents for the scale of Dodd, Donegan & McMaster (1995)

Verbal expression	Corresponding numeric value
No dominance (equal)	1
Weak or insignificant dominance	1,132
Moderate dominance	1,287
More than moderate dominance	1,477
Strong dominance	1,720
More than strong dominance	2,060
Very strong dominance	2,600
Very-very strong dominance	3,732
Extreme dominance	9

A comparative study of the above-mentioned scales and optimization model for selection of scales are set forth in Dong (2008). A constructive original approach to classification of ratio scales and linking them to each other was recently suggested by William Wedley (2010).

In contrast to the research described in the listed publications (containing useful ideas to arm ourselves with), we suggest choosing a different scale for each single pair comparison and not for all pair comparisons. In the experimental part of our research we will focus on three particular scales: integer-value fundamental scale with 5 grades, 9 grades, and on a "mixed" scale, where an expert can chose the type of scale and the number of grades (from 2 to 9) for every single pair comparison.

3. Hypotheses/objectives

The purpose of the present study is to prove that in order to ensure that thorough and undistorted expert information on the relation between objects (on estimates provided during pair comparisons) is obtained, an expert should be given an opportunity to use scales with different degrees of detail (accuracy). This hypothesis is based on a presumption that in every issue under consideration (and in every pair comparison) an expert has a different level of knowledge/competency/awareness. Each expert's competence level can correspond to a respective estimation scale: the higher the expert's competence, the more detailed scale (s)he can use to adequately present his(her) knowledge. According to the same principle, an uninformed/incompetent expert should have an opportunity to use a scale with a smaller number of grades (including ordinal scale with only two values – "more" or "less") for pair comparisons, or even refuse to compare objects in a pair because of incompetence. It is understandable that an expert judgment provided in a more detailed scale should be considered more significant than that same judgment provided in a less detailed scale because in the first case the expert is more confident, and his self-estimated competence in the issue under consideration is higher. Consequently, if during pair comparisons an expert considers objects equal, this judgment can be considered the same as a refusal to conduct this particular comparison (inability to evaluate preference of objects in a pair due to doubts/low competency in the issue under consideration). As we see, in verbal scales there is no real need for a grade "equal"/"no preference", because if an expert chooses this value, (s)he might as well "skip" (refuse to estimate) the respective preference. Anyway, the choice of "equal" preference value does not introduce any additional information about the relation between objects.

Proof (confirmation) of any hypothesis in a weakly structured domain (in which we are conducting our research) is problematic, as there are absolutely no benchmarks to compare results with. That is why the only way to confirm the hypothesis is an experiment using estimates provided by experts. Such an experiment is described in section 5 of this paper.

4. Research design/methodology

During the research a methodology and respective software tools were developed to conduct expert estimation based on the abovementioned approach. In group estimation every expert is offered the opportunity to provide pair comparisons in verbal scales with different degrees of detail. Each particular pair comparison starts with the scale including only two values («Less» and «More») with an opportunity to refuse to provide the judgment – «No idea» (Figure 1).

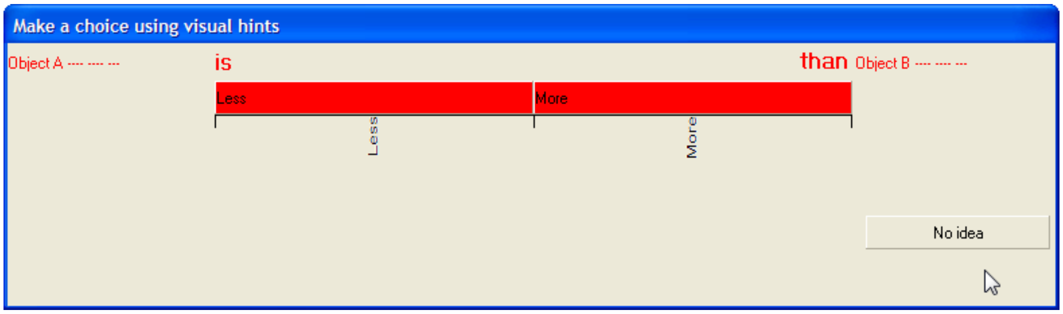

Figure 1. Initial estimation in ordinal scale

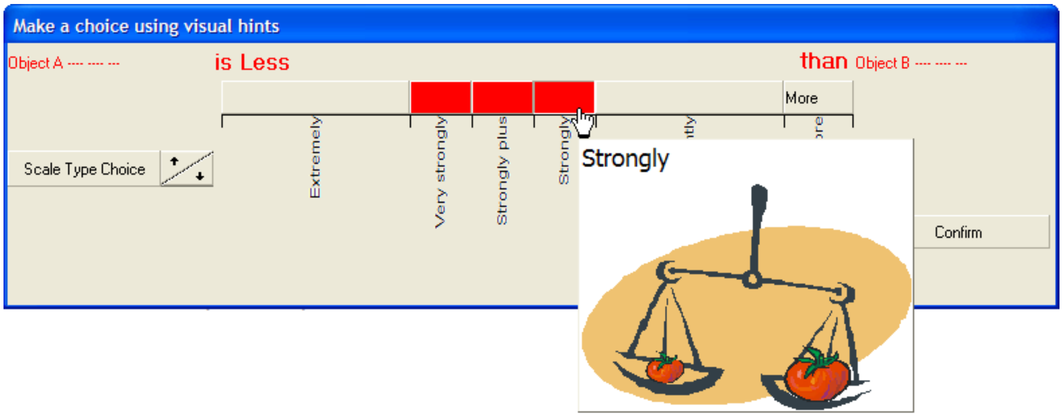

Figure 2. Procedure of gradual estimate precision increase

If ordinal comparison is provided (one of the values «Less» or «More» is selected) the expert is given the opportunity to gradually make the estimate more precise, and stop estimation at any stage («Confirm» button on Figure 2). In the process of this iterative procedure the final estimate is conducted in the scale which most adequately corresponds to the expert's competence about the issue of defining the preference relation between two particular objects. The final estimate may be provided in a scale including 2 to 8 grades.

It should be noted that the developed tool allows an expert to be sure that the quantitative equivalent really corresponds to this or that verbal expression from the estimation scale. Such confidence is achieved through providing the user (expert) with interactive graphic tips (hints) which allow him to imagine the approximate relation between objects and thus improve the degree of correspondence between the expert's personal notions and the information (s)he inputs during pair comparisons.

For aggregation of incomplete comparison matrices provided by a group of experts, when different comparisons can be conducted in scales with different accuracy, we suggest using the method based on enumeration of all spanning trees with further averaging of priority vectors calculated for every tree (Tsyganok, 2010). Before calculation of priority vectors, all pair comparison matrix elements (judgements) are brought to a unified (most detailed) scale. During this process, weights of particular judgements (pair comparisons) are taken into consideration. The weights depend on the degree of detail of scales the comparisons were provided in.

4.1 Problem statement

The formal statement of the alternative weight calculation problem in our case can be shown as follows.

What is given:

- A_i , $i \in [1..m]$ – expert pair comparison matrices (PCM) with dimensionality of $n x n$, which have the following properties: 1) matrices are reciprocally-symmetrical, that is why we are going to use only the elements above the principal diagonal; 2) matrices are multiplicative, i.e. each element a_{ij} shows how many times an object with index i is better than object with index j according to some criterion; 3) in the general case, matrices are incomplete, because an expert can, for some reasons, abstain from providing some pair comparisons; 4) every single element of a PCM is obtained in some scale, which is assigned a weight coefficient s_j, $j\in[0..8]$)

- c_l, $l\in[1..m]$ – relative competence of experts in the group.

We should find:
The resulting object (alternative) weight vector (priority vector) w_k, $k\in[1..n]$.

4.2 Unification of estimates

The problem of aggregating individual expert estimates includes bringing estimates provided in different scales to a unified form. At this stage, we consider it appropriate to bring estimates provided by different experts in different scales to a single scale, the most informative (detailed) one. Bringing the estimates to less informative scales is irrelevant, because in this case the information given in scales with a larger number of grades will be lost.

One of the ways to solve the problem of unification of estimates is to define clear correspondence between each of the grades of the less informative scale and some sub-set of grades in the more informative scale, and subsequently to bring the estimates to the more informative scale through selecting respective grades from these scales. In order to establish this clear correspondence, we should keep in mind that in the case when the grade of a less informative scale covers some range of grades in a more informative scale (one grade corresponds to a range of grades), when a respective value is selected on the more informative scale, it should be the value which is equally distant from the limits of this range. In this case a certain value, most likely provided by the expert, will be selected.

If we consider all possible estimate values corresponding to some scale grade, random values, distributed according to some law, which is close to normal, then when the information is unified and the estimates are brought to the more informative scale the estimate given in the less informative scale should be replaced by the mathematical expectation of the range of grades in the more informative one. For symmetric distribution laws (which include normal distribution), it is appropriate to take the simple average of lower and upper border (limit) values of the range of grades of the more informative scale, lying within the limits of the grade of the less informative scale.

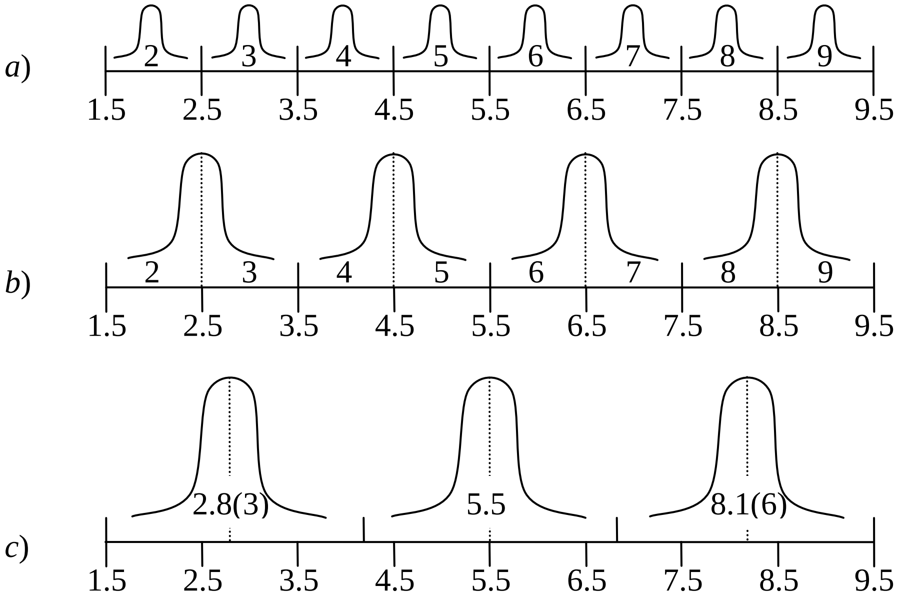

Figure 3. Numeric equivalents for grades in scales of: *a*) 8; *b*) 4; *c*) 3 grades

Figure 3a displays the standard integer-value scale of 9 grades (maximal number of grades among the scales under consideration), which includes numeric values from 2 to 9 (Elliot, 2010; Saaty 2006). Based on the abovementioned considerations, during pair comparisons in scales, which include from 2 to 9 grades of dominance of one alternative over another, a numeric value corresponding to a certain grade of a scale with a smaller number of grades (the less informative one) is calculated as a simple average of respective values of "range limiting" grades of the more informative scale. For instance, the value, corresponding to the 2nd grade from the left of the scale of 4 grades equals 4.5 (Figure 3b) in a scale of 9 grades. The step-by-step explanation is as follows. Due to psycho-physiological constraints, the largest possible number of grades in a scale equals 9. Each grade in a 9-grade scale corresponds to an integer value from the set {2..9}, while the limits of each of these grades are as follows: for the 2nd grade – 1.5 on the left, 2.5 on the right, for the 3rd grade –2.5 on the left, 3.5 on the right, etc. (Figure 3a). Consequently, for the 2nd grade of the scale of 4 grades (which corresponds to combined 4th and 5th grades of the scale of 9 grades), the average value is $\frac{3.5+5.5}{2}=4.5$, where 3.5 is the left border value of the 4th grade of the scale with maximal number of grades, while 5.5 is the right border value of the 5th grade.

Calculation of numeric equivalents for the scale of 3 grades is displayed on Figure 3c. The general formula for calculating these values for any given number of grades looks as follows:

$$M_i^n = l + \left(i - \tfrac{1}{2}\right)\frac{p-l}{n}, \qquad (1)$$

where M_i^n is the numeric equivalent of the *i*-th of *n* scale grades; *l* is the left border of the scale (l = 1.5); *p* is the right border of the scale (p = 9.5).

Consequently, the numeric equivalent for the rightmost (3rd) grade of the scale of 3 grades, is calculated as follows: $M_3^3 = 1.5 + \left(3 - \frac{1}{2}\right)\frac{9.5 - 1.5}{3} = 8.1(6)$, as shown on Figure 3*c*.

4.3 Weighting of estimates

Besides the problem of unification of expert estimates provided in different scales, processing of unified estimates is also an important task within the process of estimate aggregation. We suggest an approach to processing of unified estimates based on the idea of assigning different weights to estimates provided in different scales. An estimate provided in the more informative scale should weigh more than the estimate provided by an expert using the less detailed scale. The presumption is based on our belief that usage of a more detailed scale requires more competence in the issues under consideration from the expert. In fact, an expert using the scale with larger number of grades is in a way using a "more precise device" for measurement (estimation) of objects than an expert using a less detailed scale.

The specific form of dependence between weight (significance) of expert estimates provided in some scale and the number of grades in the scale needs to be addressed in a separate study. We can state however, that the dependence of the weight of an estimate provided in some scale on the number of grades in this scale is a monotonously increasing function. Besides that, it can be seen that the significance of adding more grades to a scale decreases with the increase of the number of grades. Based on these considerations, we suggest linking the scale weight coefficient (indicating the scale's degree of detail or informative content) to the quantity of information which can potentially be obtained from an expert providing an estimate in the given scale. In fact, such an indicator shows to what extent the usage of a certain scale for expert estimation decreases the general entropy of the system (subject domain description model).

One of the options for a simplified calculation of such informative content indicator, based on the assumption that selection of any scale grade by an expert is equally probable (probability is evenly distributed), is calculation of the quantity of information according to Hartley's formula (Hartley, 1928):

$I = \log_2 N$, where N is the number of expert estimation scale grades.

According to this formula, the following weight coefficients are "assigned" to the scales used: 0 (for N=1) – the expert cannot define the preferences among alternatives; 1 (for N=2) – the expert defined only ordinal preferences among alternatives ("better" or "worse"); $\log_2 3$ – after defining ordinal preference the expert defined the degree of dominance using 2 additional preference grades (for instance, "strong" or "weak" preference); $\log_2 4 .. \log_2 9$ – after defining ordinal preference of one alternative from a pair over the other, the expert specified the degree of dominance using the scale with respective number of grades (3 to 8). The advantages of the given approach to scale weight calculation are its simplicity and gradual decrease of significance of adding new grades to the scale. As we see, the estimates provided by experts during pair comparisons of alternatives are assigned weight coefficients depending on the scales they used for estimation.

4.4 Aggregation of estimates

After the estimates are unified and weighted, the PCM obtained from different experts can be aggregated. In order to fully utilize the redundancy of expert information, we suggest aggregating individual PCM using the so-called combinatorial method (Tsyganok, 2010). Before aggregation of estimates, completeness and consistency of individual PCMs should be checked. Completeness and consistency are ensured as follows. In the combinatorial method, priority vectors are calculated based on some basic sets of pair comparisons. If such basic sets are not complete, the priority vector cannot be calculated, so in such a case the expert should be re-addressed with a request to provide the missing basic pair comparisons, and thus ensure the completeness of the set.

If the matrices are not consistent enough, the results of the whole expert examination, even if they can be obtained, will be less credible. In order to check consistency (and, if necessary, improvement) we suggest using a spectral consistency coefficient (Zgurovsky, Totsenko & Tsyganok, 2004). After the aggregation is done, alternative weights can be calculated based on the aggregate PCM using one of the numerous approaches available (Tsyganok 2010). At this point the problem posed in the beginning of this section is solved.

5. Data/model analysis

To confirm the hypothesis set forth in section 3 of this paper, experimental research has been conducted with real experts. The description of the experiment itself is provided below. The purpose of the experiment is to compare the suggested technology with existing approaches for obtaining relative factor weights particularly with AHP-based ones (Saaty, 2008). We suggest comparing the degrees of correspondence between subject domain objects (factors) and relative factor weights, calculated using this or that expert estimation technology. In fact, we are comparing the results obtained using expert estimation technologies with the “model” values formed in the expert’s mind.

Results (relative factor weights), obtained based on the suggested technology we are testing, are compared with weights calculated based on pair comparisons provided in integer-value scales with 5 and 9 grades. Weights of criteria (factors) for pair comparisons in the fundamental scale are calculated using Saaty’s eigenvector method, based on PCM.

We suggest conducting the experiment in the following 4 stages:

1. *Formulation of the goal (problem) and factors which influence it.* At this stage the expert formulates the problem he or she considers him or herself competent in, and lists 5 to 7 mutually independent factors which in his\her opinion are most significant for solving this problem.

2. *Input of pair comparisons in three different scales*. All pair comparisons, which must be input throughout the experiment (for all three technologies, respectively, using scales with 5, 9 grades and a "mixed" scale) are offered to the expert for input in random order.

3. *Calculation of factor weights*. Relative weights of objects (factors) are calculated based on PCMs obtained at the previous stage which are processed using the respective methods. 3 priority vectors are calculated. The Eigenvector method is used to process matrices built using the first two approaches, while the so-called combinatorial (or spanning tree enumeration) method is used to define a priority vector based on a matrix including comparisons provided in different scales, (Tsyganok, 2010) (particularly, its modification allowing for usage of different weights for different estimation scales).

4. *Ranking of calculated weight vectors*. At this stage the expert ranks the three previously calculated factor weight vectors in their decreasing order of relevance, i.e. their correspondence to the expert's actual notions. Vectors in the form of unnamed bar diagrams are displayed on the screen in random order without technology specification. Each respondent (experiment participant) is offered the opportunity to rank the vectors according to their correspondence to his/her perceptions of quantitative relations between impacts of the formulated factors.

A quantitative indicator, formed as a result of experiment, is the degree of preference of this or that expert estimation technology (i.e., frequency of this technology being rated as number one in the ranking of all technologies under consideration).

5.1 Experiment integrity aspect

In order to ensure experiment integrity certain steps were taken.

1) Subject domain chosen by the respondent.
At the first of the listed stages we suggest that an expert select a subject domain, in which he or she considers him or herself competent in. The expert is offered the opportunity to formulate a problem which, in his or her opinion, is most understandable. Independent subject domain choice guarantees that expert examination organizers are not biased in any way (as the subject domain is chosen by the expert). The expert examination organizer (knowledge organizer) does not "impose" a subject domain upon the expert, so there can be no situation in which the expert is not sufficiently competent to conduct estimation in the subject domain. Because of the described feature, the experiment becomes more universal than the experiment conducted by Elliot in 2010.

2) Same factors, different technologies.
Once the problem is formulated, the expert is offered the opportunity to list a set of factors which describe it. This same set of factors is used to evaluate different expert estimation technologies in the process of the experiment, thus ensuring the relevance and credibility of comparison of results obtained using different technologies.

3) Independence of factors.
When criteria (factors) are formulated, it is required that they must describe the problem most thoroughly, and at the same time be mutually independent (no "intersections" or

mutual impacts between criteria are allowed). This is the necessary condition for obtaining credible results with PCM processing methods.

4) Random order of factor enumeration.
Another critical feature embedded in the experiment is the opportunity to name the factors in random order: in the first place the expert, intuitively, recalls (and, consequently, names) the factors he or she considers most important. This order of factors remains the same for all methods. This condition is also important for ensuring equal credibility of results obtained based on pair comparisons (as in pair comparison method the number of comparisons a given object is featured in depends on the order of objects).

5) Limitation on the quantity of factors.
The number of factors must not exceed 7±2. This condition is determined by psycho-physiological constraints of an average individual (expert) (Miller, 1956). Besides, the condition also plays an important role when ranges and numbers of grades for used scales are defined.

6) Random order of comparisons.
Random choice of pairs of objects (factors) to be presented to the expert for comparison, as well as random choice of estimation technology, *at the second stage*, allow for a decrease in correlation between estimates of ratios obtained using different methods for the same pairs of objects during one experiment session (instance). In this way we can ensure mutual independence of specific pair comparisons.

7) “Blind” ranking of estimation technologies.
At the fourth stage, the expert is presented a bar diagram, on which he or she can chose one of three factor weight vectors. The type of the scale in which the estimation has been conducted is not specified. In this way, we can guarantee that the rankings of expert estimation technologies according to their correspondence to expert’s own notions of the problem are unbiased.

5.2 Interpretation of experiment results

The experiment was conducted using a specially designed software application. The result of each experiment represented a file including the following information:

1. Expert’s name
2. Problem title
3. List of factors
4. Matrices of paired comparisons of factors, obtained using different technologies of expert estimation
5. Time, spent by the expert on every question (comparison)
6. Ranking of expert estimation technologies (criterion weight vectors)
7. Expert’s own explanation of the ranking

Every expert can participate in the experiment several times as long as he(she) formulates different problems. Once a particular examination (expertise) is done, the relevance of its results is checked because further statistical processing of results obtained from different experts can be done only under the condition that the set of expert estimation technology rankings is adequate. A relevance (adequacy) check includes two stages:

1. During the analysis, results which can be interpreted as “careless” are filtered off. These include a) results which took the expert too little time to obtain (any expert requires at least several seconds to formulate a well-considered answer or estimate); b) PCMs where most estimates coincide; c) files with resulting data which were corrected manually; d) incomplete PCMs, based on which priority vectors cannot be calculated due to lack of basic pair comparisons.

2. Factor weight vectors are filtered off if PCMs, based on which these weights were calculated, contain significant contradictions (inconsistency level is too high). For example, if one of the three resulting weight vectors is significantly different from the other two, the expert might ignore this vector and the respective technology, and not include it into the ranking. Even if the vector and the respective technology are included in the ranking as the lowest-ranking (most inadequate) one, the whole ranking will not bear information on correspondence between weight vectors and expert’s notions of the problem.

5.3 Statistical relevance of experiment’s results

The minimal number of experiment instances necessary for achievement of desired statistical relevance of results has been calculated as follows. The estimate of statistical relevance was based on central-limit theorem. Under a confidence probability level P_β=0.9 (probability that the random value in question falls into the confidence interval β), and confidence interval size β=0.1, selected for the experiment, the necessary number of experiment instances is calculated from the inequality:

$$n \geq \frac{p \cdot (1-p)}{\beta^2}\left(F^{-1}\left(P_\beta\right)\right)^2, \qquad (2)$$

where F^{-1} is a inverse Laplace function; p is the frequency of repetition of the resulting random characteristic.

We select the values of p based on the data from the table of experimental results already obtained (see Table 7) as the worst (closest to 0.5) probability (frequency). In our case, “worst” frequency means that the characteristic assumes some value as frequently as it does not, so that probabilities p and $(1 – p)$ are equal. If we look at the second column of the table {10/63≈0.16; 12/63≈0.19; 41/63≈0.65} we can see that frequency p =0.65 is the “worst” one in a sense. This value should be put into the formula above.

If we replace the variables with the actual values, we get the following results:

$$F^{-1}(0.9) \approx 1.65, \ \left(F^{-1}(0.9)\right)^2 \approx 2.72, \ n \geq \frac{0.65 \cdot (1-0.65)}{0.1^2} 2.72, \ n \geq 61.9 .$$

This means that in order for the experiment results (i.e. conclusions regarding the preference of this or that decision support technology) to be relevant it is sufficient to analyze data from at least 62 instances of the experiment.

5.4 Numeric results of the experiment

The experts/respondents in the experiment were represented mostly by graduate IT students at one of the universities in Kyiv, Ukraine. So far, around 100 respondents participated in the experiment. After some results were filtered off, as described above, the set was limited to 63 rankings of factor weight vectors (and respective expert estimation technologies), and 2 rankings turned out to be incomplete (included two, but not three vectors). The results are presented in Table 7.

Table 7
Comparative experimental research results

Name of pair comparison technology	Number of participants, who assigned the specified rank to respective technology		
	"1"	"2"	"3"
Fundamental preference scale with 5 grades	10	15	37
Fundamental preference scale with 9 grades	12	33	17
Technology suggested in the paper	41	15	7

As we can see, most respondents preferred the suggested expert estimation technology based on aggregation of results of pair comparisons provided in different scales. Based on the results of the experimental research, we can conclude that in most of the analyzed cases expert estimates obtained using the suggested technology are more consistent with an expert's individual perceptions of the examination subject, in comparison to estimates, based on traditional estimation techniques (where fixed number of verbal scale grades is used). Consequently, wide implementation of the suggested pair comparison instrument in decision support technologies (including those using AHP/ANP) seems adequate.

6. Limitations

Usage of the suggested tool for pair comparisons may require a longer time during expert estimation and as a result more resources than traditional methods. This may result from the fact that more actions are required from experts during each pair comparison. However, this also results in higher credibility of expert estimates and recommendations to decision makers.

7. Conclusions

As a result of the research, we have suggested an expert estimation mechanism which allows experts to use scales of different accuracy for each pair comparison. Relevance of the suggested approach is experimentally proven. It has been demonstrated that usage of the respective tool for pair comparisons allows us to improve the degree of correspondence between an expert's estimates and his notions of the examination subject. This improvement results from the fact that experts use scales whose accuracy is most consistent with their competency in every issue under consideration.

The suggested approach can be considered an extension of (and not an alternative to) AHP/ANP approaches. Implementation of the suggested expert estimation technology in combination with pair comparison matrix aggregation methods (including group methods) improves the credibility of AHP/ANP-based recommendations given to decision makers.

The technology's basic advantages are universality and flexibility. These features allow it to be utilized in existing and new decision support tools, particularly in those areas where multiple criteria (factors, attributes) of different natures, both qualitative and quantitative, should be taken into account and where expert data can be the only credible source of information. Such areas include managerial decisions, project selection, personnel evaluation, logistics, strategic planning and others.

REFERENCES

De Felice, F., & Petrillo, A. (2013a). Multicriteria approach for process modelling in strategic environmental management planning. *International Journal of Simulation and Process Modelling, 8(1)*, 6–16. doi: http://dx.doi.org/10.1504/IJSPM.2013.055190

De Felice, F., & Petrillo, A. (2013b). Decision-making analysis to improve public participation in strategic energy production management. *Studies in Fuzziness and Soft Computing 305*, 129–142. doi: 10.1007/978-3-642-35635-3_11

Dodd, F.J., Donegan, H.A., & McMaster, T.B.M. (1995). Scale horizons in analytic hierarchies. *Journal of Multi-criteria Decision Analysis, 4*, 177–188. doi: 10.1002/mcda.4020040304

Yucheng Dong, Yinfeng Xu, Hongyi Li, & Min Dai (2008). A comparative study of the numerical scales and the prioritization methods in AHP. *European Journal of Operational Research, 186*, 229–242. doi:10.1016/j.ejor.2007.01.044

Elliott, M.A. (2010). Selecting numerical scales for pairwise comparisons. *Reliability Engineering and System Safety, 95*, 750–763. doi:10.1016/j.ress.2010.02.013

Hartley, R.V.L. (1928). Transmission of information. *Bell System Technical Journal, 7*, 535–63. doi: 10.1002/j.1538-7305.1928.tb01236.x

Ji, P., & Jiang, R. (2003). Scale transitivity in the AHP. *Journal of the Operational Research Society, 54(8)*, 896–905. doi:10.1057/palgrave.jors.2601557

Kalika, V.I., & Rossinsky, G. (2003). Methodology of multi-criteria decision making accounting for uncertainty and some applications. *International Journal of Management and Decision Making, 4(2/3),* 240 – 271. doi: http://dx.doi.org/10.1504/IJMDM.2003.003507

Kannan, G., Noorul Haq, A., & Sasikumar, P. (2008) An application of the Analytical Hierarchy Process and Fuzzy Analytical Hierarchy Process in the selection of collecting centre location for the reverse logistics. Multicriteria Decision-Making supply chain model. *International Journal of Management and Decision Making, 9(4)*, 350 – 365. doi: http://dx.doi.org/10.1504/IJMDM.2008.019360

Lootsma F.A. (1989). Conflict resolution via pairwise comparisons of concessions. *European Journal of Operational Research, 40,* 109–116. doi:10.1016/0377-2217(89)90278-6

Lootsma F.A. (1991). *Scale sensitivity and rank preservation in a multiplicative variant of the AHP and SMART. Report 91–67.* Faculty TWI. Delft University of Technology. Delft. The Netherlands.

Ma, D., & Zheng, X. (1991). Scale method of AHP. *Proceedings of the Second International Symposium on the AHP, 1*, 197–202.

Miller, G.A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *The Psychological Review, 63(2)*, 81–97. http://dx.doi.org/10.1037/h0043158

Noorul Haq, A., & Kannan, G. (2007). A hybrid normalised multi criteria decision making for the vendor selection in a supply chain model. *International Journal of Management and Decision Making, 8 (5/6)*, 601–622. doi: http://dx.doi.org/10.1504/IJMDM.2007.013421

Saaty, T.L. (2006). *Fundamentals of decision making and priority theory with the analytic hierarchy process*. Pittsburgh, PA: RWS Publications.

Saaty T.L. (2008). Relative measurement and its generalization in decision making. Why pairwise comparisons are central in mathematics for the measurement of intangible factors. The Analytic Hierarchy/Network Process. *Statistics and Operations Research, 102(2)*, 251–318. doi:10.1007/BF03191825

Salo, A.A., & Hamalainen, R.P. (1997). On the measurement of preferences in the analytic hierarchy process. *Journal of Multi-criteria Decision Analysis, 6*, 309–319. doi: 10.1002/(SICI)1099-1360(199711)6:6<309::AID-MCDA163>3.0.CO;2-2

Stevens, S.S. (1957). On the psychophysical law. *Psychology Review, 64*, 153–181. doi: http://dx.doi.org/10.1037/h0046162

Tsyganok, V.V. (2010). Investigation of the aggregation effectiveness of expert estimates obtained by the pairwise comparison method. *Mathematical and Computer Modelling, 52(3-4)*, 538–544. doi:10.1016/j.mcm.2010.03.052

Vaidogas, E.R., & Zavadskas, E.K. (2007). Introducing reliability measures into multi-criteria decision-making. *International Journal of Management and Decision Making, 8(5/6)*, 475 – 496. doi: http://dx.doi.org/10.1504/IJMDM.2007.013413

Wedley, W.C., & Eng Ung Choo (2010). *A taxonomy of ratio scales*. OR-52 Keynotes and Extended Abstracts, Operational Research Society Ltd., Royal Holloway University of London, UK 7-9/09/2010, 199–203.

Zandi, F. & Tavana, M. (2010). A multi-attribute group decision support system for information technology project selection. *International Journal of Business Information Systems, 6(2)*, 179 – 199. doi: http://dx.doi.org/10.1504/IJBIS.2010.034353

Zgurovsky, M.Z., Totsenko, V.V., & Tsyganok, V.V. (2004). Group incomplete paired comparisons with account of expert competence. *Mathematical and Computer Modelling, 39(4)*, 349–361. doi:10.1016/S0895-7177(04)90511-0