International Journal of Approximate Reasoning 83 (2017) 356–370

Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

www.elsevier.com/locate/ijar

Exploiting collective knowledge with three-way decision theory: Cases from the questionnaire-based research ☆

Federico Cabitza [a,b,*], Davide Ciucci [a], Angela Locoro [a]

[a] Dipartimento di Informatica, Sistemistica e Comunicazione, Università degli Studi di Milano-Bicocca, Via Bicocca degli Arcimboldi 8-20126, Milano, Italy
[b] I.R.C.C.S. Istituto Ortopedico Galeazzi, Via Riccardo Galeazzi, 4-20161, Milano, Italy

ARTICLE INFO

Article history:
Received 8 July 2016
Received in revised form 7 October 2016
Accepted 16 November 2016
Available online 23 November 2016

Keywords:
Three-way decision
User study
Online questionnaires
Collective knowledge

ABSTRACT

Two methods are proposed for collective knowledge extraction from questionnaires with ordinal scales and dichotomous questions.
Both methods are based on a three-way decision procedure and a statistical method aimed at attaining statistical significance of the above decision. One method is aimed at giving an (absolute) assessment of "objects" according to a given "criterion" and the other one at producing a relative ranking of the "objects". A criterion can be related to one or more questionnaire items (usually questions or statements). In this latter case a method to compose ordinal items in aggregate scores is also given. The paper also presents two various case studies that illustrate the methods and give motivations for their application in different domains where the knowledge of a community or any distributed group of experts can be externalized (in terms of users' perceptions, attitudes, opinions, choices) with a structured closed-ended questionnaire.

1. Introduction

Three-way decision (3WD) theory is based on a trisecting-and-acting paradigm and it aims to classify a finite collection of objects in three regions (positive, negative, boundary) in order to act differently on objects belonging to different regions [1, 2]. It was introduced in 2009 by Yiyu Yao [3] and despite the fact that it is a young research field, it is gaining a consensus in the rough-set community and beyond [4] leading towards a standalone discipline. Due to its relative novelty, a number of issues are still open and Yao in [2] outlines some directions for future works. With respect to trisection of the universe, which is the central point of the methodology, open problems include

> to consider a ranking of objects, which is a fundamental notion in theories of decision-making. Three regions correspond to the top, middle, and bottom segments of the ranking.

Moreover, a further research direction Yao identified was to "seek for a statistical interpretation of a trisection". First steps to this aim can indeed be found in [5], where basic statistical measures such as average and standard deviation are used to partition the universe.

☆ This paper is part of the virtual special issue on Tri-partition, edited by Davide Ciucci and Yiyu Yao.

* Corresponding author.
E-mail addresses: cabitza@disco.unimib.it (F. Cabitza), ciucci@disco.unimib.it (D. Ciucci), angela.locoro@disco.unimib.it (A. Locoro).

http://dx.doi.org/10.1016/j.ijar.2016.11.013
0888-613X/

Here, we are going to delve into the two research concerns mentioned above in an integrated manner, namely, to provide a three-level ranking of objects using a statistical method and interpretation. The general criterion that we propose to divide a universe U into three pair-wise disjoint regions is the *collective perception*[1] of a population of domain experts.

This requires to ask n experts (i.e., a sample from a population) if an object x (e.g., a scientific conference, a diagnosis, a medical treatment) on the basis of some criteria (e.g., the quality of the conference x, the plausibility of diagnosis x, the appropriateness of therapy x) should be classified, that is put, in a specific partition of U (e.g., class of quality, level of plausibility, strength of appropriateness).

Then, a simplistic way would be to ask each expert to put object x in a partition (or to abstain to move it from the border) and then to adopt the majority criterion: the object x is moved to a partition if the majority of the sample decided to put it there. However this procedure exhibits two main shortcomings. First, it is subject to what we could call "dichotomous choice" bias, that is the bias coming from the fact that people have to decide between two alternatives but they are not really convinced and take the decision almost by coin flipping or, worse yet, subject to some *cognitive effect* (like *priming* [7]). In this case, even a small effect can be amplified by the number of subjects involved. Second, the partition would express the collective decision (deliberation) of a specific group of people (the sample), but nothing could be said of the overall population. A priori, a different sample could bring to a completely different partitioning.

To go beyond this latter shortcomings (probably the worst one) an alternative approach would be to estimate whether the population, from which the sample has been extracted, expresses any majority-ruled consensus. To this aim, the real proportions of both two kinds of responses (that is "x should be put in U_i" or it should not) can be estimated from the proportions extracted from the sample of respondents by performing a hypothesis test (e.g., the chi-squared test).

In this paper, we adopt a similar approach of statistical inference, while also addressing the dichotomous choice bias. To this latter aim, we do not ask directly the experts in which partition of U they would put x; rather, for each respondent we infer her/his classification by first analyzing her/his absolute evaluations of x recorded on a psychometric ordinal scale (e.g., a Likert scale [8,9]). This, if the scale is sufficiently wide, could give the respondents a way to express either a strong attitude and convinced resolution in either directions (one partition or the latter) by choosing the outer or extreme values of the scale or, conversely, express a weaker conviction by choosing the middle values of the scale. And then by considering the aggregate classification through a statistical procedure. This latter one is aimed at rejecting the *null hypothesis* that no partitioning decision can be made beyond the effects due to chance (that is for biases due to the specific sample under consideration, in the assumption that it has been randomly extracted from the reference population and it is representative of the whole population for the characteristics of interest, e.g., expertise).

More specifically the statistical hypothesis test would provide a number to be compared with an acceptance threshold and hence it would give an indication on whether the perceptions collected from the sample of respondents allow us to reject the null hypothesis above (assumed as true), or not. Thus, the result of the hypothesis test represents the so-called *evaluation function* of the three-way method [2]. The reader should note that in our approach there is not a "right" classification, so we cannot compute the thresholds by minimization/maximization of a quantity measure, as it is the case in other applications of three-way decision theory. We rather aim at the *most representative* classification of a community of experts, which is inherently 3-partitioned for the difficulty of reaching a large consensus especially in non-trivial matters.

As a first step of our method, we thus produce what we could call an *absolute* classification of objects in three regions. We remark that this classification is obtained through two main resources: the *collective knowledge* [10] that is available in a community of experts, and a *statistical procedure* through which to tap in this knowledge [11].

Further, also a *relative* classification in terms of a ranking of objects is obtained, again by a three-way decision process. In this case, the three regions of U are those related to the capability to put an object into any of the "first positions" (e.g., the first three positions, like in a podium), or in any of "the other positions" *beyond any doubt* (that is beyond the effect on the above decision that is due to chance), or to the statistical incapability to do so (because the user responses do not allow to reject the hypothesis stating that the object could not put in either of the above partitions). Also in this case we propose to use ordinal values (i.e., categories from a total order set), but indirectly: by conjecturing that respondents are better in giving absolute judgments than rankings, especially when the entities to rank are many and hence differences among them subtle (if any), we derive a ranking for each respondent, and then we look for any tendency in the ranking distribution beyond the effects due to chance.

The paper is structured as follows. In Section 2, we give an overall picture of our approach and explain the method to compose ordinal variables together. Section 3 contains the two main methods of our contribution: the three-way assessment and ranking procedures based on questionnaire (ordinal) responses. These two methods are then applied to two case studies in Section 4: the quality evaluation of scientific conferences and the management of difficult medical cases, which we chose for their heterogeneity and controversial or difficult practices of decision making, respectively. Finally, in Section 5 we present the conclusions and outlines for future research.

[1] "A perception is defined as the result of a cognitive process whereby a person interprets information" [6]. A collective perception is then the perception that can be inferred from a collective of people, each of whom expresses an individual perception. To our aims such an abstraction can be expressed in terms of a stochastic variable.

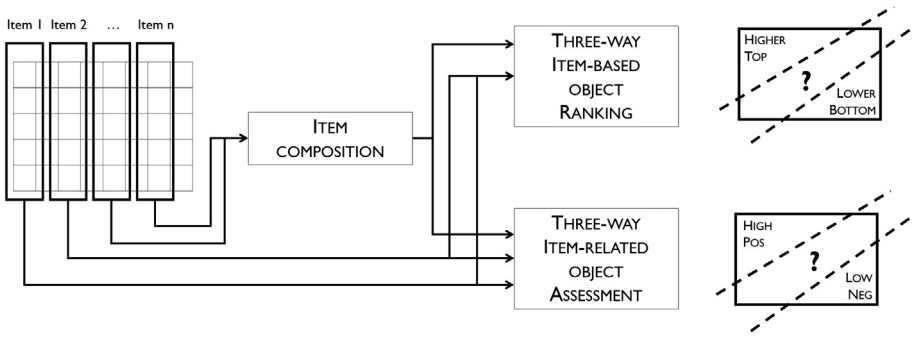

Fig. 1. Flowchart of the assessment and ranking methods. Responses are intended item-wise, that is by column with respect to Fig. 2.b.

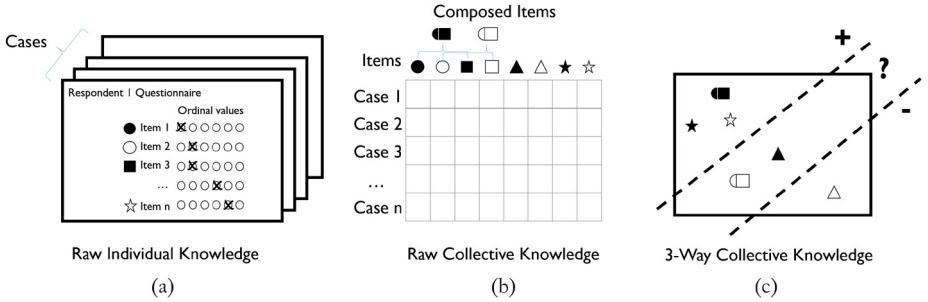

Fig. 2. Three ways to see the same knowledge on particular object of interest as it is represented in terms of experts' responses to structured questionnaires. In (a) this collective knowledge is still tacit and expressed at individual level; in (b) this knowledge is objectified in a matrix of ordinal values. In (c) this knowledge has been transformed in suitable indications (tripartition classification) for interpretation and action.

2. Preliminary notions and variable composition

In this section, we define our setting and explain the method of ordinal variable composition. In questionnaire-based research, respondents are asked to respond to questions or express their agreement with respect to some statement so as to share their perceptions about what are called *items*, that is aspects that the questionnaire addresses, usually a measurable characteristic of an object, be this physical or abstract. To our aims, questionnaire items are variables. Given the responses to a set of questionnaire's items, their values can be composed into a unique aggregate variable. Both single and composed items can be assessed through a three-way method (Section 3.1) or used to produce a three-way ranking (Section 3.2) of objects that are described (to some extent) by those items. The overall schema of our approach is depicted in Fig. 1, where objects are represented in terms of characteristics that are in their turn expressed in terms of items; each item can be seen as the set of the responses given by the respondents in regard to the above characteristic. Two or more characteristics (i.e., items) can be composed in a higher-level construct that subsumes them in some way. Any single object characteristic, expressed in terms of either an original item or any new one created by composition of the original ones, can be assessed according to the three-way decision of putting the object in either a high or low level with respect to the characteristic of interest; two or more objects can be ranked by considering one item (or characteristic) at a time, that is be put in either a higher (top) level or lower (bottom) level according to a three-way decision that is based on the responses given to a specific original item or to the values of any composed item.

In Fig. 2 we depict how questionnaire responses can be transformed into a collective three-way classification of objects (from left to right). In particular Fig. 2.a depicts a set of questionnaires filled in by a sample of respondents (usually experts in some knowledge domain); each questionnaire allows for the collection of opinions, beliefs, perceptions, attitudes expressed by a respondent about some characteristic of some object in terms of ordinal values. Each questionnaire is a case (see Fig. 2.b), that is a row of the response matrix, where each column represents an item (i.e., the characteristic of some object). Colors (black and white) indicate different "objects" to be assessed and ranked according to some common characteristics. In Fig. 2.b we represent the raw collective knowledge of the expert sample regarding two objects, or more precisely, regarding five common characteristics of theirs, four natively represented by items of the questionnaire (namely 'circle', 'square', 'triangle', 'star') and one higher-level construct created by composition (namely, 'circle-square'). These names have been chosen abstract purposely: for instance, the circle item can be *conference selectivity*; the square item can be *proceedings quality*; the circle-square item can be *overall quality*; black object is conference "Acme", the white object is conference "Ace". In Fig. 2.c the ranking process can put the white object in the positive region with respect to the star item, in the negative region for the triangle item, and in the boundary region for the composed item circle-square. The black

object can be put in the positive region for the composed item and the star item, and in the border region for the triangle item.

2.1. Assumptions

Let us consider k variables, also called items, that are ordinal in nature [12] and can have v possible values (with V the set of these values). In what follows we will assume that the cardinality of V is 6 and the values will be denoted as $V = \{v_1, \ldots, v_6\}$ with $v_1 < v_2 < \ldots < v_6$. The adoption of a six-value scale is motivated by the main fact that, in so doing, three natural levels have two distinct values each; therefore, no middle point is given, to minimize *central tendency bias*.[2] Moreover, the total number of values is not too high, which would increase noise (because the same span or spectrum of a characteristics is divided in more narrower intervals), and yet not too low, so as to have sufficient variability and give the respondents sufficient choice to express their perceptions.

Let S be a set of *cases* s_j (see Fig. 2.b), each case representing all the responses of a single respondent to a questionnaire, i.e., S is the collection of all the responses of all respondents (the matrix in Fig. 2.b). An element of S is a k-value tuple, each element being an element from V or the NULL value (when a respondent has given no answer to a question, that is she skipped the question) or the DK value (when a respondent has given a "Don't know" answer to a question, admitting not to have an opinion to that respect). Thus, $S = \{s_1, \ldots, s_n\}$, where any s_j are the responses of a respondent $s_j = \langle x_1^j, \ldots x_k^j \rangle$ and $x_i^j \in V \cup \{NULL, DK\}$ are the responses to each item of the questionnaire.

Example 2.1. Let us consider as an example the conference evaluation that will be discussed in Section 4.1. Each of the 21 conferences will be evaluated according to some criteria. Thus, a variable x_i will represent the response of a respondent in regard to a conference (an object) with respect to a given criterion (or characteristic of the object). For instance, the researcher s_j interest with respect to the conference *AAAI-Conference on Artificial Intelligence* is a variable x_i^j with values on the scale 1 *(very far)* to 6 *(very close)*.

Finally, we introduce a totally ordered set O, containing three values: $O = \{L, M, H\}$ with $L < M < H$. This set will be used to map an expert's response from the six levels in V to three levels in O, which therefore can be interpreted as Low, Medium and High. To be more precise, we consider the union set $O' = O \cup \{NULL, DK\}$ as the available symbols to represent the users' responses. As hinted above, the symbol DK (acronym of the phrase "Dont' Know") is used whenever users do not want (or cannot) to make their perception explicit (on the supplied ordinal scale). The element 'NULL' is associated with the absence of a response. Therefore, NULL represents what is usually called the *system missing*, while DK denotes the so-called *user missing* [13].

Finally, given m variables, we denote the set of all of the 5^m possible dispositions of O' as T.

2.2. Composition method for ordinal variables

Respondents are called to give a symbolic representation of their perception for a specific dimension of interest on an ordinal scale, by choosing an element from V, totally ordered set. They can also select one value that does not belong to V, i.e., DK or $NULL$.

An overall respondent's judgment, or case (see above), is composed of different perceptions, one perception along each dimension.[3] A judgment (or case) is then represented in terms of a tuple of values from O'. We denote the set of all these tuples as T, and remark that T is a partially ordered set.

Example 2.2. Let us consider again the case study on the quality of scientific conferences. In a questionnaire, each conference is to be evaluated by expert scholars along two dimensions: *selectivity* and *quality of selected papers*. These two variables can be composed into a unique score, the *overall quality* of the conference.

We now describe a method to compose the variables of this tuple by giving a unique value in V through a three steps procedure:

1. *Case downsampling* from V to O. Each (of the six) value in V is mapped to one of the three values in O;
2. *Item composition:* a tuple T of values in O' is mapped to a unique value in a set R;
3. *Score standardization:* each value in R is mapped back to a value in V.

[2] Central tendency bias occurs when raters or respondents avoid using extreme response categories on ordinal scales and rather tend to choose the options in the middle of the scale not because they actually could not decide, but rather to respond faster and with less cognitive effort.

[3] Dimension, attribute, characteristic, variable and item are synonyms and treated as such in this paper. More precisely, the extensional definition of a variable would be: an unordered set of values, each representing (on an ordinal scale) the single perception of a single respondent.

In the following we give the details of these three steps.

1) Case downsampling. In this phase we define a non-injective and non-surjective function between V and O that maps each value of each case (i.e., perception of the respondent) from v' ($v' \in V$) into one value of O. This is done to reduce the number of possible combinations in the step 2, and facilitate the adoption of the most appropriate heuristics of score normalization in step 3. If the cardinality of V is a number that is divisible by 3 (e.g., $|V| = 6$) the downsampling mapping is trivial: for instance, if V contains 6 symbolic numbers, 1s and 2s are mapped into L, the 3s and 4s into M, and 5s and 6s into H. Formally, we defined a function $cd : V \mapsto O$ as:

$$cd(v) = \begin{cases} L & \text{if } v = v_1 \text{ or } v = v_2 \\ M & \text{if } v = v_3 \text{ or } v = v_4 \\ H & \text{if } v = v_5 \text{ or } v = v_6 \end{cases}$$

2) Item composition. In this phase, for a given tuple from T, we build a composite score from k variables to a unique value on an ordinal scale R.

First of all, a tuple $t \in T$ containing a NULL or a DK is mapped directly to NULL/DK.

So, once discarded the tuples with NULL/DK values, we remain with the 3^k words of length k on the alphabet $O = \{L, M, H\}$[4] (that is, for $k = 3$, there are 3^3 values).

These values are then partitioned according to an equivalence relation we are going to define and R contains a value for any equivalence class.

Given two tuples $t_1 = (o_1^1, o_2^1, \ldots, o_k^1)$ and $t_2 = (o_1^2, o_2^2, \ldots, o_k^2)$ we compute the value $D(t_1, t_2) = \sum_{i=1}^{k} d(o_i^1, o_i^2)$ where $d(o_i^1, o_i^2)$ is the number of positions by which o_i^1 and o_i^2 differ in O. If $O = \{L, M, H\}$ then $d : O \mapsto \{-2, -1, 0, 1, 2\}$ is defined as

$$\begin{cases} d(L, M) = d(M, H) = -1 \\ d(L, H) = -2 \\ d(M, L) = d(H, M) = 1 \\ d(H, L) = 2 \\ d(x, x) = 0 \text{ for } x \in O \end{cases}$$

Two tuples are equivalent if $D(t_1, t_2) = 0$ and hence they are put in the same class. Moreover, the set $R = \{r_1, \ldots, r_f\}$, representing the different classes, is an ordered set such that $d(t_i, t_j) = 1$ iff r_i and r_j are two consecutive classes that differ for one position only.

Example 2.3. Let us consider the case $k = 4$. Then, $t_1 = (L, M, L, H)$ and $t_2 = (M, M, M, L)$ belong to the same class. Indeed $D(t_1, t_2) = -1 + 0 - 1 + 2 = 0$. The "following" class of the one containing t_1 and t_2 contains, for instance, $t_3 = (M, M, M, M)$, since $D(t_1, t_3) = D(t_2, t_3) = 1$.

Finally, the cardinality of R is $f = 2k + 1$.

In Table 1, the explicit conversion for $k = 3$ is given, where classes are denoted by Roman numbers. We notice that the set R must contain an ordinal value for the lowest level and one ordinal attribute for the greatest level. In our example with $k = 3$, $|R| = 7$ and R contains a value associated to the triple (L, L, L) and one to the triple (H, H, H).

3) Score standardization. In this phase, all of the unique values of R are mapped back into V, that is an equivalent numeric scale with contiguous values to enable statistical standard techniques. This mapping is performed heuristically. We detected at least two ways to this aim.

(ss1) In one case, the number of times a value can occur is considered and we assign values so that the probability of occurrence is most evenly distributed among all the possible values.

(ss2) In the second case, the n values in R are mapped into the m values in V so as to make the resulting groups as much evenly distributed as possible in terms of number of values in R being mapped into a single value of V.

In our case studies, we have adopted the latter heuristic (namely, ss2), but we also assigned the lowest and highest values biunivocally, that is the lowest in R mapped into the lowest in V, and the same for the highest, and then the other ones by following the main heuristic above. Thus, in our example with $|R| = 7$ and $|V| = 6$ the mapping is as in Table 2.

If there is a need to compose composite indicators together, the algorithm can be reiterated from step no. 1.

3. Three-way assessment and ranking

The two methods for (absolute) assessment and (relative) ranking, based on a statistical three-way decision, are given.

[4] Also named permutations with repetition of L, M and H.

Table 1
The explicit mapping for the composition function from T to R when $k = 3$. For brevity's sake, we show only the 3^3 disposition that are mapped into a value different from NULL and DK.

L	*L*	*L*	↦	I	*M*	*H*	*L*	↦	IV
L	*L*	*M*	↦	II	*H*	*L*	*M*	↦	IV
L	*M*	*L*	↦	II	*H*	*M*	*L*	↦	IV
M	*L*	*L*	↦	II	*L*	*H*	*H*	↦	V
L	*L*	*H*	↦	III	*M*	*M*	*H*	↦	V
L	*H*	*L*	↦	III	*M*	*H*	*M*	↦	V
L	*M*	*M*	↦	III	*H*	*L*	*H*	↦	V
M	*L*	*M*	↦	III	*H*	*M*	*M*	↦	V
M	*M*	*L*	↦	III	*H*	*H*	*L*	↦	V
H	*L*	*L*	↦	III	*M*	*H*	*H*	↦	VI
L	*M*	*H*	↦	IV	*H*	*M*	*H*	↦	VI
L	*H*	*M*	↦	IV	*H*	*H*	*M*	↦	VI
M	*L*	*H*	↦	IV	*H*	*H*	*H*	↦	VII
M	*M*	*M*	↦	IV					

Table 2
Mapping from values in R to values in V, for $R = 7$ and $V = 6$. The value IV is evenly mapped either to v_3 or v_4.

I	↦	v_1
II	↦	v_2
III	↦	v_3
IV	↦	(v_3 or v_4) with $p = .05$
V	↦	v_4
VI	↦	v_5
VII	↦	v_6

3.1. Method of three-way item (absolute) assessment

The aim of this method is to assess the most representative level category for a single item of a questionnaire, as a sort of *absolute* assessment (that is not affected by the other items in the same questionnaire). This is usually done by calculating a central tendency parameter (like median and mode) of the distribution of ordinal responses associated with an item. We take a different approach: we look into the ordinal responses that the respondents have given to this item and take a (three-way-)decision on what the most representative cumulative level for that item is.

To this aim we propose to perform a procedure employing a statistical method to establish the most representative region, denoted as *high level*,[5] *uncertain level* and *low level*, for each variable.

At first, let us recall that the collection of all of the responses of a questionnaire is $S = \{s_1, \ldots, s_n\}$, where each respondent's response is represented as $s_j = \langle x_1^j, \ldots x_k^j \rangle$, with $x_i^j \in V \cup \{NULL, DK\}$. We now select some of the k responses or a composition of responses obtained according to the procedure described in Section 2.2. Let us suppose to have l such variables.

Example 3.1. In the case study on conference (see Section 4.1), we have $l = 21$. Indeed, we want to assess the composite variable *overall quality* (see Example 2.2) for all the 21 conferences.

In order to assess these l items, the following steps are performed.

1. For i from 1 to l, that is for each variable (question or composition) collect all the n values $\{x_i^j : j = 1 \ldots n\}$ that can be found in the corresponding columns of S.
2. For each case from $j = 1$ to n, count the number of times that x_i^j is lower than (or equal to) a threshold *neg* and the number of times that is higher (or equal to) a threshold *pos*. Then, neg_i is the number of times the respondents associated a negative assessment to the i-th variable; pos_i the number of times the respondents associated a positive assessment to the variable.
   If the cardinality of V is 6 as assumed above, there are three ways to define *pos* and *neg*. If the researchers believe that central tendency bias is not negligible and/or the number of *uncertain respondents* (that is people assigning the variable

[5] According to the conceptual dimension related to the item, the generic term level can be substituted by a more specific one, like quality, plausibility, appropriateness, and the like.

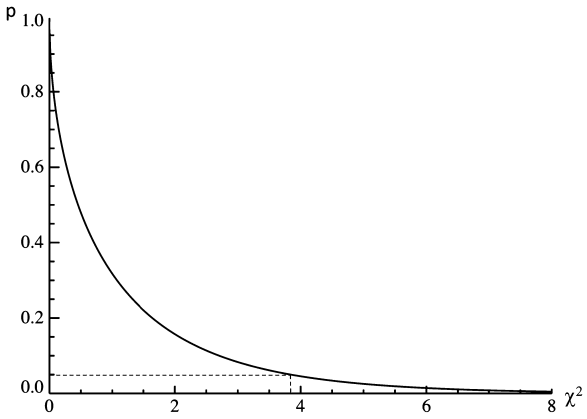

Fig. 3. Chi-squared distribution for 1 degree of freedom (χ^2 is on the *x*-axis, the *P*-value on the *y*-axis). The dotted line crossing the *x*-axis indicates the threshold value for the observed significance level of .05. Intuitively the *P*-value indicates the probability of having collected the responses that were actually collected (or more extreme responses than those ones) if it is assumed that no three-way partition could be performed.

either v_3 or v_4) is high, then $neg = v_2$ and $pos = v_5$; if they believe that acquiescence bias[6] can affect the collected responses, then $neg = v_3$ and $pos = v_5$ (in short, only the v_4s are discarded). In all of the other cases, $neg = v_3$ and $pos = v_4$. In the following we will assume that no biases are present and use this last one.

3. Perform a Pearson's chi-squared test (or a binomial test if the former cannot be applied) on the null hypothesis that there is no significant difference between pos_i and neg_i, besides the difference due to chance. At a 95% confidence level and a significance level[7] set conventionally at .05 (or .01) this means to calculate the chi-squared test statistic χ^2 for each variable with the following formula.

$$\chi^2 = \frac{(obs_1 - E_{obs_1})^2}{E_{obs_1}} + \frac{(obs_2 - E_{obs_2})^2}{E_{obs_2}} \tag{1}$$

where $obs_1 = pos$, $obs_2 = neg$, E_{obs_1} and E_{obs_2} are the number of *expected* observations ($E_{obs_1} = E_{obs_2} = n/2$ if $neg = v_3$ and $pos = v_4$). If the value of χ^2 is greater than the critical value of the chi-squared distribution, that is either $\chi^2 = 3.841$ or $\chi^2 = 6.635$ for the significance level of .05 and .01 respectively (see Fig. 3), the difference between pos_i and neg_i can be considered statistically significant.

4. Partition the initial set of l variables into three areas. One area, denoted as "high level", is where to put the variables for which the corresponding pos_i was significantly higher than the corresponding neg_i. Another area, denoted as "low level", is where to put variables for which neg_i was significantly higher than pos_i; the third area, denoted as "uncertain level", is where to put the remaining variables for which pos_i and neg_i were not found to be significantly different, that is for which we fail to reject the null hypothesis mentioned above.
   The reader should notice that each pair of thresholds neg_i and pos_i can be associated with a different three-way partition. The choice of the three-way partition to be considered as informative depends on the research context, scope and aims. Generally speaking, this choice can be undertaken according to either *a priori* or *a posteriori* reasoning. In the first case, the researcher can take the respondent-related biases into consideration: for instance, in a scale of 6 values (like ours) he or she can choose *neg* and *pos* so that the two central values are not considered, i.e., $neg \leq v_2$ and $pos \geq v_4$ in order to eliminate central tendency bias; or he/she can consider $neg \leq v_4$ and $pos \geq v_5$ in order to minimize acquiescence bias. Alternatively, researchers can reason *a posteriori*: they can produce $3n + 1$ threshold-pairs (and hence tri-partitions) and then assign a variable to a single partition by either majority or according to the partition with the *lowest* P-value (see above). In our case studies, we reasoned according to *a priori* considerations and hence $n = 0$.
   The procedure described above can be applied iteratively, using monotonic thresholds, i.e., $neg' < neg$ and $pos' > pos$, so that the interval $[neg, pos]$ increases monotonically within the ordinal scale. In so doing, the uncertain level can be partitioned in three further areas: significantly tending towards the low/high level, and uncertain tendency. This means that we can follow a sequential three-way decision method [14] in order to refine our classification.

3.2. Method of three-way item ranking

Starting from a set of l variables, each representing either a set of answers of the sample to a single item, or the composition of some items performed with the method described in Section 2.2, we want to produce a ranking of these

[6] This is the response bias that is created when the respondents to a survey have a tendency to agree with the statements, especially when in doubt, or to indicate the positive answers, also for the so-called "social desirability" tendency.

[7] That is the probability of rejecting the null hypothesis given that it is true, also denoted as α.

variables. This could be considered a complementary information to extract from the responses with respect to the absolute assessment of each item described in Section 3.1.

Usual ways to rank items entail to compute and rank their corresponding central tendency parameters (e.g., means, medians and modes) and then rank the corresponding items accordingly. Like in the case of the item absolute assessment, this can be plainly wrong (for instance, in the case of means see [15]) or incapable to distinguish whether any detected difference is due to chance or not. Like in the case of the absolute assessment (see Section 3.1), we rather propose to employ a three-way strategy based on statistical significance to put a variable in either the higher/lower partition or in the uncertain region between them.

Now, given a set of l variables (for instance corresponding to the overall quality of the 21 conferences in the previous examples), and the answers by n users to these variables, as usual denoted as $S = \{s_1, \ldots, s_n\}$, $s_j = \langle x_1^j, \ldots x_l^j \rangle$ the ranking procedure consists of the following steps.[8]

1. Count the number of times each variable x_i $(i = 1, \ldots l)$ is ranked first, second, third (and so forth) among the variables x_i^j, according to the "standard competition ranking" strategy applied to the values (different from NULL/DK) of all tuples s. This is a strategy by which features that compare equal receive the same ranking number, and a gap is left in the ranking numbers (such as "1224" where two items are ranked second, and so the next one is ranked fourth – indeed this strategy is also known as the "1224" strategy);

   Example 3.2. In the conference example, let us suppose to have only three conferences evaluated by four users, hence $l = 3$ and $n = 4$. The answers of the respondents are: $s_1 = \langle v_2, v_3, v_6 \rangle$, $s_2 = \langle v_3, v_6, v_6 \rangle$, $s_3 = \langle NULL, v_3, v_4 \rangle$ and $s_4 = \langle v_4, v_3, v_1 \rangle$. Hence, the ranking of each user is respectively the following: $rank(s_1) = (x_3, x_2, x_1)$; $rank(s_2) = (x_3 = x_2, x_1)$; $rank(s_3) = (x_3, x_2)$; $rank(s_4) = (x_1, x_2, x_3)$.

2. Normalize the sum of all rankings thus associated with each item, by the number of times that item was actually evaluated (thus creating a sort of *mean rank*);

   Example 3.3. Continuing the previous example, we have for the variable x_1: $ncount(x_1) = ((3 + 3 + 1)/(4 - 1))$, since the first variable was ranked only three times out of four: twice as the third one, and once as the first one.

3. Partition the initial set of k variables into three areas. One area, denoted as "higher value", is where to put the items that should be considered of higher value or priority with statistical significance. Another area, denoted as "lower value", is where to put the items that are of lower value (significantly); the third area, denoted as "uncertain level", is where to put the remaining values (if any) that cannot be put in either the former or the latter area with statistical significance. This three-way partition is accomplished according to a Pearson's chi-squared test (see Equation (1)) – or a binomial test if the former cannot be applied – executed by comparing the number of times each single item was ranked in the first p positions ($p < k$; if $k \geq 6$ we usually make $p = 3$, as if it were a *podium*) and the number of times the same item was not ranked in the first p position, that is was ranked in one of the other m $(m = k - p)$ positions, where in Equation (1) $E_{obs_1} = E_{obs_2} = n/2$ (with n number of times the single items was evaluated), if $p = k/2$, and $E_{obs_1} = n/(k/p)$, $E_{obs_2} = n/(k/(k - p))$ otherwise.
   As a result, we have a ranking of the objects in three levels. Also a finer-grained ranking inside each of these three partitions of the variable set can be produced. This can be obtained either by counting the absolute number of victories of each variable (i.e., the number of times each variable ranked first in the case set S) or, simply, by the mean rank (see above point 2).[9]

4. Case studies

In order to both illustrate and validate the methods presented in the previous sections, we designed two user studies. While our aim was to conceive these studies to be different and regarding complementary concerns, they share the same method of data collection and analysis. In both cases, we collected the opinions of a community of experts through a short, closed-ended questionnaire. The first case study regards the International Rough Set Society (IRSS), a non-profit organization counting more than 600 scholars worldwide whose research interests are somehow related to the Rough Set Theory. We applied the above methods to extract the collective knowledge of that community [10,11] about the perceived quality of

[8] We developed a Python script that implements the method. This is available for free and research-oriented uses at https://github.com/federicocabitza/rankEvaluations, under the Creative Commons license (CC BY-SA).

[9] Usually these two ways produce orders that differ slightly, and can be chosen according to the importance that is attached to "being the preferred" object, even as joint winner with other objects. However, both ways produce an internal order within the three-way partitions that should be taken with a bit of caution: the differences between the single ranks inside each partition are seldom statistically significant (that is, they depend on the respondent sample or can be even due to chance). This ranking internal to each partition is proposed to allow for the convenient, if not informative, representation of the three-way-partition in tabular form.

Table 3
The list of the conferences considered in the IRSS case study.

Acronym	Title	CORE 2014 rank[a]
AAAI	Conference on Artificial Intelligence	A*
CIIT	Communications-Internet-and Information Technology	n/a
FSKD	Fuzzy Systems and Knowledge Discovery	C
FUZZ	IEEE International Conference on Fuzzy Systems	A
GRC	IEEE International Conference on Granular Computing	C
IAT	IEE/WIC/ACM International Conference on Web Intelligence	B
IC3K	International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management	C
ICCI	IEEE International Conference on Cognitive Informatics	C
ICML	International Conference on Machine Learning	A*
ICPR	International Conference on Pattern Recognition	B
ICTAI	International Conference on Tools with Artificial Intelligence	B
IJCAI	International Joint Conference on Artificial Intelligence	A*
IJCRS	International Joint Conference on Rough Sets (formerly RSCTC+RSFDGRC+RSKT)	C
IPMU	Information Processing and Management of Uncertainty	C
ISMIS	International Symposium on Methodologies for Intelligent Systems	C
KDD	ACM International Conference on Knowledge Discovery and Data Mining	A*
KES	International Conference on Knowledge-Based and Intelligent Information and Engineering Systems	B
NIPS	Neural Information Processing Systems	A*
PAKDD	Pacific-Asia Conference on Knowledge Discovery and Data Mining	A
PREMI	Pattern Recognition and Machine Intelligence	n/a
SMC	IEEE Conference on Systems-Man and Cybernetics	B

[a] The CORE four ranks, ranging from A* to C, were extracted from the official portal: http://portal.core.edu.au/conf-ranks/. Last visited on Nov. 17, 2016.

their reference conferences. On the basis of their aggregated opinions, a three-way-partitioning is shown to be a feasible and democratic approach to conference ranking that is alternative with respect to those based on either bibliometric analysis or the consensus of small panels and committees.

The second case regards another kind of expert: medical doctors. We proposed difficult clinical cases to two groups of doctors: the residents of the last three years of a postgraduate training program in Orthopedy and Traumatology; and their teachers at the same teaching hospital. We probed these respondents on the plausibility of diagnostic and therapeutic options in order to distinguish between sound and unlikely options with a three-way-partition of a wide list of alternatives. Moreover, we split the overall sample in two random halves: one group could assess the appropriateness of the treatment options along a single ordinal item; respondents from the other group were required to assess the perceived efficiency, effectiveness and safety, which are traditional dimensions related to appropriateness, in addressing a specific condition. This was done to understand if the composition of the above sub-dimensions of appropriateness into a single, comprehensive "appropriateness score" by means of the method described in Section 2.2 would yield different results from those carried out by using the responses given to the single item denoted as 'appropriateness'.

In what follows we briefly illustrate the two case studies by describing the method first, then reporting the results, and finally we discuss conclusions drawn from the analysis of those results.

4.1. The IRSS study on conference ranking

The questionnaire conceived for the IRSS case study contained some typical questions about the respondent profile: more precisely, gender, IRSS membership, research experience and main areas of research interests. These areas had been selected from the ACM classification (2012)[10] by a panel of IRSS senior members and experts including one of the authors, and respondents could indicate at most three areas as their preferred ones. Then the questionnaire showed a list of 21 conferences that the above panel had selected for their relevance in the IRSS community (see Table 3). For each conference listed therein the questionnaire would ask the respondent to indicate on an ordinal scale (from 1 to 6):

- the perceived degree of proximity of their research interests to the main recurring topics of the conference (on a scale from 1 'very far' to 6 'very close');
- the general level of quality of the papers that are usually published in the related proceedings (on a scale from 1 'very low quality' to 6 'very high quality');
- the selectivity of the conference, that is how much "hard" is usually to have a paper accepted there on a scale from 1 'very low selectivity' to 6 'very high selectivity');
- whether they have published at least one full research paper in that conference in the last 5 years (possible answers: Yes, No, I do not remember);
- whether they have been member of the Program Committee at least one time in the last 5 years (possible answers: Yes, No, I do not remember);

[10] http://www.acm.org/about/class/2012. Last visited Nov. 17, 2016.

Table 4
The results for the conference case study.

	Conference	Ranking level$^{(sig.)}$	Quality level$^{(sig.)}$	Median mode
1	AAAI	Higher$^{(***)}$	High$^{(***)}$	5, 5
2	ICML	Higher$^{(***)}$	High$^{(***)}$	5, 5
3	NIPS	Higher$^{(***)}$	High$^{(***)}$	5, 5
4	KDD	Higher$^{(***)}$	High$^{(***)}$	5, 5
5	IJCRS	Higher$^{(***)}$	High$^{(***)}$	5, 5
6	ICPR	Higher$^{(***)}$	High$^{(***)}$	5, 5
7	FUZZ	Higher$^{(**)}$	High$^{(***)}$	4, 5
8	IJCAI	Higher$^{(**)}$	High$^{(***)}$	5, 5
9	PREMI	Higher$^{(*)}$	High$^{(***)}$	4, 5
10	PAKDD	Uncertain$^{(NS)}$	High$^{(***)}$	4, 5
11	FSKD	Uncertain$^{(NS)}$	High$^{(***)}$	4, 5
12	GRC	Uncertain$^{(NS)}$	High$^{(**)}$	4, 5
13	SMC	Uncertain$^{(NS)}$	High$^{(**)}$	4, 5
14	IPMU	Uncertain$^{(NS)}$	High$^{(**)}$	4, 3
15	ISMIS	Uncertain$^{(NS)}$	High$^{(**)}$	4, 3
16	ICCI	Uncertain$^{(NS)}$	High$^{(***)}$	4, 5
17	IAT	Uncertain$^{(NS)}$	High$^{(**)}$	4, 3
18	IC3K	Uncertain$^{(NS)}$	High$^{(***)}$	4, 4
19	KES	Uncertain$^{(NS)}$	Uncertain$^{(NS)}$	4, 3
20	ICTAI	Lower$^{(*)}$	Uncertain$^{(NS)}$	3, 3
21	CIIT	Lower$^{(*)}$	Uncertain$^{(NS)}$	3, 3

- the degree of agreement with the related rank assigned in 2014 by the Computing Research and Education Association of Australasia (CORE) in renown international conference ranking.

In March 2016, we sent an invitation to fill in an online questionnaire to the email addresses enlisted in the IRSS mailing list, that is 635 addresses. We closed the survey after three weeks from the invitation and one reminder, and collected usable responses from 157 questionnaires, 125 of which were complete in every question and item. This study did not have any census-like ambition; since its main aim was to collect the informed opinion of domain experts on the quality of their main conferences of interest, we considered that the survey would be successful if we had involved more than one hundred respondents from all over the world, as it actually happened.

In the respondent sample almost everyone (97%) was (at the time of the survey) or had been in the past a member of the IRSS. Two respondents out of three were from Asia; one fourth of the respondent sample was from Europe. The other ten percent of the respondents were from the Americas. Four respondents out of five were male (82% vs. 18%). The preferred research areas were various: the most represented areas was 'Data Mining', chosen by half of the sample (53%), then Knowledge representation and reasoning and Machine Learning (both selected by 46% of the respondents); approximately one third of the sample indicated that their research area was Artificial Intelligence and Decision Support Systems (34% and 28% respectively). Discrete mathematics and Logic were keywords chosen by much fewer respondents (11% and 15% respectively). The average expertise of the respondents in these areas was quite high: exactly half of the sample stated to have been working in the main research field indicated above for more than 10 years (7% for more than 25 years, while only 5% for less than 3 years).

In general, the majority of the involved experts considered the CORE ranking adequate (the response "I agree with the assigned rank" was chosen 59% of the times). It is difficult, if not impossible, to estimate the so called acquiescence bias [10] in regard to this response proportion. Even assuming this latter negligible, exactly one third of the times the respondents believed that the actual rank should rather be higher (and in particular, the 27% of these times, "much higher") than the rank assigned by CORE, while only in the 7% of the times they believed it should be lower. In particular a large majority of the respondent sample (67%) believed that the IJCRS conference should have been ranked higher (now C in CORE),[11] while this was the case for approximately half of the sample also for FSKD, GRC and IC3K conferences (all C, see Table 4).

By applying the composition method described in Section 2.2, we created an aggregated variable representing the overall quality of a conference by composing the perceived average quality of its papers[12] and its perceived selectivity. With respect to the procedure described in Section 2.2, each item assumes values in $V = \{v_1, \ldots, v_6\}$ and elements in V are then mapped in $O = \{L, M, H\}$. We have to compose two items, thus $k = 2$ and consequently $|R| = 5$. Finally, the five values in R are mapped back to the six values in V.

[11] Actually in the CORE ranking we can find RSCTC (Rough Sets and Current Trends in Computing), IJCRS is now the major conference on rough sets which brings together four conferences, including RSCTC.

[12] The reader should notice that doing the other way round is common practice, that is to infer the quality of a paper from the venue where it has been accepted. Here we infer the quality of the conference from the perceived quality of the papers that it usually accepts.

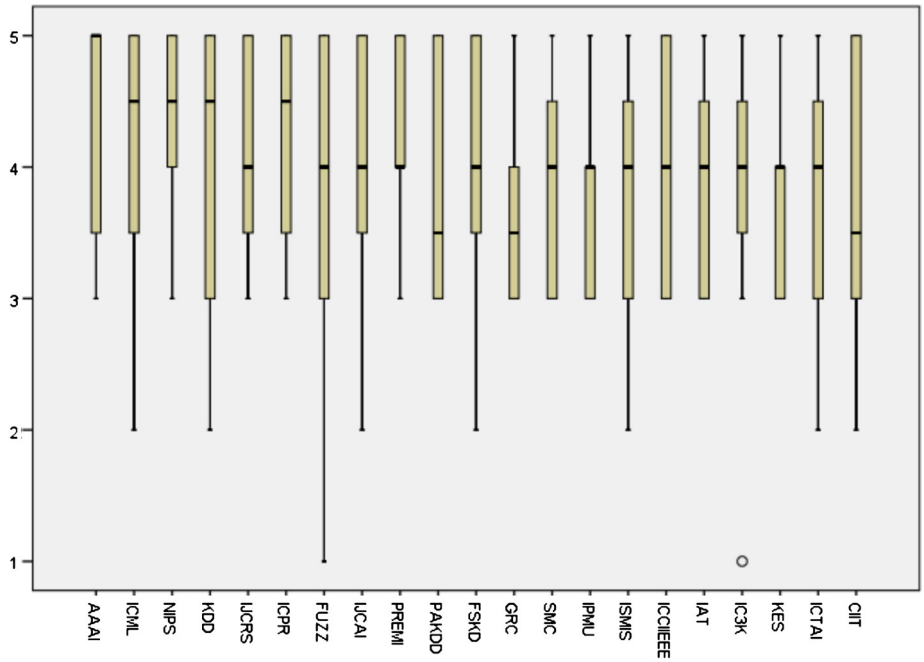

Fig. 4. The boxplots of the ordinal responses for the IRSS conference case study. For each conference, the median values are indicated by the thicker line between the two bars, which indicates the interquartile ranges; whiskers indicate the highest and lowest values.

In column "quality level" of Table 4, we show the result of the three-way-assessment of above overall quality[13] obtained by the method described in Section 3.1. With respect to the assessment algorithm, in this case we have that $l = 21$ (the number of conferences) and $n = 125$ (the number of respondents).

Since the overall quality is an ordinal variable, the traditional methods to analyze it would entail to compute the central tendency parameters of the response distribution, like the median and the mode. However, in order to rank the conferences this can come out of little use. To make this point clearer, in the rightmost columns of Table 4 we also report the medians and modes of the overall quality: the response selected by the majority of the respondents (i.e., the mode) is 5 for two thirds of the conferences considered; the median is above 3 in 9 cases out of 10. Even looking at the response distribution is of little use (see Fig. 4) to understand what are the best conferences and how to divide them in top, medium and low quality venues. In some extreme cases, as this case study showed, also a three-way absolute assessment accomplished with the method described in Section 3.1 can be of little information (see column Quality level in Table 4). On the contrary, the ranking method described in Section 3.2 is robust enough to detect relative differences among the items, even if their assessment is polarized towards one extreme of the ordinal scale (see column Ranking level in Table 4).

In order to understand the role of the context in these results and detect significant correlations between the considered variables, we selected a subset of conferences from the list reported in Table 3. To this aim, We chose the conferences that were both the most frequently assessed ones (which got between 105 and 123 evaluations) and make the set representative of all of the CORE ranks (i.e., A*, A, B and C[14]). This set encompassed the following conferences: KDD, IJCRS, PAKDD, IC3K, KES, FUZZ. In regard to this representative set, the perceived proximity of conference topics and research interests was positively correlated with the perceived paper quality and selectivity (Alpha = .49** and .31**, respectively). This does not indicate necessarily a conflict of interest. Indeed, this kind of conflict could be ruled out in regard to more delicate contextual factors, like having been members of the conference's Program Committee or being an author of a paper published in its proceedings, which were not found to be significant correlated with either the perceived quality nor the selectivity of the conference. No correlation was found between these latter variables and neither gender nor geographical area. On the other hand, conference selectivity and paper quality resulted to be strongly correlated with each other, as expected, but not redundantly (.55**). This can be interpreted as that these dimensions cover different aspects of an overall quality construct, and hence motivated us to create a score from composing them together. Each of these two dimensions was strongly correlated with this latter overall quality score (.76** and .82**) but the fact that the correlation coefficient is lower than 90%

[13] To express statistical significance we use the usual convention to report three asterisks if the level of observed significance, or P-value is lower than .001, two asterisks if P-value is lower than .01, one asterisk if P-value is lower than .05; NS otherwise. Intuitively, the more the number of the asterisks, the farther is the point from the border of the partition.

[14] From the CORE Website (http://www.core.edu.au/conference-portal; last visited on Nov. 17, 2016): A* is assigned to "flagship conferences, leading venues in a discipline area"; A to "excellent conference, and highly respected in a discipline area"; B to "good conferences, and well regarded in a discipline area"; and C to "other ranked conference venues that meet minimum standards".

comes out on the side of the legitimacy of the composition procedure, as extracting an aspect that is completely intrinsic to neither subcomponent.

Comparing the rankings extracted from our study and the CORE one,[15] we see a matching rate of less than 50% (9 matches out of 19). When the rankings do not match, in most of the cases the IRSS community expressed a better ranking for the conferences under consideration (8 times out of 10). The most notable cases regard the upgrade of IJCRS (from CORE C to our higher level) and the downgrade of PAKD (from CORE A to uncertain level).

That said, a general comparison between the CORE ranking and the ranking that resulted from the IRSS study is a way to assess the *face validity*[16] of this latter study, rather than a way to contest the ranking of the former institution, by observing the plausibility of the outcome. That notwithstanding, single comparisons at the conference level should be taken with caution, although we are aware that they could be useful triggers for reflection and discussion among experts and policy makers: the CORE conference rankings are determined by an executive committee of relatively few people that considers a mix of indicators, among which "citation rates, paper submission and acceptance rates, and the visibility and research track record of the key people hosting the conference and managing its technical program".[17] On the other hand, the IRSS ranking could be considered a more collective exercise, involving anonymous experts in Rough Set Theory and leveraging the subjective perception of what is considered valuable in a scientific community from the perspective of its knowledgeable members (that is of the people who write, read and review the papers that are presented in those conferences and who actually attend them[18]).

4.2. The Orthopedy Post Graduate School study, on medical diagnosis and therapy ranking

The questionnaire conceived for the Orthopedy Post Graduate School study contained just two questions regarding the respondent profile: work experience (in terms of years practicing) and gender. The former item also allowed to distinguish between school teachers and residents. In its first page the questionnaire presented a clinical case through a standard summary regarding a 55-year woman suffering from right shoulder pain for one month. This case summary (which we do not report as clearly out of the paper's scope) had been previously made up by three senior professors collaborating with one of the authors by combining some real difficult cases to make the case a real conundrum and a border-line case to be presented to test the skills of the residents. Then, the questionnaire presented a number of examinations that the respondents had to assess on a six-value semantic differential from 'very low' to 'very high' and along four complementary dimensions: informativity, cost, risk, trouble for the patient to carry out the test. Our intention was to compose these dimensions in a recommendability variable (using the composition variable method with $k = 4$), increasing informativity (for the case at hand) and minimizing the other dimensions. Then, the questionnaire presented six alternative diagnoses ($l = 6$ in both three-value decision methods) to be evaluated in regard to their plausibility for the case at hand on a six-value semantic differential from 'very unlikely' to 'very plausible'. The questionnaire then asked the respondents to assume a particular diagnosis to be the correct one and to evaluate the appropriateness of nine alternative treatments ($l = 9$). One half of the respondents (selected randomly) had to evaluate appropriateness directly through a single item depicting an ordinal scale from 1 (very inappropriate) to 6 (very appropriate); the other half indirectly, through three ordinal scales to evaluate the safety, efficiency and effectiveness of each treatment.

The aim was to understand if a group of doctors of varying expertise could be consulted to prioritize exams, identify plausible diagnoses and rank these, identify appropriate treatments and rank them, even for difficult and purposely border-line cases. The last question asked the respondents how difficult they had found the case on a scale from 1 (trivial) to 6 (very difficult).

In March 2016, we sent an invitation to fill in the online questionnaire to the email addresses provided by the Post Graduate School secretariat: these were of 36 post-graduate students and 26 teachers in Orthopedy and Traumatology. We closed the survey after three weeks from the invitation and one reminder to completion, and collected usable responses from 54 questionnaires, 52 of which were complete in every question and item (thus, $n = 52$ in the three-way decision methods). The sample was evenly split between residents and senior specialists (26 vs. 26). Four respondents out of five were male doctors. Almost 9 out of 10 teachers declared to have been specialists for more than 15 years. The case was found slightly more difficult by the residents than by the teachers (as it was expected) but the difference was not statistically significant ($t(50) = .392$, $p = .7$). The results are shown in Tables 6, 7 and 9. In these tables, we show the result of the three-way-ranking method (see the column Ranking level, where we distinguish between the higher partition, the uncertainty region, and the lower partition); and the three-way-assessment of the pertinent dimension (in those tables see the rightmost column, where we distinguish between the high partition, the uncertainty region, and a low partition).

[15] Our study provides three levels, while the CORE four. For this reason we considered only the conferences that were considered of either high or uncertain quality by the IRSS community (see Table 4) and considered A* and A CORE conferences assimilated to the IRSS higher level, B to the uncertain level and C to the lower one.

[16] This is commonly intended as the degree to which an assessment procedure appears effective (usually to experts, even in a subjective way) in terms of coverage of the concept it aims to measure.

[17] CORE Conference Portal, http://www.core.edu.au/conference-portal; last visited Nov. 17, 2016; permanently archived at http://archive.is/fB0vz.

[18] For a longer discussion on this topics, the interested readers can see [11].

Table 5
The examination ranking, considering only their informativity.

	Exams	Ranking level$^{(sig.)}$	Recommendability$^{(sig.)}$
1	Magnetic resonance imaging	Higher$^{(***)}$	High$^{(***)}$
2	Arthro-magnetic resonance imaging	Higher$^{(***)}$	High$^{(*)}$
3	RX	Higher$^{(***)}$	High$^{(**)}$
4	Ultrasound	Higher$^{(*)}$	Uncertain$^{(NS)}$
5	Lab exams	Lower$^{(*)}$	Low$^{(***)}$
6	Computerized axial tomography	Lower$^{(*)}$	Low$^{(**)}$

Table 6
The examination ranking considering the composition of informativity and negative factors.

	Exams	Ranking level$^{(sig.)}$	Recommendability$^{(sig.)}$
1	Arthro-magnetic resonance imaging	Higher$^{(***)}$	High$^{(***)}$
2	Magnetic resonance imaging	Higher$^{(***)}$	High$^{(***)}$
3	Computerized axial tomography	Higher$^{(**)}$	Uncertain$^{(NS)}$
4	RX	Higher$^{(**)}$	Uncertain$^{(NS)}$
5	Ultrasound	Higher$^{(***)}$	Uncertain$^{(NS)}$
6	Lab exams	Uncertain$^{(NS)}$	Low$^{(***)}$

Table 7
The diagnosis ranking for the orthopedy case study.

	Diagnoses	Ranking level$^{(sig.)}$	Plausibility$^{(sig.)}$
1	Supraspinatus tendonitis	Higher$^{(***)}$	High$^{(***)}$
2	Periarthritis	Higher$^{(***)}$	High$^{(***)}$
3	Isolated rupture of supraspinatus and tendinopathy	Higher$^{(***)}$	High$^{(***)}$
4	Adhesive capsulitis	Higher$^{(**)}$	Uncertain$^{(NS)}$
5	Massive rotator cuff tears	Uncertain$^{(NS)}$	Low$^{(***)}$
6	Polymyalgia rheumatica	Lower$^{(*)}$	Low$^{(***)}$

The results were then reviewed by two expert orthopedists to understand if the composition and the ranking were reasonable and sound with respect to their knowledge. They found the group of experts involved capable of converging on the most recommendable, plausible, and appropriate options (for the exams, diagnoses and therapies, respectively).

A remarkable finding regards the fact that the group of residents reached the same conclusions of the senior specialists: thus collectively the residents' performance, that is the choice of doctors with less than 3 years of practice, is equivalent to that of the teachers, that is expert surgeons with more than 15 years of experience. This suggests that a *second opinion*[19] service could be organized by involving a relatively small sample of young doctors (like residents are) because their collective decisions, extracted with the methods presented in this paper, are equivalent to the advice that can be extracted by a group of expert senior doctors, thus coming out on the side of the cost-effectiveness and reliability of such a service. Moreover, the residents assessed the least plausible diagnoses much more plausible than the teachers did, namely Massive rotator cuff tears and polymyalgia rheumatica ($p = .030$ and $p = .006$ in a Mann Whitney test comparing mean ranks between residents and teachers). This can be interpreted as an evidence that traditional ways to assess ordinal variables through central tendency parameters can be misleading, and that the ranking method presented in Section 3.2 can help distinguish relative ranking better, much alike in the conference case study, the relatively good marks that the conferences received, with few exceptions, did not affect the representativeness of the ranking.

In regard to the utility of the composition method presented in Section 2.2, we notice some differences in ranking and absolute value in Tables 5 and 6. This could be interpreted in favor of considering the negative side of exam prescriptions (that is costs, risks and annoyance for the patient) in prioritizing exams to do.

For the treatment ranking we can notice two different phenomena. The first two treatments in the rankings reported in Tables 8 and 9 are the only ones that got a positive appropriateness with statistical significance. In regard to these, we did not detect any difference between the appropriateness assessed by means of a single item and the appropriateness assessed as composition of its traditional dimensions, that is efficiency, effectiveness and safety. However, for the other treatments, which were not considered significantly appropriate, we detected a significant difference between the group that was presented with the single item, and the group that was presented with the three subscales from which then we created the composite score (with p values ranging from .002 to .009 after a Mann Whitney test): the former group expressed a significant *lower* perception of appropriateness. This could be interpreted in favor of considering relevant dimensions, like

[19] A second opinion service [16] is a service by which either a patient or a doctor can present a medical case to another doctor to get her opinion as an external consultant.

Table 8
The treatment ranking, in the single item group.

	Treatments	Ranking level$^{(sig.)}$	Appropriateness$^{(sig.)}$
1	Instrumental physical therapies and assisted physiokinesis therapy	Higher$^{(***)}$	High$^{(*)}$
2	Intra-articular injections and assisted physiokinesis therapy	Higher$^{(***)}$	High$^{(**)}$
3	Arthroscopic capsular release	Uncertain$^{(NS)}$	Uncertain$^{(NS)}$
4	Hyaluronic acid intra-articular injections and assisted physiokinesis therapy	Uncertain$^{(NS)}$	Low$^{(**)}$
5	Arthrotomic capsular release	Lower$^{(***)}$	Low$^{(***)}$
6	Arthroscopic needling	Lower$^{(***)}$	Low$^{(***)}$
7	Inverse shoulder prostheses	Lower$^{(***)}$	Low$^{(***)}$
8	Arthrotomic suture of rotator cuff	Lower$^{(***)}$	Low$^{(***)}$
9	Arthroscopic suture of rotator cuff	Lower$^{(***)}$	Low$^{(***)}$

Table 9
The treatment ranking, in the composed items group.

	Treatments	Ranking Level$^{(sig.)}$	Appropriateness$^{(sig.)}$
1	Instrumental physical therapies and assisted physiokinesis therapy	Higher$^{(***)}$	High$^{(***)}$
2	Intra-articular injections and assisted physiokinesis therapy	Higher$^{(***)}$	High$^{(**)}$
3	Hyaluronic acid intra-articular injections and assisted physiokinesis therapy	Higher$^{(*)}$	Uncertain$^{(NS)}$
4	Arthroscopic capsular release	Higher$^{(**)}$	Uncertain$^{(NS)}$
5	Arthroscopic needling	Uncertain$^{(NS)}$	Uncertain$^{(NS)}$
6	Arthroscopic suture of rotator cuff	Uncertain$^{(NS)}$	Low$^{(**)}$
7	Arthrotomic capsular release	Uncertain$^{(NS)}$	Low$^{(***)}$
8	Inverse shoulder prostheses	Uncertain$^{(NS)}$	Low$^{(**)}$
9	Arthrotomic suture of rotator cuff	Lower$^{(*)}$	Low$^{(***)}$

appropriateness, as composed by other related and more specific dimensions on which to probe the respondents, because otherwise the risk of underestimation of the overall concept is concrete.

5. Conclusions

In this paper we have presented two methods by which to extract indications from the collective knowledge [10] of even large communities of experts by means of structured questionnaires that encompass ordinal scales and two-value (e.g., yes–no) questions [9] to represent the respondents' attitudes, perceptions and opinions. These two methods adopt a three-way-decision strategy to both assess and rank the items expressed in the questionnaires and characterized by the experts in terms of ordinal values.

We presented the methods formally, and then illustrated their application to two purposely challenging case studies that involved two real communities of experts. One case, the IRSS conference study, was challenging both for the number and heterogeneity of the experts involved and for the notorious vagueness of the object of study, that is the quality of scientific international conferences and their ranking. The latter case was challenging for the relevance of the potential impact: detecting the best course of action to address a health problem (i.e., what exam would optimize informativity while also taking costs and risks into account), to identify its causes (what is the most plausible diagnosis even when only a few textual indications are available, like in second opinion services), to manage its treatment (what is the most appropriate therapy, by optimizing efficiency and effectiveness?).

These are just two examples of a three-way approach to a qualitative assessment and ranking of ordinal variables. We refer to the qualitative nature of this kind of assessment and ranking to highlight the reliance on the sensitivity and expertise of people involved through soundly designed structured questionnaires [9] and not to overlook the importance of quantitative analyses, which rather our approach extensively use in terms of statistical inference techniques.

These user-driven and three-way methods are proposed as alternatives with respect to methods providing parametric-based assessment and ranking with a clear-cut and dichotomous classification of the items under consideration. As the user studies presented in this work show, this approach is feasible, applicable also to relatively small data sets, and looks preferable in those cases where there is a need to involve domain experts and tap in their informed and knowledgeable opinion in debates and subjective (or hard to objectify) matters (like in the case of conference ranking [11,17], or in the case of idea prioritization [18]); or when it is important to get as many informed opinions as possible (as well as in a cost-effective manner) to improve the efficiency and effectiveness of decision making and the related activities in critical domains (as healthcare and medical work is [10]).

In this line, the methods presented in this work could also be applied in initiatives that involve a large number of people who do not necessarily share any membership to the same community, like in crowdsourcing Internet marketplaces (e.g., Amazon Mechanical Turk[20]), in which questions are asked to identify the "right" answer, both when this latter one exists,

[20] https://www.mturk.com/.

and even when no one (individually) still knows this answer [19], but a group of people can find and discover it, in order to realize the full potential of what is generally called *collective intelligence* [20].

Acknowledgements

The authors would like to thank Prof. Giuseppe Peretti, head of the Post Graduate School in Orthopedy and Traumatology of the University of Milano, and Mrs. Manganaro from the School secretariat, for the continuous help and support in achieving the highest participation by the residents and teachers of their school in our case study. We are also grateful to Andrea Casella, MD, who first drafted the case summaries, and to Prof. Pietro Randelli, who later revised them and the questionnaire. We are also grateful to all of residents and teachers who willingly volunteered in filling in the questionnaire and giving us their informed opinions. We also thank the members of the Executive Board of the IRSS (namely, Chris Cornelis, Guoyin Wang and Yiyu Yao) for their availability and willingness to help, and all of the respondents of the conference ranking survey for believing in the value of the initiative.

References

[1] Y. Yao, An outline of a theory of three-way decisions, in: J. Yao, Y. Yang, R. Slowinski, S. Greco, H. Li, S. Mitra, L. Polkowski (Eds.), Proc. RSCTC 2012, in: Lect. Notes Comput. Sci., vol. 7413, 2012, pp. 1–17.
[2] Y. Yao, Rough sets and three-way decisions, in: D. Ciucci, G. Wang, S. Mitra, W. Wu (Eds.), Proc. RSKT 2015, in: Lect. Notes Comput. Sci., vol. 9436, 2015, pp. 62–73.
[3] Y. Yao, Three-way decision: an interpretation of rules in rough set theory, in: P. Wen, Y. Li, L. Polkowski, Y. Yao, S. Tsumoto, G. Wang (Eds.), Proc. RSKT 2009, in: Lect. Notes Comput. Sci., vol. 5589, 2009, pp. 642–649.
[4] H. Fujita, T. Li, Y. Yao, Advances in three-way decisions and granular computing, Knowl.-Based Syst. 91 (2016) 1–3.
[5] Y. Yao, C. Gao, Statistical interpretations of three-way decisions, in: D. Ciucci, G. Wang, S. Mitra, W. Wu (Eds.), Proc. RSKT 2015, in: Lect. Notes Comput. Sci., vol. 9436, 2015, pp. 309–320.
[6] L.A. Geer, B.A. Curbow, D.H. Anna, P.S. Lees, T.J. Buckley, Development of a questionnaire to assess worker knowledge, attitudes and perceptions underlying dermal exposure, Scand. J. Work, Environ. & Health (2006) 209–218.
[7] P.M. Herr, Priming price: prior knowledge and context effects, J. Consum. Res. 16 (1) (1989) 67–75.
[8] M.C. Kaptein, C. Nass, P. Markopoulos, Powerful and consistent analysis of likert-type ratingscales, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM, 2010, pp. 2391–2394.
[9] F. Cabitza, A. Locoro, Questionnaires in the design and evaluation of community-oriented technologies, Int. J. Web Based Commun. 13 (2017) 1.
[10] F. Cabitza, C. Simone, Investigating the role of a web-based tool to promote collective knowledge in medical communities, Knowledge Management Research & Practice 10 (4) (2012) 392–404.
[11] F. Cabitza, A. Locoro, Exploiting the collective knowledge of communities of experts: the case of conference ranking, in: IC3K 2015, Proceedings of the 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management, vol. 3, Lisbon, Portugal, November 12–14, 2015, 2015, pp. 159–167.
[12] S. Stevens, On the theory of scales of measurement, Science 103 (2684) (1946) 677–680.
[13] E. Huizingh, Applied Statistics with SPSS, Sage, 2007.
[14] Y. Yao, Granular computing and sequential three-way decisions, in: P. Lingras, M. Wolski, C. Cornelis, S. Mitra, P. Wasilewski (Eds.), Proc. RSKT 2013, in: Lect. Notes Comput. Sci., vol. 8171, 2013, pp. 16–27.
[15] J. Robertson, Likert-type scales, statistical methods, and effect sizes, Commun. ACM 55 (5) (2012) 6–7.
[16] H.S. Ruchlin, M.L. Finkel, E.G. McCarthy, The efficacy of second-opinion consultation programs: a cost-benefit perspective, Med. Care (1982) 3–20.
[17] M. Saarela, T. Kärkkäinen, T. Lahtonen, T. Rossi, Expert-based versus citation-based ranking of scholarly and scientific publication channels, J. Informetr. 10 (3) (2016) 693–718.
[18] C. Riedl, I. Blohm, J.M. Leimeister, H. Krcmar, Rating scales for collective intelligence in innovation communities: why quick and easy decision making does not get it right, in: Proceedings of Thirty First International Conference on Information Systems, 2010.
[19] E. Bonabeau, Decisions 2.0: the power of collective intelligence, MIT Sloan Manag. Rev. 50 (2) (2009) 45.
[20] A. Kittur, B. Smus, S. Khamkar, R.E. Kraut, CrowdForge: crowdsourcing complex work, in: Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology, ACM, 2011, pp. 43–52.