Scientometrics
DOI 10.1007/s11192-015-1770-8

Research project evaluation and selection: an evidential reasoning rule-based method for aggregating peer review information with reliabilities

Wei-dong Zhu[1] · Fang Liu[2,3] · Yu-wang Chen[3] · Jian-bo Yang[2,3] · Dong-ling Xu[2,3] · Dong-peng Wang[2]

Received: 19 November 2014

Abstract Research project evaluation and selection is mainly concerned with evaluating a number of research projects and then choosing some of them for implementation. It involves a complex multiple-experts multiple-criteria decision making process. Thus this paper presents an effective method for evaluating and selecting research projects by using the recently-developed evidential reasoning (ER) rule. The proposed ER rule based evaluation and selection method mainly includes (1) using belief structures to represent peer review information provided by multiple experts, (2) employing a confusion matrix for generating experts' reliabilities, (3) implementing utility based information transformation to handle qualitative evaluation criteria with different evaluation grades, and (4) aggregating multiple experts' evaluation information on multiple criteria using the ER rule. An experimental study on the evaluation and selection of research proposals submitted to the National Science Foundation of China demonstrates the applicability and effectiveness of the proposed method. The results show that (1) the ER rule based method can provide consistent and informative support to make informed decisions, and (2) the reliabilities of the review information provided by different experts should be taken into account in a rational research project evaluation and selection process, as they have a significant influence to the selection of eligible projects for panel review.

Keywords Research project evaluation and selection · Evidential reasoning · Reliability · Confusion matrix

✉ Fang Liu
liu_fang2014@163.com

[1] School of Economics, Hefei University of Technology, 193 Tunxi Road, Hefei 230009, Anhui, China

[2] School of Management, Hefei University of Technology, 193 Tunxi Road, Hefei 230009, Anhui, China

[3] Manchester Business School, The University of Manchester, Manchester M15 6PB, UK

Published online: 13 October 2015

Introduction

Research project evaluation and selection is a common and significant task for many companies and research funding agencies. Its main objective is to determine appropriate projects for implementation (Mahmoodzadeh et al. 2007). However, the rapidly changing pace of technology development, together with the increasing complexity, has made the research project evaluation and selection a challenging decision making process (Solak et al. 2010; Tavana et al. 2013). Particularly, it involves multiple evaluation criteria and multiple peer review experts or decision makers. In addition, both quantitative and qualitative assessment criteria need to be taken into account simultaneously in the decision making process. Quantitative criteria can be easily assessed by numerical values, while qualitative criteria may be assessed by a set of different linguistic evaluation grades, such as, poor, average, good, and excellent.

Due to the importance of research project evaluation and selection, numerous methods and techniques have been proposed by researchers for evaluating and selecting research projects in the past few decades. The typical methods include peer review (Jayasinghe et al. 2006; Južnič et al. 2010), fuzzy logic (Coffin and Taylor 1996; Wang and Hwang 2007), fuzzy analytic hierarchy process (AHP) method (Hsu et al. 2003; Huang et al. 2008), the technique for order preference by similarity to ideal solution (TOPSIS) (Mahmoodzadeh et al. 2007; Khalili-Damghani et al. 2013), data envelopment analysis (Linton et al. 2002), and so on.

These studies provide support for decision makers to make informed decisions in the research project evaluation and selection process. However, previous research rarely pays attention to the aggregation of evaluation information, and especially most methods use simple additive methods to add up scores which do not convey rich information to differentiate a large number of projects under evaluation and are lack of the ability to capture the true performance profile of the projects. Thus these may have a negative impact on the quality of the decision made in research project evaluation and selection. In addition, the coordination and aggregation of evaluation information from multiple peer experts often becomes an obstacle in practical research project evaluation and selection, as experts may provide inconsistent and even conflicting evaluation on the same project. Therefore, it is extremely necessary to take into account the reliability of the review information in a sensible way, but most existing aggregation methods failed to do it.

In order to deal with these issues, the paper presents an effective method for evaluating and selecting research projects by using the recently-developed evidential reasoning (ER) rule (Yang and Xu 2013). The proposed ER rule based evaluation and selection method includes the following main components: belief distributions to represent review information, a confusion matrix to generate experts' reliabilities, utility based information transformation to handle evaluation criteria with different evaluation grades, and aggregation of multiple experts' evaluation information on multiple criteria using the ER rule. A case study is conducted to demonstrate the applicability and effectiveness of the proposed ER rule based evaluation and selection method.

The remainder of the paper is organized as follows. "Literature review" section reviews previous research on project evaluation and selection. "Problem description and formulation" section briefly introduces the research background and problem formulation. "The ER rule for project evaluation" section presents the application of the recently-developed evidential reasoning (ER) rule for evaluating and selecting research projects. "An experimental study" section conducts a case study to illustrate the proposed ER rule based method and compares the results with the existing ones on the research project evaluation

and selection of the National Science Foundation of China (NSFC). "Conclusions" and perspectives of the proposed ER rule are given in the last section.

Literature review

The evaluation of a research project can be divided into three different stages: ex-ante evaluation, monitoring and ex-post evaluation. The evaluation criteria, as well as evaluation approaches, are usually different for the three different stages of evaluation (Bulathsinhala 2014). The ex-ante evaluation is conducted before project start-up, while monitoring is for an ongoing project, and the ex-post evaluation performs an assessment to a project after it has fully completed (Olsson et al. 2010). In relation to different context for project evaluation, studies have shown that project evaluation usually involves complex processes in large organizations (Oral et al. 2001). This is the case particularly for explorative research projects due to the complexity of the objectives of the program and the phase of development of the technology (Olsson et al. 2010; Horrobin 1996). In this work, the evaluation and selection are regarded as the objective assessment of a research project and the aggregation of evaluation information for supporting project selection at the stage of ex-ante evaluation.

There are a wide range of studies on research project evaluation and selection. One important question in research project evaluation and selection is to determine the criteria for project evaluation. Traditional evaluation and selection criteria are mainly about financial benefits and costs which are quantitative in nature, such as net present value and internal rate of return. Although financial aspects are very important for businesses and have been explored widely in project selection, organizations conducting research projects may have different aims and needs, and they should be taken into consideration seriously and be reflected in certain evaluation criteria. These criteria are complex and more difficult to quantify. For example, many public research funding bodies and organizations usually also consider qualitative criteria including science and technology development strategy or corporate strategy, qualitative benefits and risks, desires of different stakeholders when making the project selection (Meade and Presley 2002).

Many methods and techniques have been presented to deal with research project evaluation and selection, which tend to be either qualitative or quantitative. According to Henriksen and Traynor (1999), project evaluation and selection methods can be categorized into unstructured peer review, scoring, mathematical programming, economic models, decision analysis, interactive methods, artificial intelligence, portfolio optimization, etc. An organizational decision support system (ODSS) architecture has been proposed to support R&D project selection from organizational decision-making perspective, which focuses on the whole life cycle of the selection process (Tian et al. 2005). Data envelopment analysis (DEA) has been illustrated to be a useful method for dividing projects into different groups, and it does not require variables to have the same scale or conversion weight. It is an ideal solution for the comparison of research projects that potentially have many different non-cost and non-numeric variables (Linton et al. 2002). Huang et al. (2008) employs fuzzy numbers to represent subjective expert judgments, and the fuzzy analytic hierarchy process method is utilized to identify the most important evaluation criteria. For selecting an appropriate portfolio of research projects, a fuzzy mixed integer programming model for valuing options on R&D projects is developed, and future cash flows are considered to be trapezoidal fuzzy numbers (Carlsson et al. 2007).

When there is uncertain and flexible project information, the method of fuzzy compound options is used to maximize the target portfolio value (Wang and Hwang 2007). To support the selection of projects, a social network-empowered research analytics framework is proposed to capture the social connections and productivity of researchers (Silva et al. 2013). A hybrid project selection model composed of financial considerations, risk analysis and scoring model is studied and a field test is conducted in a small to medium-sized enterprise by Lawson et al. (2006). This work makes it possible to transfer the model into an applicable form for a small engineering company (Lawson et al. 2006).

The basic project evaluation and selection process usually can be carried out through several steps, namely, proposal completion, proposal submission, preliminary examination, peer review, summary of comments, panel review and final decision (Feng et al. 2011; Silva et al. 2014). Previous approaches for R&D project selection analyzed only one of the steps such as proposal clustering, reviewer assignment or automated workflows (Silva et al. 2014). In addition, most previous R&D evaluation studies focused on describing the mechanisms of the techniques and on analyzing their strengths and weakness based on the nature of R&D projects, and very few have gained wide acceptance in the real world situation (Hsu et al. 2003). There are the following inherent limitations in the currently proposed models (1) rarely paying attention to the aggregation of evaluation results. (2) No explicit recognition and incorporation of the reliabilities of review information from multiple experts.

From the above discussion, it is evident that research project evaluation and selection is still a challenging task even with the use of the above methods. However, the evidential reasoning (ER) rule has an inherent capability of aggregating information from multiple decision makers, and it has a great potential to resolve the issues discussed above in the project evaluation and selection process. In the ER rule, different pieces of evidence are associated with weights and reliabilities in the aggregation process. However, the weight and reliability are not differentiated clearly in many aggregation methods (Smarandache et al. 2010; Yang and Xu 2013). The ER algorithm was developed for multiple criteria decision analysis (MCDA) on the basis of Dempster-Shafer evidence theory (Shafer 1976; Yang and Singh 1994), and it has been widely studied and applied by researchers for information aggregation (Yang and Xu 2002).The rule and utility based techniques has also been proposed for transforming various types of information (Yang 2001). The application areas of the ER algorithm covers engineering design; reliability, safety and risk assessment; business management; project management and supply chain management; environmental and sustainability management; smart homes management; policy making and group decision making (Xu 2012). As a generalization of the ER algorithm, the ER rule is a generic probabilistic reasoning process and can be used to combine multiple pieces of independent evidence with both weight and reliability of the evidence considered. It has been proved that the classical Dempster-Shafer evidence theory and the ER algorithm are special cases of the ER rule (Yang and Xu 2013). The ER rule uses a weighted belief distribution with reliability (WBDR) structure for profiling a piece of evidence, which further improves the basic probability assignment process in the Dempster-Shafer evidence theory and the ER algorithm. It further employs the orthogonal sum operation on the WBDR to combine multiple pieces of evidence, in which each piece of evidence can play a limited role relative to its weight and reliability (Yang and Xu 2013). The reliability is used to represent the quality of the information source and its ability to provide the correct assessment or solution for a given problem. The weight is used to reflect the relative importance of a piece of evidence in comparison with other evidence and determined according to who uses the evidence. This means that weight can be subjective and whereas

the reliability is the inherent property of the evidence. The ER rule has the features of managing importance and reliability of sources separately and handling highly or completely conflicting evidence. The ER rule is thus ideally suitable to evaluate a project where a group of experts are involved in providing evaluation information.

Problem description and formulation

As a main program funding body for supporting fundamental research, the NSFC plays an important role in fostering research innovation and development in China. One of the core functions of the NSFC is to perform research project evaluation and selection, and it concerns whether the government's investments in science and technology can be effectively utilized.

Generally the NSFC has an annual call for proposals. In 2013, there were 161,888 research proposals submitted to the NSFC and among which 38,920 were funded. The NSFC has a few research program categories among which the general program is the main one. Under the general program scheme, the NSFC sets up eight funding departments in terms of different disciplines, including mathematical and physical sciences, chemical sciences, life sciences, earth sciences, engineering and materials sciences, information sciences, management sciences, and health sciences.

The structure model of research project evaluation and selection in the NSFC

The research project submission, evaluation and selection process in the NSFC, as shown in Fig. 1, mainly includes seven steps. (1) Principal investigators (PIs) prepare proposals according to the NSFC call for proposals. (2) The supporting institutions collect the applications and submit them to the NSFC correspondingly. (3) Each discipline-specific department of NSFC conducts a preliminary screening, which is of a purely administrative nature and should be completed within 45 days after the submission deadline. (4) The successful proposals from the preliminary screening are sent electronically out for peer review. (5) The program directors in each department consider the review information provided by all the peer reviewers and then make an initial ranking. (6) Then a certain number of highly ranked applications are evaluated again in a panel meeting. (7) Finally, the NSFC council approves the selected applications from the panel meeting. The participants or organizations involved at each step are also listed in Fig. 1.

In reality, some constraints should be taken into consideration for research project selection. For example, the available funds decide how many projects can potentially be

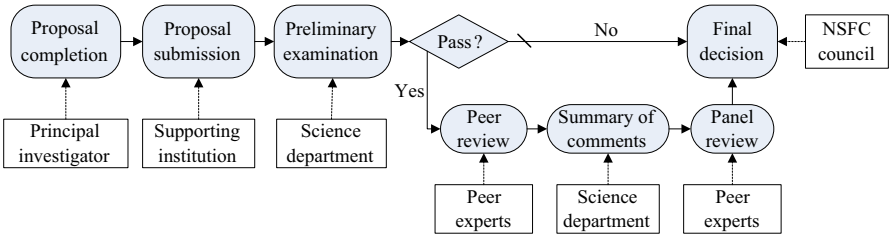

Fig. 1 The NSFC research project submission, evaluation and selection process

funded, and the political and scientific priorities affect the allocation of resources to different research disciplines. Furthermore, in accordance with the provision of the NSFC, PI may not apply more than once per year to any single NSFC programme or should not hold more than three NSFC grants at the same time. In principle, proposals are limited in length to 8000 Chinese characters but in practice this limit is frequently being exceeded without incurring any penalty. In this paper, we aim to develop a systematic method of prioritizing and selecting research projects for panel review through using the peer review information from multiple experts.

Evaluation criteria

In the fourth peer review step, some instructions of NSFC on the selection rules and evaluation criteria are also sent to experts along with the proposals to be reviewed. For the general programme, experts need to consider each proposal from the following five aspects: (1) scientific value and potential for application, (2) academic novelty, (3) research content and its appropriateness, (4) rationality and feasibility of research plan, (5) capacity of research team.

Experts normally review the projects and fill out the review forms according to the instructions provided. The review form has two evaluation criteria related to the project quality in the NSFC research project evaluation and selection scheme, namely "Comprehensive evaluation level" and "Funding recommendation" as listed in Fig. 2. The peer review system of the NSFC is similar to the practice of research councils in the UK and the USA. Taking management sciences department for example, the program directors randomly choose three to five experts from the database of experts working on the same or relevant field for reviewing each project.

Peer reviewers rate projects on the "Comprehensive evaluation level" criterion using a four-point scale, namely Excellent, Good, Average and Poor, while on the "Funding recommendation" criterion by a three-point scale: Fund with priority, Fund, and Not fund. Each expert chooses one grade on each of the two criteria.

Aggregation of results

In order to take into consideration the review information on the two evaluation criteria from multiple experts, the NSFC simply adopts an additive approach. Specifically, in the management sciences department they assign the values of 4, 3, 2 and 1 to the four subjective grades of the "Comprehensive evaluation level" criterion, and the values of 2, 1, and 0 to the three subjective grades of the "Funding recommendation" criterion respectively. Then, the average of the evaluation scores from all experts can simply be calculated on each criterion. The sum of the two average scores represents the overall performance of a project proposal (Chen 2009). After the calculation of the overall score, the projects will be categorized into

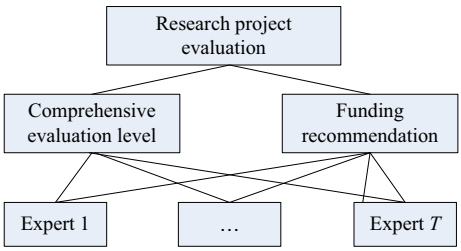

Fig. 2 Hierarchical structural model for research project evaluation in the NSFC

six categories: $A(\geq 4.8)$, within which projects will be considered as funding with priority, $A-$ (≥ 4.6), and B (≥ 4.0), within both of which projects are regarded as potentially fundable, and E, within which projects are regarded as non-unanimous and should be further discussed in panel review meeting. The eligible projects within the above four categories will be forwarded to panel review. There are the other two categories: C within which projects are not eligible for panel review, and D within which projects are for direct rejection. This evaluation method, which is simple and easy to use in practice, has been implemented for many years in the evaluation of the NSFC projects, and it plays an important role to ensure equity and fairness in the research project evaluation and selection process.

Unfortunately, the existing evaluation and selection method does not capture the richness of the actual evaluation information of reviewers and it has two main limitations: (1) the single overall score cannot capture the true performance profile of each project, given that it is evaluated from different perspectives by multiple experts. (2) The limited number of adding up scores lacks the ability to distinguish ten thousands of proposals. There are only twenty-six possible average scores with five valid review forms, and the number of the scores is even smaller with a smaller number of valid forms. Therefore many projects get the same score and it is not effective to differentiate the true quality of proposals.

The ER rule for project evaluation

If the research project selection should be in strict accordance with peer experts' opinions, how to use peer experts' comments becomes a key point. As the proposed projects are of exploratory nature, it is also likely for peer experts to review projects beyond their expertise. In addition, peer experts usually evaluate a project according to several criteria, which may use different evaluation grades. In such circumstances, evaluation information cannot be fully reliable, leading to complexity and uncertainty. In this section, we employ the ER rule to aggregate multiple experts' evaluation information in the NSFC project evaluation process. It mainly includes the following steps: (1) modelling the research project evaluation information on each criterion using a belief structure. (2) Deciding the relative weight of each criterion and the reliability of each piece of evidence, as it may not be fully reliable. (3) Information aggregation to combine multiple experts opinions on each criterion with taking into consideration both expert weight and reliability. (4) Information of aggregated expert opinions on each of the two criteria obtained in step (3) transformation and aggregation to the top criterion; information transformation is necessary because the assessment on each of the two criteria uses a different set of evaluation grades.

Belief structure for review information representation

Let Θ be a set of collectively exhaustive and mutually exclusive hypotheses, which is called the frame of discernment. The hypotheses in the context of project evaluation are the evaluation grade on each criterion, such as "Funding recommendation (for a project)" is to "Fund with priority". A basic probability assignment (bpa), called a belief structure, is a mass function $m{:}2^{\Theta} \rightarrow [0, 1]$. It satisfies the following two conditions:

$$\sum_{A \subseteq \Theta} m(A) = 1, \quad 0 \leq m(A) \leq 1 \tag{1}$$

$$m(\phi) = 0 \tag{2}$$

where ϕ is an empty set, and 2^{Θ} is the power set of Θ. $m(A)$ is a probability mass to A, a subset of Θ, which represents the degree to which the evidence supports A. $m(\Theta)$ is called the degree of ignorance, which measures the probability mass assigned to Θ.

In the research project selection problem, a belief structure can be used to describe subjective evaluation information in a more informative scheme. The review information which each expert gives to each project on each criterion is treated as a piece of evidence. The main notations used in this paper are listed in the following table.

Using the notation of $H_{n,i}$ and $\gamma_{n,i,t}$ in Table 1, the assessment of a project on the ith basic evaluation criterion by the tth expert can be described by the following belief distribution:

$$s(e_{i,t}) = \left\{\left(H_{n,i}, \gamma_{n,i,t}\right)\right\}, \quad i = 1, \ldots, L, \quad n = 1, \ldots, N_i, \quad t = 1, \ldots, T$$
$$\text{with} \quad 0 \le \gamma_{n,i,t} \le 1 \quad \text{and} \quad \sum_{n=1}^{N} \gamma_{n,i,t} \le 1. \tag{3}$$

where $(H_{n,i}, \gamma_{n,i,t})$ is an element of evaluation evidence $s(e_{i,t})$, representing that the evidence points to grade $H_{n,i}$ to the degree of $\gamma_{n,i,t}$. $\left(H_{n,i}, \gamma_{n,i,t}\right)$ will be referred to as a focal element of $s(e_{i,t})$ if $\gamma_{n,i,t} > 0$. L, N_i and T denote the number of basic criteria, the number of evaluation grades on the ith basic evaluation criterion and the number of experts assigned to assess each project respectively.

Calculation of weights and reliabilities

Weight assignment methods

There are a number of ways to elicit the weights of criteria in the context of multiple criteria decision analysis, such as direct assignment, swing weights, pairwise comparisons and analytic hierarchy process (Agarski et al. 2012). The direct assignment method is easy to carry out, and it consists of the following steps: (1) Identify the most important criterion

Table 1 Notations on project evaluation

Notation	Definition
e_i	The ith basic evaluation criterion
a_l	The lth project
$H_{n,i}$	The nth evaluation grade for the ith basic evaluation criterion
$\gamma_{n,i,t}$	The degree of belief to which the ith basic evaluation criterion is assessed to $H_{n,i}$ by the tth expert
H_j	Evaluation grades for the overall evaluation criterion
w_i	The weight of the ith basic evaluation criterion
v_t	The weight assigned to the tth expert
r_t	The reliability of the tth expert
$\tilde{v}_t$	The new weight of a piece of evidence after taking into consideration the tth expert's reliability, and $\tilde{v}_t = v_t/(1 + v_t - r_t)$
$\beta_{j,i}$	The degree of belief to which a project is assessed to the grade H_j on the ith basic evaluation criterion

and assign a weight $\omega_1 (\leq 1)$ to it, and then the remaining weight is reduced to $1 - \omega_1$. (2) Identify the next most important criterion and assign a weight to it out of the remaining weight, denoted by ω_i. The remaining weight is $1 - \sum \omega_i$. (3) Repeat the above steps until each criterion has been assigned with a weight. (4) To check consistency, the process can be re-started from the least important criterion, or indeed from one of any other criterion in principle.

The above methods for calculating weights can be used for generating both weights of criteria and weights of experts.

Confusion matrix for generating experts' reliabilities

In general, the reliability of an information source reflects its ability to provide the correct assessment or solution of the given problem. In this work, the reliabilities of experts can be measured to some extent by their past review performances as many experts have reviewed many projects previously. In the research project evaluation and selection process, peer experts make final recommendation into two main categories, "Fund (including with or without priority)" or "Not fund". The actual funding outcomes also fall into two categories, "Funded" or "Unfunded". As the number of projects in the "Unfunded" category is much higher than that in the "Funded" category and the data set is unbalanced. The ratio of the number of "Funded" or "Unfunded" projects to the total number of projects is not a reliable metric to measure the performance of an expert. For example, if there were in total 20 projects, including 18 "Unfunded" projects and only 2 "Funded" projects in the data set, the expert could easily achieve 90 % reliability by recommending all the projects as "Not fund". Therefore, to overcome the issue of imbalanced data set, we propose to use a confusion matrix (Provost and Kohavi 1998) to measure the reliability of peer review experts. The confusion matrix is a square matrix that represents the count of a classifier's class predictions with respect to the actual outcome on some labeled learning data set. The reliability of a peer expert can be evaluated by a confusion matrix with a two class classifier as shown in Table 2.

The entries in the confusion matrix have the following meanings in this research: TP and FP denote the number of correct and incorrect "Fund" decisions, and FN and TN the number of incorrect and correct "Not fund" decisions respectively, compared with the actual outcomes. If an expert gives a "Fund" decision, the rate of matching the final decisions can be measured by the true positive rate. Otherwise, the true negative rate should be used.

$$\text{True positive rate} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \tag{4}$$

$$\text{True negative rate} = \frac{\text{TN}}{\text{TN} + \text{FN}}. \tag{5}$$

Table 2 A confusion matrix for generating experts' reliabilities

	Expert's decisions	
	Fund	Not fund
Actual outcomes		
Funded	True positive (TP)	False negative (FN)
Unfunded	False positive (FP)	True negative (TN)

This method can be used to measure reliabilities of the review information provided by each expert and can be updated every year. The mean value of all experts' reliabilities can be used for peer review experts who join the review team in the present year, who have had a very small number of proposal reviewed previously, or whose historical review information is not available.

The ER rule for aggregating experts' evaluation information on each criterion

In the ER rule based aggregation method, the basic probability masses for e_i are assigned as follows

$$\tilde{m}_{n,i,t} = \tilde{v}_t \gamma_{n,i,t}, \quad n = 1, \ldots, N_i \quad \text{and} \quad \tilde{m}_{P(\Theta),i,t} = 1 - \tilde{v}_t \tag{6}$$

where $\tilde{v}_t = v_t c_{rw,i}$ and $c_{rw,i} = 1/(1 + v_t - r_t)$. $P(\Theta)$ represents the power set of Θ. $\tilde{m}_{n,i,t}$ measures the degree of support on $H_{n,i}$ from evidence $e_{i,t}$ with both the weight and reliability taken into account. $\tilde{m}_{P(\Theta),i,t}$ is the residual support of evidence $e_{i,t}$ due to its weight and reliability. A weighted belief distribution with reliability can be used to represent a piece of evidence as follows:

$$m = \left\{ \left(H_{n,i}, \tilde{m}_{n,i,t}\right), \quad \forall H_{n,i} \subseteq \Theta; \quad \left(P(\Theta), \tilde{m}_{P(\Theta),i,t}\right) \right\} \tag{7}$$

The above distribution is called weighted belief distribution with reliability (WBDR). Then the combined degree of belief $\gamma_{n,i}$ to which T pieces of independent evidence $e_{i,t}$ with weight v_t and reliability $r_t (t = 1, \ldots, T)$, jointly support proposition n is given by

$$\gamma_{n,i} = \frac{\hat{m}_{n,e(i,T)}}{\sum_{n=1,\ldots,N_i} \hat{m}_{n,e(i,T)}}, \quad n = 1, \ldots, N_i \tag{8}$$

$$\hat{m}_{n,e(i,t)} = \left[(1 - r_t) m_{n,e(i,t-1)} + m_{P(\Theta),e(i,t-1)} m_{n,i,t}\right] + \sum_{B \cap C = n} m_{B,e(i,t-1)} m_{C,i,t}, \quad \forall n \subseteq \Theta \tag{9}$$

$$m_{P(\Theta),e(i,t)} = (1 - r_t) m_{P(\Theta),e(i,t-1)} \tag{10}$$

The above equation is the recursive combination rule of the ER rule. It satisfies the commutativity and associativity of multiplication. The ER rule provides a rational way for handling with conflicting evidences through: (1) allocating conflicting beliefs to the power set of the frame of discernment; and (2) modifying the initial belief function to better represent original information by using the WBDR.

The utility based transformation method for combining two evaluation criteria with different grades

Different sets of evaluation grades are usually used to assess different qualitative criteria in a real decision environment. In order to get the overall performance, the original assessments need to be transformed to a unified format. When utilities can be estimated explicitly, a utility based information transformation technique can be applied to implement the transformation process (Yang 2001).

Suppose the utilities of all grades have been estimated by a panel of decision makers and denoted by $u(H_j)$, $\left(u(H_{j+1}) > u(H_j), \quad j = 1, \ldots, N\right)$ and $u\left(H_{n,i}\right)$, an original

evaluation $\{(H_{n,i}, \gamma_{n,i})\}$ can be transformed to an equivalent expectation $\{(H_j, \beta_{j,i})\}$ using the following equations:

$$\beta_{j,i} = \begin{cases} \sum_{n\in\pi_j} \gamma_{n,i}\tau_{j,n}, & \text{for} \quad j = 1, \\ \sum_{n\in\pi_{j-1}} \gamma_{n,i}(1-\tau_{j-1,n}) + \sum_{n\in\pi_j} \gamma_{n,i}\tau_{j,n}, & \text{for} \quad 2 \le j \le N-1, \\ \sum_{n\in\pi_{j-1}} \gamma_{n,i}(1-\tau_{j-1,n}), & \text{for} \quad j = N, \end{cases} \tag{11}$$

$$\text{and} \quad \tau_{j,n} = \frac{u(H_{j+1}) - u(H_{n,i})}{u(H_{j+1}) - u(H_j)} \quad \text{if } u(H_j) \le u(H_{n,i}) \le u(H_{j+1}), \tag{12}$$

$$\pi_j = \begin{cases} \{n | u(H_j) \le u(H_{n,i}) < u(H_{j+1}), \quad n = 1, \ldots, N_i\}, & j = 1, \ldots, N-2, \\ \{n | u(H_j) \le u(H_{n,i}) \le u(H_{j+1}), \quad n = 1, \ldots, N_i\}, & j = N-1. \end{cases} \tag{13}$$

The utilities of grades can be determined using the decision maker's preferences. If preferences are not available, the utilities of evaluation grades can be assumed to be linearly distributed in the normalized utility space, that is, $u(H_j) = (j-1)/(N-1) \quad (j = 1, \ldots, N)$.

After the aggregated peer evaluation information on the two basic criteria has been transformed into a common set of evaluation grades, the proposed ER rule based aggregation method can be implemented again to aggregate them together in a consistent way, assuming that the weight and reliability of each criterion are known. Then the combined belief distribution can be used to represent the overall performance profile of a project. Suppose β_j is the combined degrees of belief to H_j, and a distributed assessment for the overall performance of a project a_l can be described by

$$S(y(a_l)) = \{(H_j, \beta_j), \quad j = 1, 2, \ldots N\} \tag{14}$$

Finally, the expected utility can be used for ranking projects, which is calculated as follows

$$z = \sum_{j=1}^{N} u(H_j)\beta_j \tag{15}$$

The results of the ER rule based aggregation method can be used to support the selection of research proposals for further panel review, and it can also provide a more consistent and informative way compared with the existing practice of showing the simple additive peer review scores of proposals. On the other hand, it can save considerable time for panel review meetings, as some proposals will be screened out from peer review process.

An experimental study

Introduction to the project evaluation dataset from the NSFC

In this experimental study, the dataset on the project proposals of general programs is collected from the NSFC information center. The dataset consists of two parts including the project review information and the project approval information. The project review information includes project number, comprehensive evaluation level, and funding

recommendation, and the amount of money under the "Fund" recommendation category. The project approval information contains project number, project approval number, amount of money approved, etc. To check whether a project has been successful, we can match the project number between the review information and the approval information.

The NSFC requires that the proposed research should be of scientific significance and has research merits, good theoretical basis, new academic ideas, clear research objectives, specific research contents and feasible research schemes. According to these requirements, the review form is designed with the emphasis on two criteria, namely "Comprehensive evaluation level" and "Funding recommendation". The NSFC gives brief explanation on the four grades of the first criterion, as listed in Table 3.

From the above discussion, it can be noted that the evaluation and selection of projects are mainly based on two types of information in the review form. In this paper, we take into account both the weights and the reliabilities of the evaluation information in order to combine the evidence rigorously. The logical relationship of the evaluation problem is shown in Fig. 3.

As shown in Fig. 3, experts assess a project against the two evaluation criteria simultaneously, and the quality of evaluation information is influenced by the experts'

Table 3 Explanation of evaluation grades

Grades	Reference standard
Excellent	Strong innovation; important scientific significance or application prospects; explicit research purpose; appropriate research content; feasible overall scheme; good research foundation and conditions
Good	Novel idea; important scientific significance or application prospects; good research content and overall scheme; some research foundation and conditions
Average	Some scientific research value or application; fair research content and overall scheme, but need to be modified
Poor	Obviously insufficient in key aspects

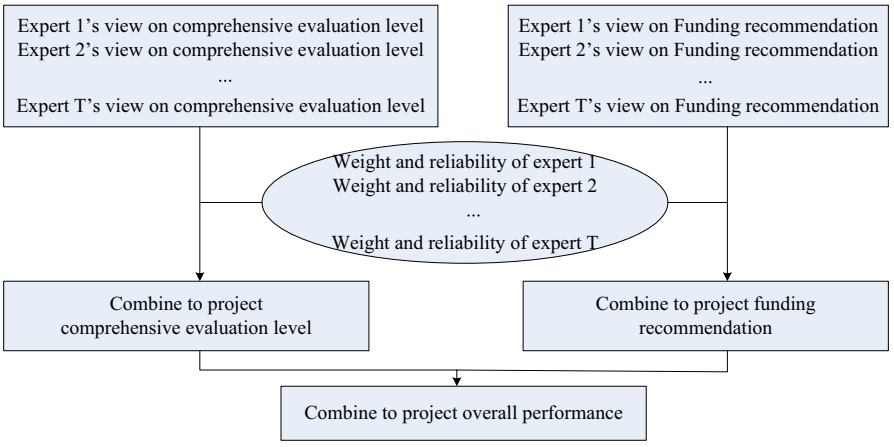

Fig. 3 Logical relationship diagram of peer review form

reliabilities and weights. The overall aggregation process will be discussed in the following sections.

The process for research project evaluation and selection

The proposed research project evaluation and selection process has six steps. At steps 1 and 2, the data are preprocessed and transformed into evidence for further analysis. At steps 3–5, the evaluation model is formulated. At step 6, the projects are ranked with the utility function. A MATLAB program is designed to implement the following processes and generate the final results.

Step 1: Denote the review information of each project on each criterion by each expert by a belief distribution, which is also called a piece of evidence.
Step 2: Calculate the reliabilities and weights of experts. Use the confusion matrix as shown in Table 2 and the historical review matching rate to calculate the reliabilities. Each expert has two reliability measures, one for the positive recommendations (i.e., Fund or fund with priority) and one for the negative recommendation (i.e., Not fund). The same reliability measure is applied to all criteria. If some experts give positive and some give negative, true positive and true negative rates are used as reliability measures accordingly. At present, as each expert plays an equally important role in the evaluation process, the weight of each expert is set as one divided by the number of experts of evaluating the project.
Step 3: Aggregate multiple experts' review information on each criterion using the ER rule. It takes weight and reliability of experts into account and the direct assignment method can be used for generating the weights. It should be noted that in the practice of project evaluation in the NSFC, no expert can dominate the evaluation result even if his or her reliability is equal to 1. To aggregate experts' review information with weights and reliabilities by using the ER rule, the normalisation factor is revised to be $c_{rw,i} = 1/(1 + w_i - w_i r_i)$, where $w_i r_i$ in the normalisation factor sets a bound within which r_i can play a limited role (Wang et al. 2015).
Step 4: Transform the belief distributions on basic criteria to a unified format using the utility based transformation technique proposed previously.
Step 5: Aggregate the transformed information on the two evaluation criteria to get the overall performance using the ER rule. The weight of the two evaluation criteria should be taken into account in this step. With the use of the direct assignment method, the weights of the two evaluation criteria in this paper are assigned as $\omega_1 = 2/3$ and $\omega_2 = 1/3$ respectively.
Step 6: Sort the multiple projects with utility function.

Implementation of the proposed ER rule based method for project evaluation

In this section, we will take a research project, namely R_1, for example and illustrate the evaluation process step by step. The original review information is as shown in Table 4. The overall score of project R_1 under the existing method in the NSFC is 4.4.

Step 1: A set of evaluation grades to assess the comprehensive evaluation level is denoted by $H_{:,1} = \{\text{excellent}, \text{good}, \text{average}, \text{poor}\} = \{H_{4,1}, H_{3,1}, H_{2,1}, H_{1,1}\}$. In terms of the second basic evaluation criterion funding recommendation, the following set of evaluation grades is denoted by $H_{:,2} = \{\text{fund with priority}, \text{fund}, \text{not fund}\} =$

Table 4 Original evaluation information by experts

	Expert 1	Expert 2	Expert 3	Expert 4	Expert 5
Comprehensive evaluation level	Excellent	Average	Excellent	Good	Excellent
Funding recommendation	Fund	Not fund	Fund	Fund	Fund with priority

Table 5 Historical review information of experts for calculating reliability

	Expert 1	Expert 2	Expert 3	Expert 4	Expert 5
No. of projects reviewed	15	5	10	11	20
Negative	3	1	2	5	6
Positive	12	4	8	6	14
TN	3	1	2	4	6
TP	3	1	3	2	6
True positive rate	0.25	0.25	0.375	0.3333	0.4286
True negative rate	1	1	1	0.8	1
Reliability	0.25	1	0.375	0.3333	0.4286

Table 6 Results generated by combining the five experts' evaluation of a project on each criterion

Comprehensive evaluation level	Excellent	Good	Average	Poor
	0.6311	0.1703	0.1986	0
Funding recommendation	Fund with priority	Fund	Not fund	
	0.1741	0.6269	0.1990	

$\{H_{3,2}, H_{2,2}, H_{1,2}\}$. Then the above evaluation information can be denoted by the following two sets of five belief distributions: $(H_{4,1}, 1); (H_{2,1}, 1); (H_{4,1}, 1); (H_{3,1}, 1); (H_{4,1}, 1)$ and $(H_{2,2}, 1); (H_{1,2}, 1); (H_{2,2}, 1); (H_{2,2}, 1); (H_{3,2}, 1)$ respectively. It is evident that some pieces of evidence from difference experts are inconsistent or conflicting with each other.

Step 2: As Experts 1, 3, 4 and 5 give positive recommendation, the true positive rates are used as the reliabilities of these four experts. Expert 2 gives negative recommendation, the true negative rate is used as reliability measures. It can be seen from Table 5 that the reliabilities of the five experts are 0.25, 1, 0.375, 0.3333 and 0.4286 respectively. If the historical information is not available for experts, the average reliability can be used to replace the missing value.

Step 3: The direct assignment method is used for weight generation. Suppose the five experts are of equal importance, i.e., $\omega_i = 1/5$, $(i = 1, \ldots, 5)$. Using the information of weights and reliabilities, the assessment of individual experts for each project are aggregated into an assessment on each criterion, as shown in Table 6.

Step 4: As preference information is not available, the utilities of evaluation grades for different criteria can be assumed to be linearly distributed in the normalized utility space, that is, $u(H_n) = (n-1)/(N-1) \quad (n = 1, \ldots, N)$. Then the utilities of the six grades of the top level criterion can be $u(H_6) = 1$, $u(H_5) = 0.8$, $u(H_4) = 0.6$,

Table 7 Transformed assessment distribution on the final 6 overall categories

	A	*A*−	*B*	*E*	*C*	*D*
Comprehensive evaluation level	0.6311	0.0568	0.1135	0.1324	0.0662	0.0000
Funding recommendation	0.1741	0.0000	0.3135	0.3135	0.0000	0.1990
Results from the ER rule	0.5430	0.0422	0.1561	0.1723	0.0493	0.0370

$u(H_3) = 0.4$, $u(H_2) = 0.2$, $u(H_1) = 0$. In the same way, the utilities of the basic criteria are $u(H_{4,1}) = 1$, $u(H_{3,1}) = 0.67$, $u(H_{2,1}) = 0.33$, $u(H_{1,1}) = 0$ and $u(H_{3,2}) = 1$, $u(H_{2,2}) = 0.5$, $u(H_{1,2}) = 0$ respectively. Use the proposed utility based information transformation technique, the assessment with different grades in Table 6 can be transformed to a unified format, as shown in Table 7.

Step 5: As there are just two evaluation criteria involved in this experimental study, direct assignment method is used for generating weights. As discussed previously, the weight of the first criterion is 2/3 and of the second is 1/3. Then the ER rule can be employed to calculate the overall degrees of belief, as included in Table 7.

Step 6: The expected utility of the project is calculated by Eq. (15) and the final utility is 0.7493.

Results and comparative analysis

Statistical analyses of the whole data set

The NSFC receives thousands of applications every year, but not all of them can be submitted for panel review. Some projects are rejected directly for administrative reasons. The remaining projects are rated by three to five experts. To keep consistency, projects with five valid reviewers are picked up. There are 1225 projects in the data set, and among which 210 of them were funded and 1015 were not. Using the existing method, the range of scores of the 1225 projects is [1.0, 5.8]. Using the discrete step size of 0.2, the histogram distributions of the funded and unfunded projects on the scores are shown in Fig. 4.

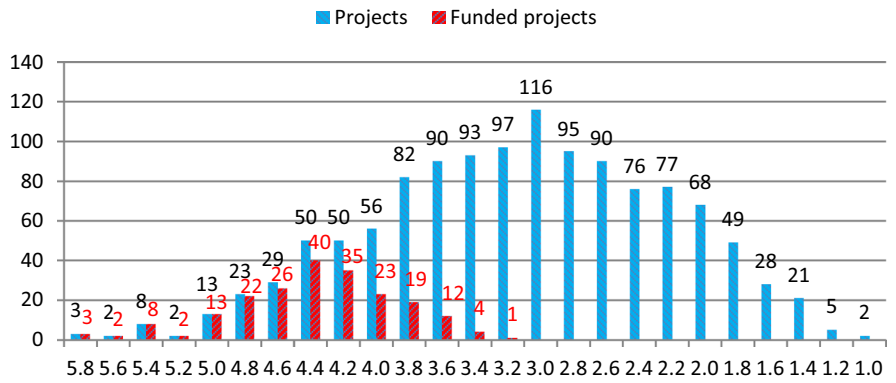

Fig. 4 The histogram distribution of the funded and unfunded projects under the existing method

It can be seen from Fig. 4 that there are 3 projects achieving the highest score of 5.8 and 2 projects having the lowest score of 1.0. A large number of projects, 116 had a score of 3.0 and none of them was funded. Within the same score categories, such as from 4.8 to 3.2, some projects were funded, but some others were not. Specially, take the 56 projects having the score of 4 for example, 23 of them were funded and 33 of them were not. It means that the scores alone cannot differentiate a large number of projects, and the existing method lacks the ability to capture the true performance profile of these projects.

Illustration using representative projects

Ten representative projects from different assessment categories are then selected and the original assessments provided by peer experts are shown in Table 8. The detailed information for experts' reliabilities is presented in "Appendix 1".

Generally, given the peer review information, it is assumed that the closer the aggregated results to the final funding outcomes, the more rational and reliable the processing method is. Therefore we take the final funding outcomes as the basis, and compare the proposed method with the existing method of the NSFC. The scores generated using the existing method and the ER rule with reliability are shown in Table 9 headed by x and y respectively. The rankings of projects are headed by O_x and O_y respectively.

As shown in Table 9, there are some differences in the rankings of the projects when applying the proposed method. The most obvious difference happens in Projects 5 and 6. Project 5 falls to the 6th in the ER rule based method from the 4th in the existing method and Project 6 rises from the 6th in the existing method to the 5th in the ER rule based method. Project 6 gets a higher utility score than Project 5 in the ER rule based method, while the order of their utility scores is in the other way round in the existing method. The main reason for Project 6 getting a high score in the ER based method is that Expert 5 gave an "Excellent" recommendation on the project and the reliability of Expert 5, which is the true positive rate 0.5714, is quite high. Expert 5 is regarded as more reliable and his opinion is given a more important role in the aggregation process. In contrast, the main reason for Project 5 getting a higher score in the existing method and a lower score in the

Table 8 Original performance assessment of projects represented by belief structure

	Expert 1	Expert 2	Expert 3	Expert 4	Expert 5
1	$(H_{3,1},1)(H_{2,2},1)$	$(H_{3,1},1)(H_{2,2},1)$	$(H_{3,1},1)(H_{2,2},1)$	$(H_{4,1},1)(H_{3,2},1)$	$(H_{4,1},1)(H_{3,2},1)$
2	$(H_{4,1},1)(H_{2,2},1)$	$(H_{2,1},1)(H_{1,2},1)$	$(H_{4,1},1)(H_{2,2},1)$	$(H_{3,1},1)(H_{2,2},1)$	$(H_{4,1},1)(H_{3,2},1)$
3	$(H_{3,1},1)(H_{2,2},1)$	$(H_{3,1},1)(H_{2,2},1)$	$(H_{4,1},1)(H_{3,2},1)$	$(H_{2,1},1)(H_{1,2},1)$	$(H_{4,1},1)(H_{3,2},1)$
4	$(H_{4,1},1)(H_{2,2},1)$	$(H_{3,1},1)(H_{2,2},1)$	$(H_{2,1},1)(H_{1,2},1)$	$(H_{4,1},1)(H_{3,2},1)$	$(H_{3,1},1)(H_{1,2},1)$
5	$(H_{4,1},1)(H_{2,2},1)$	$(H_{3,1},1)(H_{2,2},1)$	$(H_{1,1},1)(H_{1,2},1)$	$(H_{4,1},1)(H_{3,2},1)$	$(H_{3,1},1)(H_{2,2},1)$
6	$(H_{2,1},1)(H_{1,2},1)$	$(H_{3,1},1)(H_{1,2},1)$	$(H_{4,1},1)(H_{2,2},1)$	$(H_{3,1},1)(H_{1,2},1)$	$(H_{4,1},1)(H_{3,2},1)$
7	$(H_{2,1},1)(H_{1,2},1)$	$(H_{2,1},1)(H_{2,2},1)$	$(H_{3,1},1)(H_{2,2},1)$	$(H_{3,1},1)(H_{2,2},1)$	$(H_{4,1},1)(H_{3,2},1)$
8	$(H_{2,1},1)(H_{1,2},1)$	$(H_{3,1},1)(H_{2,2},1)$	$(H_{2,1},1)(H_{1,2},1)$	$(H_{4,1},1)(H_{3,2},1)$	$(H_{2,1},1)(H_{1,2},1)$
9	$(H_{3,1},1)(H_{3,2},1)$	$(H_{2,1},1)(H_{1,2},1)$	$(H_{2,1},1)(H_{1,2},1)$	$(H_{3,1},1)(H_{2,2},1)$	$(H_{2,1},1)(H_{1,2},1)$
10	$(H_{4,1},1)(H_{3,2},1)$	$(H_{2,1},1)(H_{1,2},1)$	$(H_{1,1},1)(H_{1,2},1)$	$(H_{3,1},1)(H_{2,2},1)$	$(H_{2,1},1)(H_{1,2},1)$

Table 9 Results generated by the two methods

Project	x	O_x	y	O_y	Funded
1	4.8	1	0.7755	1	Yes
2	4.4	2	0.7493	2	Yes
3	4.4	2	0.7102	3	Yes
4	4	4	0.6629	4	Yes
5	4	4	0.6187	6	No
6	3.8	6	0.6371	5	Yes
7	3.8	6	0.5558	7	No
8	3.2	8	0.4401	8	No
9	3	9	0.4097	9	No
10	3	9	0.3881	10	No

ER rule based method is that the reliability of Expert 4 is 0, although Expert 4 gave an "Excellent" recommendation on the project. Therefore, the review information of Expert 4 was given a large discount when aggregated with others. As can be seen from the final funding outcomes, Project 6 was funded while Project 5 was not. This indicates the applicability and effectiveness of the proposed ER rule based method for research project evaluation and selection. More statistics about the performance of the ER rule based method are given in the following sub-sections.

It can also be observed that Projects 2&3, 4&5, 6&7, and 9&10 have the same score using the existing method but their overall performance utilities are different using the ER rule based method. The main reason is that we take into account both the weights of the two evaluation criteria and the reliabilities of the review information provide by different experts, which is more consistent with the real-world situation.

Belief distributions of two projects In this section, two projects, namely R_2 and R_3 are selected for analysis, and the original evaluation information provided by experts and data for experts' reliabilities are listed in "Appendix 2". The overall scores of the two projects obtained from using the existing method are the same, which is 3.6. But project R_2 was funded while R_3 was not. Although the overall performance scores of the two projects under the ER rule based method are also very similar, the aggregated performance distributions generated using the ER rule based method for the two projects are very different, as shown in Fig. 5.

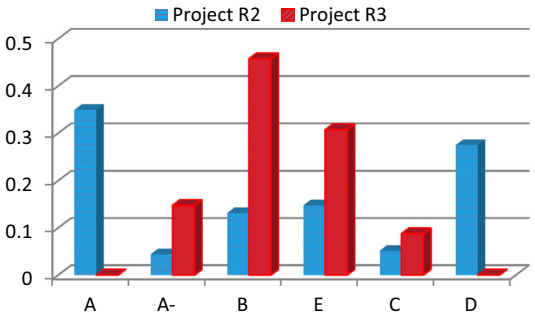

Fig. 5 Profiles generated using the ER rule based method of two projects

Table 10 Actual outcomes for top 210 projects under the ER rule based method and the existing method

	Funded	Undifferentiated	Unfunded	Total number of top projects
The ER rule based method	164	0	46	210
Existing method	151	30	29	210

In Fig. 5, the belief distributions for project R_3 are mostly centralized on the middle levels. However, those for project R_2 are on the two extremes. From the perspective of research, project R_2 is debatable and may be of higher research value. Thus, by providing the profile of projects, the ER rule based method can provide a more informative way so as to make informed decision in project evaluation and selection.

Comparative statistical analysis

For all the 1225 projects, the ER rule based method and the existing method are used to calculate the overall performances. Since there are 210 projects were funded in the actual funding outcomes, the top 210 projects ranked by the two methods are chosen for analysis. The results are shown in Table 10.

The final outcomes of the ER rule based method are significantly different from those of the existing method. In accordance with the data in Fig. 4, it can be seen that 180 projects are scored no less than 4.2. The last 30 projects, which are from top 181 to 210, are associated with the same score 4.0 and there are 56 projects scored 4.0. It means that they have to be chosen in 56 projects. As discussed before, the existing method lacks the ability to distinguish ten thousands of projects. Comparing the outcomes of the three categories under the two methods, it can be noted that none of the top 210 projects is undifferentiated in the ER rule based method. Thus the ER rule based method provides a more effective way for research project selection.

Conclusions

Research project evaluation and selection involves a complicated multi-stage decision-making process. Since the existing method for evaluation of the NSFC research project cannot make full use of multiple experts' evaluation information, and treats all experts' evaluation information equally, in this paper the ER rule based method for aggregating peer review information with reliabilities is proposed to overcome this limitation. The experimental results show that the ER rule based method provides an effective way for aggregating peer review information in the NSFC and is also capable of handling different forms of peer review information obtained. The main strengths of the proposed method can be summarized as follows: (1) it uses informative belief distributions to represent projects' performance profile, (2) it utilizes historical data to measure the reliability of an expert and takes into consideration the fact that the reliability has a high influence on the quality of review information, (3) it has the ability to differentiate the quality of research projects with continuous scores, and (4) it provides a rational alternative approach to deal with conflicting evidence.

In this paper, the utilities of evaluation grades are assumed to be linearly distributed in the normalized utility space which may not represent the decision maker's preferences precisely. Thus the modelling of the decision maker's preferences needs to be studied

further. It is worth noting that the calculation of reliabilities under study in this paper is based on the historical review performances of experts, and an appropriate set of historical data are required to obtain accurate reliability information. This work can potentially be extended to general project evaluation and selection problems for other funding agencies. The research outcome can also be used to provide decision support for governmental organizations and companies to conduct project evaluation and selection in a more rigorous and effective way.

Acknowledgments This research is partially supported by the National Natural Science Foundation of China under Grant No. 71071048 and the Scholarship from China Scholarship Council under Grant No. 201306230047.

Appendix 1: Reliabilities of experts for "Results and comparative analysis" section

As the reliabilities of some experts are not available in the data set, the average true positive rate of 0.2726 and the average true negative rate of 0.9592 are used as their reliabilities accordingly.

	No. of projects	Negative	Positive	TN	TP	True positive rate	True negative rate
Project 1/Expert 1	2		2			0	
Expert 2	12	5	7	4		0	0.8
Expert 3	11	9	2	7	1	0.5	0.7778
Expert 4	19	4	15	4	6	0.4	1
Expert 5	4	1	3	1		0	1
Project 3/Expert 1	11	8	3	5		0	0.625
Expert 2	11	6	5	6	1	0.2	1
Expert 3	18	3	15	3	4	0.2667	1
Expert 4	15	11	4	11	1	0.25	1
Expert 5	10	4	6	4		0	1
Project 4/Expert 1							
Expert 2	15	4	11	4	1	0.0910	1
Expert 3	16	13	3	12		0	0.9231
Expert 4	7	1	6	1	3	0.5	1
Expert 5	12	7	5	7	2	0.4	1
Project 5/Expert 1							
Expert 2							
Expert 3							
Expert 4	6	1	5	1		0	1
Expert 5	7	5	2	5	1	0.5	1
Project 6/Expert 1							
Expert 2	14	4	10	4	3	0.3	1
Expert 3	18	2	16	2	3	0.1875	1
Expert 4							
Expert 5	12	5	7	5	4	0.5714	1

	No. of projects	Negative	Positive	TN	TP	True positive rate	True negative rate
Project 7/Expert 1							
Expert 2	3	1	2	1	2	1	1
Expert 3	23	7	16	7	3	0.1875	1
Expert 4							
Expert 5							
Project 8/Expert 1	15	11	4	11	1	0.25	1
Expert 2	18	3	15	3	4	0.2667	1
Expert 3	11	6	5	6	1	0.2	1
Expert 4	10	4	6	4		0	1
Expert 5	11	8	3	5		0	0.625
Project 9/Expert 1	10	2	8	2	3	0.375	1
Expert 2	11	5	6	4	2	0.3333	0.8
Expert 3	5	1	4	1	1	0.25	1
Expert 4	20	6	14	6	6	0.4286	1
Expert 5	11	8	3	5		0	0.625
Project 10/Expert 1	18	3	15	3	4	0.2667	1
Expert 2	15	11	4	11	1	0.25	1
Expert 3	11	8	3	5		0	0.625
Expert 4	10	4	6	4		0	1
Expert 5	11	6	5	6	1	0.2	1

Appendix 2

	Expert 1	Expert 2	Expert 3	Expert 4	Expert 5
Original evaluation information of project R_2 by experts					
Comprehensive evaluation level	Average	Good	Poor	Excellent	Excellent
Funding recommendation	Not fund	Fund	Not fund	Fund with priority	Fund

	Expert 1	Expert 2	Expert 3	Expert 4	Expert 5
Original evaluation information of project R_3 by experts					
Comprehensive evaluation level	Average	Good	Good	Good	Average
Funding recommendation	Fund	Fund	Fund	Fund	Fund

	Experts	Projects	Negative	Positive	TN	TP	True positive rate	True negative rate
Reliabilities of experts for project R_2 and R_3								
Project R_2	Expert 3	13	4	9	4	4	0.4444	1
Project R_3	Expert 3	19	6	13	6	6	0.461538462	1
	Expert 5	12	2	10	2	3	0.3	1

The original data set is available for research use with request

References

Agarski, B., Budak, I., Kosec, B., & Hodolic, J. (2012). An approach to multi-criteria environmental evaluation with multiple weight assignment. *Environmental Modeling and Assessment, 17*(3), 255–266.

Bulathsinhala, N. A. (2014), Ex-ante evaluation of publicly funded R&D projects: Searching for exploration. *Science and Public Policy*, scu035, 1-14.

Carlsson, C., Fullér, R., Heikkilä, M., & Majlender, P. (2007). A fuzzy approach to R&D project portfolio selection. *International Journal of Approximate Reasoning, 44*(2), 93–105.

Chen, X. T. (2009). *National science fund and management sciences (1986–2008)*. Beijing: Science Press.

Coffin, M. A., & Taylor, B. W, I. I. I. (1996). Multiple criteria R&D project selection and scheduling using fuzzy logic. *Computers & Operations Research, 23*(3), 207–220.

Feng, B., Ma, J., & Fan, Z. P. (2011). An integrated method for collaborative R&D project selection: Supporting innovative research teams. *Expert Systems with Applications, 38*(5), 5532–5543.

Henriksen, A. D., & Traynor, A. J. (1999). A practical R&D project-selection scoring tool. *IEEE Transactions on Engineering Management, 46*(2), 158–170.

Horrobin, D. F. (1996). Peer review of grant applications: A harbinger for mediocrity in clinical research? *The Lancet, 348*(9037), 1293–1295.

Hsu, Y. G., Tzeng, G. H., & Shyu, J. Z. (2003). Fuzzy multiple criteria selection of government-sponsored frontier technology R&D projects. *R&D Management, 33*(5), 539–551.

Huang, C. C., Chu, P. Y., & Chiang, Y. H. (2008). A fuzzy AHP application in government-sponsored R&D project selection. *Omega, 36*(6), 1038–1052.

Jayasinghe, U. W., Marsh, H. W., & Bond, N. (2006). A new reader trial approach to peer review in funding research grants: An Australian experiment. *Scientometrics, 69*(3), 591–606.

Južnič, P., Pečlin, S., Žaucer, M., Mandelj, T., Pušnik, M., & Demšar, F. (2010). Scientometric indicators: Peer-review, bibliometric methods and conflict of interests. *Scientometrics, 85*(2), 429–441.

Khalili-Damghani, K., Sadi-Nezhad, S., & Tavana, M. (2013). Solving multi-period project selection problems with fuzzy goal programming based on TOPSIS and a fuzzy preference relation. *Information Sciences, 252*, 42–61.

Lawson, C. P., Longhurst, P. J., & Ivey, P. C. (2006). The application of a new research and development project selection model in SMEs. *Technovation, 26*(2), 242–250.

Linton, J. D., Walsh, S. T., & Morabito, J. (2002). Analysis, ranking and selection of R&D projects in a portfolio. *R&D Management, 32*(2), 139–148.

Mahmoodzadeh, S., Shahrabi, J., Pariazar, M., & Zaeri, M. S. (2007). Project selection by using fuzzy AHP and TOPSIS technique. *World Academy of Science, Engineering and Technology, 30*, 333–338.

Meade, L. M., & Presley, A. (2002). R&D project selection using the analytic network process. *IEEE Transactions on Engineering Management, 49*(1), 59–66.

Olsson, N. O., Krane, H. P., Rolstadås, A., & Veiseth, M. (2010). Influence of reference points in ex post evaluations of rail infrastructure projects. *Transport Policy, 17*(4), 251–258.

Oral, M., Kettani, O., & Çınar, Ü. (2001). Project evaluation and selection in a network of collaboration: A consensual disaggregation multi-criterion approach. *European Journal of Operational Research, 130*(2), 332–346.

Provost, F., & Kohavi, R. (1998). Guest editors' introduction: On applied research in machine learning. *Machine Learning, 30*(2), 127–132.

Shafer, G. (1976). *A mathematical theory of evidence* (Vol. 1). Princeton: Princeton University Press.

Silva, T., Guo, Z., Ma, J., Jiang, H., & Chen, H. (2013). A social network-empowered research analytics framework for project selection. *Decision Support Systems, 55*(4), 957–968.

Silva, T., Jian, M., & Chen, Y. (2014). Process analytics approach for R&D project selection. *ACM Transactions on Management Information Systems (TMIS), 5*(4), 21.

Smarandache F, Dezert J, Tacnet J. (2010). Fusion of sources of evidence with different importances and reliabilities. In: *2010 13th conference on information fusion* (*FUSION*). IEEE (pp. 1–8).

Solak, S., Clarke, J. P. B., Johnson, E. L., & Barnes, E. R. (2010). Optimization of R&D project portfolios under endogenous uncertainty. *European Journal of Operational Research, 207*(1), 420–433.

Tavana, M., Khalili-Damghani, K., & Sadi-Nezhad, S. (2013). A fuzzy group data envelopment analysis model for high-technology project selection: A case study at NASA. *Computers & Industrial Engineering, 66*(1), 10–23.

Tian, Q., Ma, J., Liang, J., Kwok, R. C., & Liu, O. (2005). An organizational decision support system for effective R&D project selection. *Decision Support Systems, 39*(3), 403–413.

Wang, J., & Hwang, W. L. (2007). A fuzzy set approach for R&D portfolio selection using a real options valuation model. *Omega, 35*(3), 247–257.

Wang, D. P., Zhu, W. D., Chen, B., & Liu, F. (2015). Analysis of the expert judgments in the NSFC peer review from the perspective of human cognition: A research based on the ER rule. *Science of science and management of S. & T., 36*(4), 22–35.

Xu, D. L. (2012). An introduction and survey of the evidential reasoning approach for multiple criteria decision analysis. *Annals of Operations Research, 195*(1), 163–187.

Yang, J. B. (2001). Rule and utility based evidential reasoning approach for multiattribute decision analysis under uncertainties. *European Journal of Operational Research, 131*(1), 31–61.

Yang, J. B., & Singh, M. G. (1994). An evidential reasoning approach for multiple-attribute decision making with uncertainty. *IEEE Transactions on Systems, Man and Cybernetics, 24*(1), 1–18.

Yang, J. B., & Xu, D. L. (2002). On the evidential reasoning algorithm for multiple attribute decision analysis under uncertainty. *IEEE Transactions on Systems, Man, and Cybernetics Part A: Systems and Humans, 32*(3), 289–304.

Yang, J. B., & Xu, D. L. (2013). Evidential reasoning rule for evidence combination. *Artificial Intelligence, 205*, 1–29.