Title Page

CE-DIFF: An Approach to Identifying and Coping with Irregular Ratings in Collaborative Decision Making

Li Yu

School of Information, Renmin University of China, 59 Zhongguancun St, Haidian, 100872 Beijing, China

E-mail: buaayuli@ruc.edu.cn

Dr. Li Yu is an associate professor and Vice Chair of the Department of Economic Information Management at Renmin University of China. He received his Ph.D. degree from the School of Economics & Management, Beihang University, in 2004. His research interests include internet marketing & viral marketing, social commerce & e-commerce, personalized recommender systems, customer relationship management, social computing, collaborative evaluation & group decision, finance data mining & analysis, etc.

Dongsong Zhang*

Belk College of Business, University of North Carolina at Charlotte, 9201 University City Blvd, Charlotte, 28223-0001, U.S.

E-mail: dzhang15@uncc.edu

Dr. Dongsong Zhang is a Belk Endowed Chair Professor in the Department of Business Information Systems and Operations Management and a professor (Courtesy) in the Department of Computer Science at the University of North Carolina at Charlotte. He received his Ph.D. in management information systems from the Eller School of Management, the University of Arizona. His research interests include social media analytics, mobile computing, e-commerce, collaborative decision making, and health IT. He has published approximately 150 research papers and received a dozen research grants and awards from National Science Foundation, National Institute of Health, and Google, etc.

---

* corresponding author

This is the author manuscript accepted for publication and has undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1111/deci.12480.

Zhe Fu

School of Information, Renmin University of China, 100872 Beijing, China. e-mail: fuzhe@ruc.edu.cn

Zhe Fu is a Master student at the School of Information, Renmin University of China.

Abstract

A key to collaborative decision making is to aggregate individual evaluations into a group decision. One of its fundamental challenges lies in the difficulty in identifying and dealing with irregular or unfair ratings and reducing their impact on group decisions. Little research has attempted to identify irregular ratings in a collaborative assessment task, let alone develop effective approaches to reduce their negative impact on the final group judgment. In this paper, based on the synergy theory, we propose a novel consensus-based collaborative evaluation method called Collaborative Evaluation based on rating DIFFerence (CE-DIFF) for identifying irregular ratings and mitigating their impact on collaborative decisions. CE-DIFF determines and assigns different weights automatically to individual evaluators or ratings based on the level of consistency of one's ratings with the group assessment outcome through continuous iterations. We conducted two empirical experiments to evaluate the proposed method. The results show that CE-DIFF has higher accuracy in dealing with irregular ratings than existing collaborative evaluation methods, such as arithmetic mean and trimmed mean. In addition, the effectiveness of CE-DIFF is independent of group size. This study provides a new and more effective method for collaborative assessment, as well as novel theoretical insights and practical implications on how to improve collaborative assessment.

Keywords: Web2.0, Group decision making, Collaborative evaluation, Irregular rating, Weight

INTRODUCTION

One of the common group decision making (GDM) tasks is to rank multiple targets based on the evaluation opinions of a group of evaluators (Hochbaum & Levin, 2006; Gao & Liu, 2017), such as refereeing a gymnastics competition or evaluating research grant proposals in a review panel. A typical group decision making problem can be defined as follows: given a set of M targets $O=\{O_1, O_2,\ldots, O_M\}$ evaluated by N evaluators $E=\{E_1, E_2,\ldots, E_N\}$ and $R=\{R_{ij} \mid i=1, 2,\ldots, N, j=1, 2,\ldots, M\}$ as the ratings or scores given by evaluators on targets, in which R_{ij} represents the evaluation score/rating of evaluator E_i on target O_j, the most commonly used method for group decision making is to compute the average of ratings provided by all evaluators as the final group rating for each target, then rank all targets based on final group ratings. With that method, the ratings of a target from individual evaluators are considered equal and therefore are assigned the same weights.

Unfortunately, the above mean-based method does not always perform well in some circumstances. First, some evaluators may have intentional or unintentional bias. For example, in a gymnastics competition, a judge may intentionally give a gymnast an unjustified high (or low) score not based on the gymnast's performance, but based on their personal relationship. Second, an evaluator is constrained by his/her domain knowledge and judgment ability, which may hinder him/her from making a fair evaluation. We define the former situation as *prejudiced evaluation*, and the latter as *unqualified evaluation*. Both will result in irregular ratings, which are referred to as ratings significantly deviated from the overall group opinion. Due to possible differences in domain knowledge, motivations, and personal bias, the credibility and quality of individual group members' evaluations may vary, which would have negative impact on group decision-making outcomes (Lima, de Souza, Moura, & da Silva, 2018) and make collaborative decisions inaccurate (Forsyth, 2006).

Therefore, it is critical to address such irregular ratings in a collaborative evaluation task to ensure that group evaluation results are fair and credible (Liu, Fan, You, & Zhang, 2018).

Some methods have been deployed to reduce the impact of potential irregular or unfair assessment of individual members on group assessment outcomes. For example, in a diving competition, the highest and lowest scores given to a diver by referees are automatically excluded from final group mean calculation. However, this strategy is arbitrary and could be unfair in many cases. On the other hand, it is not always reasonable to treat all scores equally either. It seems that deploying a systematic method to identify the possible irregularity of individual assessments and then assign different weights to them accordingly to reflect their influence on group evaluation outcomes might be a better solution.

To address those challenges and improve the quality of collaborative evaluation, this study is aimed to answer two major research questions: How to automatically identify irregular ratings, and how to determine the weight of each evaluation rating (or evaluator) in order to generate more accurate group outcomes. Based on the Synergy Theory (Haken, 2009), we posit that the contribution of an individual evaluator to the final group evaluation outcome depends on not only the ratings made by him/her, but also the ratings made by other evaluators in the group. Accordingly, we design a consensus-based collaborative evaluation approach called CE-DIFF, which automatically identifies irregular assessments and assigns different weights to individual evaluators based on the extent to which their individual assessments deviate from the whole group opinion. The results of two empirical evaluation studies demonstrate superior performance and benefits of CE-DIFF.

This research makes several novel research contributions, including (1) proposing a novel collaborative evaluation method (CE-DIFF) by using the synergy theory as a theoretical lens; (2) developing a new method for identifying irregular assessments; and (3)

designing and deploying a novel empirical evaluation methodology for assessing the effectiveness of the proposed method. This study also provides several practical implications on how to improve collaborative evaluation.

The rest of the paper will be organized as follows. First, we will introduce related work. Then, we will introduce the synergy theory, on which the collaborative evaluation notion is proposed, and presents the CE-DIFF algorithm and explains how to identify irregular ratings based on CE-DIFF. Next, the controlled lab experiments conducted to evaluate the proposed methods will be introduced, followed by the presentation of results. Finally, the paper will discuss major findings, research contributions, practical implications, and limitations of the study.

RELATED WORK

Group Decision Making

Group decision making aims at aggregating individual judgments to make a collective group decision (Appelt, Milch, Handgraaf, & Weber, 2011; de Bruin, Missier, & Levin, 2012). There are three main forms of individual preferences or opinions widely studied in the context of group decision making in literature (Contreras, 2011). The most common form is rating, where group members submit their ratings or scores for multiple targets or alternatives. The second form is pairwise comparison, where decision makers make a comparison between two targets and use a ratio to represent their relative importance to a decision. The third form is ranking, where a decision maker ranks alternatives based on certain criteria. In this study, we mainly focus on group decision making based on ratings of individual members.

In general, there are several commonly used rating-based group decision making methods, including consensus-based decision making, voting-based decision making, majority-based method, and plurality based methods (Contreras, 2011; Labella, Liu, Rodríguez & Martínez, 2018). In consensus-based decision making, all group members must converge to a consensus rating as the final group rating; with voting-based methods, each evaluator votes for one or more available options, then the option receiving the largest number of votes will be chosen as the group decision; a majority-based method requires support from more than 50% of group members; and for a plurality-based method, the final group rating is the rating cast by more group members, regardless of being an absolute majority or not, than any other ratings.

Large group decision-making (LGDM) refers to the selection of the best option from a set of feasible alternatives according to the opinions of a large number of individuals (Liu & Liu, 2018; Xu & Luo, 2019). It considers not only the opinions of individuals in a group, but also the fairness of alternatives proposed by stakeholders from different subgroups who may differ in their motivations and interest. Xu and Luo (2019) identified risks of poor large group decisions caused by preference differences among group members, and proposed a new risk measure based on information entropy theory to predict risks through clustering group members into subgroups based on their shared interest.

Weight-based Group Decision Making

Multi-criteria group decision making (MGDM) focuses on the selection of the best alternative from a finite set of alternatives based on assessments of multiple attributes (Liu & Liu, 2018). In many cases, those evaluation criteria may be of different levels of importance to a decision. For example, Talluri, DeCampos, and Hult (2013) proposed a

method to measure the performance diversity of suppliers and evaluate the efficiency of a supplier with respect to the optimal weights of its peers.

Some popular MGDM methods include the Weighted Sum Model (WSM) and Analytic Hierarchy Process (AHP). WSM is a well-known and the simplest multi-criteria decision analysis method (Zheng, Teng, & Liu, 2016), in which each decision alternative is assessed and compared with other alternatives by aggregating the multiplications of individual attribute values and the corresponding weights associated with those attributes. AHP is another widely used method that determines weights (i.e., relative importance) of individual criteria (i.e., attributes) based on pairwise criteria comparisons in terms of their importance to a decision. Weights are non-negative real numbers in the range of 0~1, with the sum of all weights equal to 1. The higher a weight value is, the more important a criterion will be.

Consensus-based Decision Models

Consensus-based decision making is a process of a group generating a unified decision in a way that respects the contributions of all participants (Labella et al., 2018; Lima et al., 2018; Bruccoleri, Riccobono, & Größler, 2019). Its basic process involves collaboratively proposing a solution, identifying unsatisfied concerns, and modifying the solution to address those concerns.

The most common method for aggregating individual ratings into a collective group rating in consensus-based decision making is the mean-based method, in which the average value of the ratings of all evaluators is used as the final group evaluation outcome. This method assumes that the ratings of individual evaluators are equally important to the final group decision, which is not always true. Sometimes one or more evaluators may make irregular assessments because of various reasons such as insufficient domain knowledge or

personal bias, resulting in the need of valuing the importance of individual evaluators and their ratings to the final group outcome differently.

Given the limitation of the Mean method, another method using a trimmed means is also often deployed (Berkes, Györfi, & Kevei, 2016), which uses the group mean after removing a number of ratings typically at the high and low ends. For example, the 25% trimmed mean (i.e., the mean after removing the lowest 25% and the highest 25% of ratings) is known as the interquartile mean. Given a set of eight ratings, it will discard the minimum and maximum ratings from the sample. The mean of the remaining six ratings will be used as the final rating of the group. Although the trimmed means method may exclude potential irregular ratings, it is relatively arbitrary because it assumes that the largest and/or smallest ratings are irregular ones, which may or may not always be the case.

The Delphi method is a general structured communication method for consensus-based group decision making, originally developed as a systematic and interactive forecasting method that relies on a panel of experts (Ishikawa, Amagasa, Shiga, Tomizawa, Tatsuta, & Mieno, 1993). Consensus Reaching Process (CRP) is a general consensus framework with a goal to bring opinions of evaluators closer to each other so as to reduce disagreement and increase satisfaction (Morente-Molinera, Pérez, Ureña, & Herrera-Viedma, 2015). In CRP, identifying and managing non-cooperative behaviors of evaluators would incur if some evaluators refuse to accept the guidance for building higher consensus (Palomares, Martinez, & Herrera, 2013). Zhang, Liang, and Zhang (2018) developed a novel consensus reaching process for multi-attribute group decision making with hesitant fuzzy linguistic term sets.

Collaboration Engineering and Technologies for Group Support systems

There have been extensive studies on group support systems (GSS) and collaboration technologies in the past few decades. Some focused on theoretical, management, and technical aspects of group support systems (e.g., Jessup & George, 1997; Walsh & Dickey, 2004), while others investigated the use of GSS and its impact on collaborative processes and outcomes, such as GSS adoption (Van Hillegersberg & Koenen 2014), the use of GSS for knowledge management (Wang, Ding, Liu, & Li, 2016; Pyrko, Eden, & Howick, 2019), and the effect of group member and communication characteristics on group performance (Dennis, Wixom, & Vandenberg, 2001; Kim 2006; Wilson, Griffin, & Jessup, 2010; Nahartyo & Utami, 2014).

Collaboration engineering is an approach for the design and deployment of collaborative technologies and collaborative processes in support of mission-critical tasks. To enable the development of such processes, Briggs, de Vreede, and Nunamaker (2003) proposed the ThinkLet concept, a codified packet of facilitation skills that can be applied by practitioners to achieve predictable and repeatable patterns of collaboration. Kolfschoten and de Vreede (2009) presented a pattern-based design approach to collaboration engineering that incorporates existing process design methods, pattern-based design principles, and insights from expert facilitators. The approach aims to design collaboration processes as a sequence of patterns, which can be defined as a pattern language such as ThinkLets. De Vreede and Briggs (2019) assessed 331 published papers on collaboration technologies and showed that the collaboration engineering approach may be viable in a number of domains and create value towards collaboration in organizations. Moreover, the approach has potential to reduce organizations' need for collaboration professionals across different application domains. Recently, researchers explored methods for making collaboration expertise explicit, codified, teachable, learnable, and replicable (de Vreede & Briggs, 2019).

In sum, almost all of the existing methods share a common assumption that all evaluators in a collaborative evaluation task have equal importance to the final group outcome, which is not always true. Although the popular trimmed mean method may discard some potential irregular ratings, it is arbitrary and therefore will not be effective in many situations.

THEORETICAL FOUNDATION

Synergy is a form of cosmic synergy, the universal constructive principle of nature. Synergistic phenomena are ubiquitous, ranging from physics and chemistry to cooperative interactions among genes in genomes, from honeybee colonies to wolf packs and human societies, as well as to many different kinds of synergies produced by socially organized groups (Haken, 2009). From a perspective of complex systems, synergy is a mechanism of indirect coordination among agents or actions that results in self-assembly of complex systems (Knvazeva & Haken, 1999). In the context of organizational behavior, following the view that a cohesive group is more than the sum of its members, synergy is the ability of a group to outperform its best individual member (Kanisauskas, 2014).

Synergy theory (Haken, 2009) posits that all targets in a system are interrelated and interdependent, and the importance of a target and the similarity between two targets depend on other targets. Therefore, the characteristics of targets, such as importance, are mutually determined. Cyclic iterations are often used to identify the characteristics of targets.

In fact, a number of algorithms in the information management field share the synergy idea, such as SimRank, K-means, and PageRank algorithms (Brin & Page, 1998). SimRank is a link-based similarity measure proposed by Jeh and Widom (2002) for searching webpages. It posits that two similar pages would be linked to many other similar pages, which means that the similarity between two pages depends on other pages. K-means is a popular

clustering method that can automatically cluster targets into K classes based on their distance to cluster centers. Each target is assigned to the cluster with the minimal distance between the target and the cluster center. At the same time, K cluster centers, not known in advance, are determined by all targets within a cluster iteratively.

Following the same line of thought, we propose a concept of Collaborative Evaluation (CE) positing that the final decision of a group should be decided by all participating evaluators, while the contribution of an individual evaluator to the final decision will be determined by the deviation of his/her assessment result from the group mean of all assessment results.

CE-DIFF: A NOVEL COLLABORATIVE EVALUATION APPROACH BASED ON RATING DIFFERENCE

Based on the concept of collaborative evaluation, we propose a novel collaborative evaluation approach to identify irregular ratings and assign different weights to individual evaluators automatically based on rating differences among evaluators. The basic idea is that the weight of a rating on a certain target depends on the difference between that rating and all other ratings on the target.

Design Rationale

In most traditional group evaluation methods, all evaluators contribute to the overall group outcome equally, meaning that each evaluator or rating shares the same weight. Unlike those methods, the proposed *Collaborative Evaluation based on rating DIFFerence* (CE-DIFF) method automatically determines and assigns different weights to individual evaluators based on differences between their ratings and the overall group rating. The weight values range from 0 to 1, which reflect the significance of contribution from individual

evaluators to the final group decision. The larger a weight value is, the more contribution an evaluator makes to the final group outcome. This methodological design hinges on an assumption that the assessment of a target by the majority of a group should be generally more reliable and accurate than that of the minority (Appelt et al., 2011; de Bruin et al., 2012). The more deviation a rating of an individual evaluator is from the group rating, the higher possibility of irregularity the rating could have, resulting in a smaller weight value for that individual or rating.

In reality, an evaluator may deliberately give an unreasonably high or low rating to an evaluated target due to personal bias or other reasons. In *CE-DIFF*, different weights can be assigned to an evaluator for different evaluated targets. Specifically, one will be assigned larger weights for targets when his/her evaluation is similar to the initial group outcome, or with smaller weights otherwise. A proxy judges whether a rating is fair or irregular based on the difference between that rating and the group average rating.

CE-DIFF adopts the iterative and reciprocal rationale of the PageRank algorithm (Brin & Page, 1998) that is used to estimate the importance of webpages a priori. According to Google, PageRank works by counting the number and quality of links to a website to determine how important the website is. The underlying assumption is that more important websites are likely to be linked to by many other websites. Specially, if a webpage *u* has a link to another page *v*, then the author of *u* is implicitly conferring some importance to *v*. For example, *Yahoo.com* is an important website, reflected by the fact that many other websites point to it. Reciprocally, webpages pointed to by *Yahoo.com* are probably important as well. Assuming N_u is the out-degree of webpage *u*, and $PR(p)$ represents the importance of webpage *p*, then the link (*u*, *v*) confers $PR(u)/N_u$ units of rank to *v*. This idea leads to the computation that yields a rank vector $Rank^*$ over all webpages. If *N* is the total number of

pages, the pages are initially assigned a value of 1/*N*. Let B_v represent the set of webpages pointing to *v*. In each iteration, the *PageRank* value of *v* at the k+1$_{\text{th}}$ iteration is propagated as follows:

$$PR_v^{k+1} = \sum_{u \in B_v} PR_u^k / N_u \quad (1)$$

The iterations work till the vector $\overrightarrow{PR}$ stabilizes. The final vector $\overrightarrow{PR^*}$ contains the PageRank values of all the pages. In the PageRank algorithm, the importance of each page depends on other pages' weights. An iteration method is used to determine the importance of all webpages. The weights of all webpages are propagated based on initial weight values. Similar to the PageRank algorithm, in CE-DIFF, the weight W_{ij}^{k+1} of the rating R_{ij} at the k+1$_{\text{th}}$ iteration is propagated as follows:

$$W_{ij}^{k+1} = f(W_{ij}^k, R) \quad (2)$$

where *f* is the CE-DIFF algorithm, which will be described in the following subsection.

The CE-DIFF Algorithm

Let us first define some concepts in CE-DIFF.

Definition 1. *Group rating*: A group rating is an overall aggregated value of the ratings given by all evaluators based on an evaluation method. In this paper, the group rating on target *j* at the k_{th} iteration is represented as $Score_j^k$, and the final group rating on target *j* is represented as $Score_j$.

Definition 2. *Average weight*: For an evaluation task, the average weight of a rating is equal to 1 divided by the number of evaluators.

Definition 3. *Evaluation weight*: The different ratings made by different evaluators on a certain target will be assigned different weight values called *evaluation weights*. An evaluation weight represents the contribution of a rating to the whole group evaluation. In general, if a rating is significantly different from most of the ratings made by other evaluators, a small weight will be assigned. Otherwise, a large weight will be assigned. The sum of evaluation weights of all ratings on a certain target by all evaluators is 1. During the implementation of CE-DIFF, evaluation weights are updated iteratively till their values converge steadily. The evaluation weight of the rating R_{ij} at the k_{th} iteration is represented as W_{ij}^{k} in this paper.

Definition 4. *Actual weight*: When an evaluation weight converges, the weight at this time is referred to as actual weight. In other words, an actual weight is the final evaluation weight. In this paper, the actual weight of the rating R_{ij} is represented as W_{ij}.

Similar to the above iterative nature of PageRank, CE-DIFF initially assigns all ratings on a target with an initial weight value equal to 1/*N* (*N* is the number of evaluators), then it computes the group rating by calculating the weighted sum of ratings of all the evaluators on the target. According to the group rating, CE-DIFF will update evaluation weights, then re-compute the group rating. The above steps will be repeated till the group rating stabilizes. Specifically, *CE-DIFF* works in the following steps:

Step 1: For each evaluated target O_j, *CE-DIFF* calculates the initial group rating by calculating the weighted sum of all ratings given by evaluators with the initial actual weight.

Step 2: the relative difference V_{ij}^{k} between the group rating of target O_j and the rating given by the evaluator E_i on O_j at the k_{th} round is calculated as follows:

$$V_{ij}^{k} = \frac{\left|R_{ij} - Score_j^k\right|}{\delta_j}, i = 1,2,\dots,N,\ j = 1,2,\dots,M \quad (3)$$

where R_{ij} represents the rating of E_i on target O_j; $Score_j^k$ represents the group rating on O_j at the k_{th} round, and δ_j represents the standard deviation of the ratings on O_j.

Step 3: Computing the actual weight of each rating by evaluator E_i on each target O_j as follows:

$$D_{ij}^{k} = \max\left(0, 1 - V_{ij}^{k}\right),\ i = 1,2,\dots,N,\ j = 1,2,\dots,M \quad (4)$$

Step 4: Normalize the weight of each evaluator for each target:

$$W_{ij}^{k} = \frac{D_{ij}^{k}}{\sum_{i=1}^{n} D_{ij}^{k}} \quad i = 1,2,\dots,N,\ j = 1,2,\dots,M \quad (5)$$

Step 5: Get the group rating of target O_j for the next iteration by summation of ratings adjusted by weights:

$$Score_j^{k+1} = \sum_{i=1}^{n} W_{ij}^{k} * R_{ij} \quad i = 1,2,\dots,N,\ j = 1,2,\dots,M \quad (6)$$

The final group rating will be obtained at the end of iterations when the evaluation score is stabilized, as shown in Figure 1. In our study, the final group rating was considered stabilized when the actual weights of evaluators on targets did not change compared with those of the last iteration, which resulted in the rank of the targets unchanged. In our study, the result was stabilized after five iterations.

Algorithm: *Collaborative Evaluation Based on Rating Difference*
Input: Evaluating Matrix R=$\{R_{ij}\}_{N\times M}$
Output: Final Group Rating

For all $O_j \in O$

$Score_j^0 = Rand()$;

End For

$Score^0 \leftarrow \{Score_1^0, Score_2^0, \ldots, Score_M^0\}$;

$Rank^0 \leftarrow$ Rank $(Score^0)$;

$k = 0$;

Do

k=k+1;

For all $R_{ij} \in R$

$V_{ij}^k = \frac{\left|R_{ij} - Score_j^k\right|}{\delta_j}$

$D_{ij}^k = \max(0, 1 - V_{ij}^k)$

$W_{ij}^k = \frac{D_{ij}^k}{\sum_{i=1}^n D_{ij}^k}$

End For

For all $O_j \in O$

$Score_j^{k+1} = \sum_{i=1}^n W_{ij}^k * R_{ij}$;

End For

$Score^k \leftarrow \{Score_1^k,\ Score_2^k,\ \ldots,\ Score_m^k\}$;

$Rank^k \leftarrow$ Rank $(Score^k)$;

While $(Rank^{k-1} \neq Rank^k)$

Return $Score^k$

Figure 1: Algorithm of CE-DIFF

Identifying Irregular Ratings Based on the 3-Sigma Rule

In *CE-DIFF*, an actual weight reflects the degree of the contribution of an evaluator to the whole group's evaluation. When the difference between an evaluator's rating and the

whole group rating is significant, the actual weight of his/her rating would be very small, implying that the evaluator's rating is irregular to some degree. Therefore, an irregular rating could be identified based on its actual weight. According to Xu and Luo (2019), decision makers' evaluations (i.e., ratings) generally follow the normal distribution, which can provide a realistic approximation of the distribution of deviations of individual decision makers' preferences in group decision making processes. Therefore, the final actual weights of the ratings made by a group of evaluators should also follow the normal distribution.

According to the 3-sigma rule in the statistics science, we classify the ratings made by all evaluators into three categories based on their actual weights, including irregular ratings, regular ratings, and close-to-group-mean ratings. Irregular ratings can be further classified into two sub-categories, namely extremely irregular ratings and moderately irregular ratings, depending on the extent of deviations from the group mean rating, as shown in Table 1. The rationale behind our classification is that extremely and moderately irregular ratings will carry less weights in determining the final group rating than regular and close-to-group ratings.

Table 1: Different Types of Ratings

Actual weights	Types of Ratings	
$(0, \mu_w - 2\sigma_w]$	Extremely irregular ratings	Irregular ratings
$(\mu_w - 2\sigma_w, \mu_w - \sigma_w]$	Moderately irregular ratings	
$(\mu_w - \sigma_w, \mu_w + \sigma_w]$	Regular ratings	
$(\mu_w + \sigma_w, 1)$	Close-to-group ratings	

Note: μ_w *and* σ_w are the mean and standard deviation of all actual weights of a target, respectively.

EVALUATION

Experiment Design

In this study, we conducted two controlled laboratory experiments for evaluating the proposed CE-DIFF, with the first one to assess the accuracy of the proposed method, and the second one to investigate the ability of the proposed method to deal with irregular ratings.

Experiment 1: estimation of height and weight of objects

This experiment included two evaluation tasks: one was to estimate people's height, and the other was the estimation of the weight of different objects. Each task was completed by three groups with different sizes: a large group with 50 evaluators, a medium-size group with 15 evaluators, and a small group with 5 evaluators, aiming to gain more insights into the potential impact of group size on the effectiveness of the proposed approach. Each participant only participated in one group.

Seventy college students were recruited as the participants of the experiment. Among them, twenty were undergraduates whose age range was between 18~22 years old; thirty-five were master students between 23~24 years old; and the remaining fifteen students were doctoral students who were between 25~28 years old. Among them, 40 were male. The majors of the participants included management information systems, management science, and engineering. No incentives were provided to the participants.

In the first experiment task, the participants were instructed to view the printed standing pictures of 10 people one by one, then write down their best estimates of those people's heights independently. Then, they were required to estimate the weights of ten different objects, such as a potted plant, a book, and a PC. Each participant (i.e., evaluator) could visually inspect and touch the objects before estimation. The evaluators were required to

write down their best estimations of weights of the ten objects independently without discussing with anyone else. Before the above two experimental tasks began, the researchers measured the true values of ten target people's heights and ten target objects' weights as the ground truth.

Experiment 2 : course presentation evaluation

The second experiment was designed to investigate the effectiveness of the proposed method in dealing with irregular ratings in collaborative evaluation. A total of 16 college students in a graduate class were recruited to participate in this experiment. None of them participated in the experiment 1. Each participant was required to give a presentation related to the content of the course. At the same time, they were required to provide ratings in the scale of 0 to 100 for all other presentations, with 100 being perfect and 0 being the worst. In order to assess the performance of the proposed CE-DIFF method in dealing with irregular ratings, by simulating the rationale of sensitivity analysis, we deployed a stepwise approach by continuously changing one rating value incrementally while keeping other ratings intact, then compared the changes of the final group rating. For example, without loss of generality, the rating $R_{3,4}$ (i.e., the rating of the participant P_3 given to the participant P4) was selected and increased from 0 to 100 step by step, with the incremental size equal to 10 (i.e., 0, 10, 20, 30, … 90, 100). For each increment, we calculated the group rating result by using CE-DIFF. We observed the change of the ranking of the evaluated targets when varying one rating. If the change was small, it meant that the ranking was not significantly affected by the irregular rating. As a result, CE-DIFF should be considered robust because it dealt with irregular ratings effectively.

The idea behind this evaluation was that if the group final result does not change significantly when an irregular rating appears, then it can be argued that the group result

generated by CE-DIFF is not significantly affected by an irregular rating (i.e., can handle it well) and is thus robust.

Baselines and Metrics

In order to assess the performance of the proposed CE-DIFF method, we selected the widely used arithmetic mean (*arith_mean*) and trimmed mean (*trim_mean*) methods as baselines. With *arith_mean*, the group evaluation result is generated by averaging the ratings given by all evaluators. The *trim_mean* method is an improved *arith_mean* method, as introduced earlier. In our experiment, we followed the most common practice by dropping 20% of all ratings in the *trim_mean* method.

For the first experiment, with regard to performance measures, researchers measured the real height and weight values of the targets before the experiment as the ground truth. Considering the different rating scales used in the two evaluation tasks, we employed *Mean Relative Error* (MRE) as a performance metric:

$$\mathrm{MRE} = \frac{1}{m}\sum_{j=0}^{m}\frac{|S_j - T_j|}{T_j} \tag{7}$$

where S_j represents an estimated value (e.g., estimated height) of a target O_j; T_j is the real value of O_j, and m is the total number of the evaluated targets.

EXERIMENT RESULT

Result of Experiment 1

For the height estimation task, the MREs of the three evaluation methods are shown in Table 2, in which the best result of each column (i.e., a target being estimated) is highlighted. For simplicity, we represent the arith_mean, trim_mean, and CE-DIFF methods by A, T, and CE, respectively, in all the result tables hereafter. As Table 2 shows, CE-DIFF consistently

achieved a better overall performance than the two baseline methods by achieving the least MREs for 8, 7, and 10 out of 10 targets in the small, medium, and large groups, respectively. We conducted pairwise t-tests to evaluate the significance of those differences in MREs among the three methods. Table 3 shows that CE-DIFF resulted in significantly lower MREs than both the *arith_mean* and *trim_mean* methods at a 0.05 significance level for the medium-size groups and at a 0.01 significance level for the large group. For the small group, the CE-DIFF had a lower MRE than the *arith_mean* method at a 0.01 significance level, and a lower MRE than *trim_mean* at a 0.05 significance level. Therefore, in the height estimation task, the CE-DIFF method consistently showed superior performance in comparison to the baseline methods across groups with different sizes.

Table 2: Mean of MREs of height estimation (%)

Groups	Methods	O1	O2	O3	O4	O5	O6	O7	O8	O9	O10
Large	*A*	6.40	4.68	3.81	2.04	3.12	1.49	3.63	3.35	6.38	2.71
	T	6.39	4.53	3.79	2.07	3.12	1.22	3.68	3.27	6.34	2.18
	CE	**6.32**	**3.98**	**3.51**	**1.98**	**2.88**	**0.32**	**3.59**	**3.20**	**6.27**	**1.68**
Medium	*A*	7.22	5.83	**4.04**	1.57	2.82	4.31	3.97	2.91	**6.74**	3.17
	T	7.29	5.87	4.19	1.81	2.90	4.18	**3.72**	2.83	6.76	2.83
	CE	**7.04**	**5.70**	4.15	**1.71**	**2.71**	**3.76**	3.90	**2.76**	6.79	**2.63**
Small	*A*	4.66	4.88	4.27	1.41	5.31	4.72	4.47	4.00	2.84	2.17
	T	4.40	6.09	**3.51**	2.50	5.51	3.65	**3.86**	3.24	**2.63**	2.86
	CE	**4.24**	**4.68**	3.81	**1.18**	**4.72**	**3.62**	4.29	**2.67**	**2.63**	**1.67**

Table 3: Comparison of mean MREs of methods for height estimation

Groups	Method 1 (M1)	Method 2 (M2)	MRE of M1	MRE of M2	Differences (M1-M2)	P Values
Large	*A*	CE	3.76%	3.37%	0.39%	0.0089
	T	CE	3.66%	3.37%	0.29%	0.0053
Medium	*A*	CE	4.26%	4.12%	0.14%	0.0215
	T	CE	4.24%	4.12%	0.12%	0.0482
Small	*A*	CE	3.87%	3.35%	0.52%	0.0012
	T	CE	3.83%	3.35%	0.48%	0.0270

For the weight estimation task, the MREs of the three evaluation methods are illustrated in Table 4, in which the best result of each column is highlighted.

Table 4: Means of MREs of weight estimation (%)

Groups	M	O1	O2	O3	O4	O5	O6	O7	O8	O9	O10
Large	A	8.00	12.05	3.96	1.66	122.42	27.84	**3.56**	5.08	18.00	1.76
	T	7.61	1.71	0.76	0.86	94.08	23.90	3.75	**5.03**	10.00	2.06
	CE	**5.93**	**0.25**	**0.24**	**0.13**	**80.21**	**2.92**	**3.56**	5.06	**5.17**	**1.33**
Medium	*A*	**13.00**	**11.90**	23.10	20.80	67.20	38.20	2.00	0.80	5.80	20.80
	T	14.20	23.40	19.80	16.00	63.20	34.50	2.10	**0.20**	6.80	21.90
	CE	15.60	17.20	**17.80**	**11.30**	**57.90**	**18.80**	**1.60**	0.30	**0.60**	**17.60**
Small	*A*	30.43	31.05	22.83	13.27	300.00	27.20	1.66	2.53	61.67	15.88
	T	43.48	53.29	39.95	19.47	373.68	48.00	0.58	4.47	87.50	30.15
	CE	**2.41**	**6.09**	**21.82**	**10.87**	**133.65**	**14.84**	**0.38**	**1.05**	**15.00**	**4.77**

According to Table 4, CE-DIFF achieved the lowest MREs with 10, 7 and 9 of the 10 targets in the small, medium, and large groups, respectively. We also conducted pairwise t-tests to evaluate the significance of differences in MREs of the three methods. The results shown in Table 5 reveal that the proposed CE-DIFF consistently resulted in lower MREs than

arith_mean at a 0.05 significance level across all three groups and lower MREs than *trim_mean* method at a 0.05 significance level for the small and large groups and at a 0.01 significance level for the medium-size group.

Table 5: Comparison of mean MREs of methods for weight estimation

Groups	Method 1 (M1)	Method 2 (M2)	MRE of M1	MRE of M2	Differences (M1-M2)	P Values
Large	*A*	CE	20.43%	10.48%	9.95%	0.0248
	T	CE	14.98%	10.48%	4.50%	0.0390
Medium	*A*	CE	20.40%	15.90%	4.50%	0.0377
	T	CE	20.20%	15.90%	4.30%	0.0098
Small	*A*	CE	50.65%	21.09%	29.56%	0.0481
	T	CE	70.06%	21.09%	48.97%	0.0281

Identifying Irregular Ratings

We also investigated how effective irregular ratings could be identified by CE-DIFF. The number and percentage of penalized ratings in each experiment are shown in Table 6. It is found that the proportion of irregular ratings for each estimation task was approximately 15%~20%.

Table 6: Numbers of ratings in different categories in height and weight estimation tasks

Group Tasks	Groups	# of Extremely Irregular Ratings	# of Moderately Irregular Ratings	# of Regular Ratings	# of Close-to-group Ratings	Percentages of Irregular Ratings (%)
Height Estimation	Large	39	43	375	43	16.4
	Medium	12	11	90	37	15.3
	Small	5	3	38	4	16.0
Weight Estimation	Large	33	41	383	43	14.8
	Medium	9	21	92	28	20.0
	Small	5	4	36	5	18.0

We further analyzed the penalty situation in the height estimation task of each target. For the small group, all of the targets had no more than one evaluator who received capital or high penalty. Additional analysis on experiments of the medium-size group showed that 3 or 4 of 15 evaluators received capital or high penalty on targets O_4, O_5, and O_9, while the number on any other target was no more than 2. Moreover, for the large group, targets O_1, O_2, O_5, O_9, and O_{10} had 9 or 10 of the 50 evaluators received capital or high penalty, while targets O_3, O_4, O_6, O_7, and O_8 had no more than 8 evaluators who received capital or high penalty. These results indicate that assigning different weights rather than equal weights to individual evaluators on different targets is necessary and beneficial.

Results of Experiment 2

In order to examine the change of rank of O_4 by various methods, we computed the variances of its rank with different values of $R_{3,4,}$ as shown in Figure 2. When using the *arith_mean* method, the target O_4 got a final group rating of 78.79, and was ranked the 11th when $R_{3,4}$ was changed to 0. The group rating increased to 79.79 with its ranking moving up to the 7th when $R_{3,4}$ was changed to 100. In general, the final group rating of O_4 would be higher as $R_{3,4}$ increased, which was demonstrated based on the observation that the rank of the target O_4 continuously moved up as $R_{3,4}$ increased, indicating that the *arith_mean* method is sensitive to and influenced by irregular ratings.

The variances in the rank of target O_4 for *arith_mean*, *trim_mean*, and CE-DIFF methods were 42.25, 0.22, and 8.02, respectively. The variance in the rank of CE-DIFF was smaller than that of the *arith_mean* method, which indicates that the evaluation result by CE-DIFF is more stable. On the other hand, the rank of O_4 by *trim_mean* almost did not change as the value of $R_{3,4}$ varied because the method dropped more ratings than *arith_mean* and CE_DIFF.

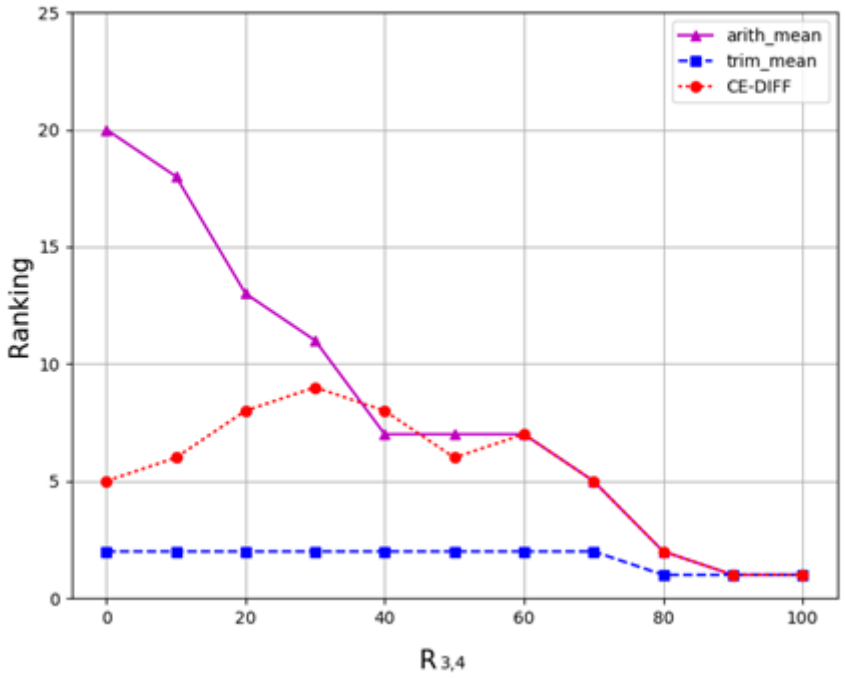

Figure 2: Rank changes of O_4 as $R_{3,4}$ increases

The linear variation of evaluation results with the increase of $R_{3,4}$ shows that the evaluation results derived by the *arith_mean* method can be easily affected by a single irregular evaluation. In contrast, CE-DIFF produced more stable evaluation results even though $R_{3,4}$ varied significantly.

DISCUSSION

There are several major findings of this study. First, it is difficult to conduct a comparative evaluation of this nature because it is extremely challenging to get the ground truth of group evaluation. In this study, we designed two experiments in order to validate our proposed method. The first experiment was mainly conducted to verify the accuracy of our method, and the second experiment was used to test the ability of our method in dealing with irregular ratings. The experimental results show that the proposed CE-DIFF method led to significantly better performance in terms of MRE than the two baseline methods across

groups with different sizes, suggesting that the effectiveness of the proposed method is independent of group size. Also, the result of the second experiment showed that the accuracy of CE-DIFF did not change significantly when an irregular rating occurs, showing good stability.

It is worth noting that in the first experiment, because evaluators were relatively familiar with estimating the height of a person and the range of height change was small, the overall evaluators' evaluation errors were small. Therefore, the errors of CE-DIFF and two baseline methods were relatively small. However, in the task of weight evaluation, because judging the weight of an object was more difficult, the evaluation errors of evaluators were bigger, and the differences in MREs between the CE-DIFF method and the two baseline methods were more obvious. It indicates that the familiarity with or complexity of collaborative evaluation tasks may have a moderating effect on the positive impact of CE-DIFF on group evaluation results, which is worthy of further investigation in the future.

In general, there are two possible reasons why irregular ratings are made: motivational bias and cognitive limitation. CE-DIFF can deal with irregular ratings caused by motivational bias well by assigning different evaluation weight values to ratings on different targets given by an evaluator. However, in many collaborative evaluation tasks, it is possible that irregular ratings are caused by insufficient domain knowledge of evaluators, or cognitive limitation. In such a situation, an evaluator may provide assessments significantly different from group assessments consistently across targets, in contrast to occasional deviated rating scores on some targets that are more likely to be caused by motivational bias.

Considering this situation, it is possible to develop a simple version of CE-DIFF (CE-DIFF-SPL) that focuses on addressing irregular ratings due to insufficient domain knowledge and reducing computational complexity. The main difference between CE-DIFF

and CE-DIFF-SPL lies in that the same evaluator may have different weight values in assessment of different targets in the former, while the same evaluator will have the same weights across all targets in the latter. In CE-DIFF-SPL, an evaluator will only be assigned one weight generated based on his/her general rating accuracy.

Another key issue is about application scenarios of CE-DIFF. In general, CE-DIFF is a generic collaborative evaluation method and can be widely applied to various group decision making tasks in practice, including large group decision making that involves a large number of participations. For example, in the e-commerce field, many customers rate or review products online. However, it is possible that a few merchants could deliberately sabotage their competitors' ratings through fake product ratings or/and reviews. So it is important that e-commerce platforms can recognize irregular ratings automatically and assign low weights to them in order to get a fair group evaluation result.

It is worth noting that if there are only two evaluators, there will not be a majority or minority opinion. The group rating is decided by them equally. Therefore, the deviation of each evaluator to the overall group rating will be the same, so are the actual weights of both evaluators, which makes it identical to the method of *arith_mean* in this special circumstance.

CONCLUSION

Collaborative evaluation is becoming increasingly popular nowadays. Although there have been a lot of studies on group decision making, up to now, most existing methods for group decision making have a common implicit assumption that all evaluators' individual judgments have the same importance to the final group assessment outcome while ignoring irregular judgments or treating irregular judgments ineffectively.

To deal with these problems, inspired by synergy theory and the collaboration notion of Web 2.0, in this paper, we first propose CE-DIFF, a novel collaborative evaluation method. In CE-DIFF, each rating has a different weight, which reflects the contribution of an individual evaluation to the collective group evaluation. The weights are collaboratively determined by all ratings of all evaluators. In other words, the weight of a rating is dependent not only on the rating itself, but also on other ratings of the same item. The closer a rating is to the group opinion, the greater its corresponding weight is. Since evaluation weights are automatically determined by all ratings, irregular ratings (e.g., extremely high and low ratings) will have minimal direct effects on final group result because the proposed method can deal with irregular ratings well. It is easy to identify irregular ratings that are significantly deviated from the group overall ratings. Furthermore, in this study, according to the 3-sigma rule in statistics, we classify all the ratings into three categories, including perfect ratings, regular ratings, and irregular ratings. The irregular rating category includes two sub-categories, namely moderately irregular ratings and extremely irregular ratings. When a rating is categorized as an extremely irregular rating, its weight is zero, which means that it would have no effect on the final group rating.

In the future, we will extend this research from two perspectives. First, this study focuses on group evaluation by collaborative rating. There are other types of group evaluation, such as comparing or ranking multiple targets. Therefore, it would be interesting to explore how CE-DIFF can be adapted or extended to other group evaluation tasks. Second, it is important to investigate the potential moderating effect of evaluators' familiarity with and complexity of group evaluation tasks on the impact of CE-DIFF.

REFERENCES

Appelt, K. C., Milch, K. F., Handgraaf, M. J., & Weber, E. U. (2011). The decision making individual differences inventory and guidelines for the study of individual differences in judgment and decision-making research. *Judgment and Decision Making*, *6*(3), 252-262.

Berkes, I., Györfi, L., & Kevei, P. (2016). Tail probabilities of St. Petersburg sums, trimmed sums, and their limit. *Journal of Theoretical Probability*, *30*(3), 1104-1129.

Briggs, R. O., De Vreede, G. J.D., & Nunamaker J. F. (2003). Collaboration engineering with thinklets to pursue sustained success with group support systems. *Journal of Management Information Systems*, *19*(4), 31-64.

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, *30*(1-7), 107-117.

Bruccoleri, M., Riccobono, F., & Groessler, A. (2019). Shared leadership regulates operational team performance in the presence of extreme decisional consensus/conflict: Evidences from business process reengineering. *Decision Sciences*, *50*(1), 46-83.

Contreras, I. (2011). Emphasizing the rank positions in a distance-based aggregation procedure. *Decision Support Systems*, *51*(1), 240-245.

De Bruin, W. B., Missier, F. D., & Levin, I. P. (2012). Individual differences in decision-making competence. *Journal of Behavioral Decision Making, 25*(4), 329–330.

Dennis, A. R., Wixom, B. H., & Vandenberg, R. J. (2001). Understanding fit and appropriation effects in group support systems via meta-analysis. *MIS Quarterly*, *25*(2), 167-193.

De Vreede, G.J., & Briggs, R. O. (2005). Collaboration engineering: designing repeatable processes for high-value collaborative tasks. *Proceedings of the 38th Annual Hawaii International Conference on System Sciences*, Big Island, HI, USA: IEEE, 1-10

De Vreede, G-J., & Briggs, R. O. (2019). A program of collaboration engineering research and practice: Contributions, insights, and future directions. *Journal of Management Information Systems*, *36*(1), 74-119.

Forsyth, D. R. (2006). Decision making. In D. R. Forsyth (Ed.), *Group dynamics (5th Ed.)*, Belmont, CA: Cengage Learning, 317-349.

Gao, J., & Liu, H. (2017). Generalized ordered weighted reference dependent utility aggregation operators and their applications to group decision-making. *Group Decision and Negotiation*, *26*(6), 1173-1207.

Haken, H. (2009). Synergetics: Basic concepts. In H. Haken (Ed.), *Encyclopedia of complexity and systems science*. New York, NY: Springer, 8926-8946.

Hochbaum, D. S., & Levin, A. (2006). Methodologies and algorithms for group-rankings decision. *Management Science*, *52*(9), 1394-1408.

Ishikawa, A., Amagasa, M., Shiga, T., Tomizawa, G., Tatsuta, R., & Mieno, H. (1993). The max-min Delphi method and fuzzy Delphi method via fuzzy integration. *Fuzzy Sets and Systems*, *55*(3), 241-253.

Jeh, G., & Widom, J. (2002). SimRank: A measure of structural-context similarity. *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Edmonton, Alberta, Canada: ACM, 538-543.

Jessup, L. M., & George, J. F. (1997). Theoretical and methodological issues in group support systems research: Learning from groups gone away. *Small Group Research*, *28*(3), 394-413.

Kanisauskas, S. (2014). The phenomenon of creativity: Philosophical and synergetic insights. *European Scientific Journal*, *10*(14), 159-168.

Kaplan, M. F., & Miller, L. E. (1978). Reducing the effects of juror bias. *Journal of Personality and Social Psychology*, *36*(12), 1443-1455.

Kim, Y. (2006). Supporting distributed groups with group support systems: A study of the effect of group leaders and communication modes on group performance. *Journal of Organizational and End User Computing*, *18*(2), 20-37.

Knyazeva, H., & Haken, H. (1999). Synergetics of human creativity. In W. Tschacher, & J. Dauwalder (Eds.), *Dynamics, Synergetics, Autonomous Agents: Nonlinear Systems Approaches to Cognitive Psychology and Cognitive Science.* Singapore: World Scientific, 64-79.

Kolfschoten, G. L., & De Vreede, G. J. (2009). A design approach for collaboration processes: A multimethod design science study in collaboration engineering. *Journal of Management Information Systems*, *26*(1), 225-256.

Labella, Á., Liu, Y., Rodríguez, R. M., & Martínez, L. (2018). Analyzing the performance of classical consensus models in large scale group decision making: A comparative study. *Applied Soft Computing*, *67*(C), 677-690.

Lima, A. S., de Souza, J. N., Moura, J. A. B., & da Silva, I. P. (2018). A consensus-based multi-criteria group decision model for information technology management committees. *IEEE Transactions on Engineering Management*, *65*(2), 276-292.

Liu, P., & Liu, X. (2018). The neutrosophic number generalized weighted power averaging operator and its application in multiple attribute group decision making. *International Journal of Machine Learning and Cybernetics*, *9*(2), 347-358.

Liu, Y., Fan, Z. P., You, T. H., & Zhang, W. Y. (2018). Large group decision-making (LGDM) with the participators from multiple subgroups of stakeholders: A method considering both the collective evaluation and the fairness of the alternative. *Computers & Industrial Engineering*, *122*, 262-272.

Morente-Molinera, J. A., Pérez, I. J., Ureña, M. R., & Herrera-Viedma, E. (2015). On multi-granular fuzzy linguistic modeling in group decision making problems: A systematic review and future trends. *Knowledge-Based Systems*, *74*(1), 49-60.

Nahartyo, E., & Utami, I. (2014). Altering rationality: The impact of group support systems and style of leadership. *Journal of Applied Management Accounting Research*, *12*(2), 41-57.

Palomares, I., Martinez, L., & Herrera, F. (2013). A consensus model to detect and manage noncooperative behaviors in large-scale group decision making. *IEEE Transactions on Fuzzy Systems*, *22*(3), 516-530.

Pyrko, I., Eden, C., & Howick, S. (2019). Knowledge acquisition using group support systems. *Group Decision and Negotiation*, *28*(2), 233-253.

Talluri, S., DeCampos, H. A., & Hult, G. T. M. (2013). Supplier rationalization: A sourcing decision model. *Decision Sciences*, *44*(1), 57-86.

Van Hillegersberg, J., & Koenen, S. (2014). Adoption of web-based group decision support systems: Conditions for growth. *Procedia Technology*, *16*, 675-683.

Walsh, K., & Dickey, M. (2004). Structured modeling group support systems: A product design theory. *Information & Management*, *41*(5), 655-667

Wang, J., Ding, D., Liu, O., & Li, M. (2016). A synthetic method for knowledge management performance evaluation based on triangular fuzzy number and group support systems. *Applied Soft Computing Journal*, *39*, 11-20.

Wilson J. L., Griffin, T. E., & Jessup, L. M. (2010). GSS anonymity effects on small group behavior. Journal of Management Information and Decision Sciences, *13*(2), 41-57.

Xu, X. H., & Luo, X. (2019). Information entropy risk measure applied to large group decision-making method. *Soft Computing*, *23*(13), 4987-4997.

Zhang, B. W., Liang, H., & Zhang, G. Q. (2018). Reaching a consensus with minimum adjustment in MAGDM with hesitant fuzzy linguistic term sets. *Information Fusion*, *42*, 12-23.

Zhang, Z., & Pedrycz, W. (2018). Intuitionistic Multiplicative group analytic hierarchy process and its use in multicriteria group decision-making. *IEEE Transactions on Cybernetics*, *48*(7), 1950-1962.

Zheng, E., Teng, F., & Liu, P. (2016). Multiple attribute group decision-making method based on neutrosophic number generalized hybrid weighted averaging operator. *Neural Computing and Applications*, *28*(8), 2063-2074.