Satish S. Salunkhe*, Ashok Deshpande and Yashwant Joshi

Degree of Certainty in Students' Academic Performance Evaluation Using a New Fuzzy Inference System

DOI 10.1515/jisys-2016-0293

Received November 10, 2016; previously published online April 12, 2017.

Abstract: The academic performance assessment of students helps teachers, administrators, and policy-makers to initiate corrective measures on academically poor students. This paper revisits the Zadeh-Deshpande formalism for evaluating students' answer scripts using the concept of the reliability of information (degree of confidence) via the "degree of match" and fuzzy inference system in students' performance evaluation. The case study infers that the overall performance of all the students is "average". Furthermore, 206 of 237 students (87%) are declared as "average" with "high degree of certainty" by the evaluators (teachers). The aim of the proposed method is not to replace the traditional method of evaluation. Instead, the proposed technique is a step forward to enrich the present system of students' performance assessment. Policymakers can use this method as it provides reliable information. A comparison between the results obtained from the academic performance of students using multiexpert and single-expert systems is also discussed in the paper.

Keywords: Academic performance assessment, degree of certainty, degree of match, examiner, fuzzy inference system, κ coefficient, probability density function, Zadeh-Deshpande (ZD) formalism.

1 Introduction

The educated young population is the knowledge potential of a country. One of the reasons why academic institutions are trying to upgrade their educational practices in elementary and secondary schools is to hone the educated youngsters with knowledge potential. The examiner and the students are the backbones of the educational grading system. It is a common belief that the elicitation and aggregation of expert's knowledge base (KB) having similar perceptions/opinions might tend to yield better results in the formulation of a multi-expert KB system in students' performance evaluation. Evaluation should be valid and reliable. It is argued that the subjective answer scripts' evaluation is often imprecise and biased. The evaluation procedure should preferably be criterion referenced, qualitative, and, whenever possible, holistic [2]. Several researchers [3, 7–9, 15, 26, 44] presented methods for students' evaluation. A two-valued logic-based probability theory and statistical techniques are suggested for computing the similarity between examiners [28, 30, 36]. Reliable information helps in taking better decisions [47, 49, 50]. With this backdrop, the first step in the evaluation of student's academic performance is to compute the similarity between the teachers having somewhat similar perceptions. An approach for the similarity computation among teacher/examiner/evaluator using "cosine amplitude", one of the fuzzy similarity measures, was studied by the authors [39, 40]. The study infers that 11 of 20 evaluators/examiners are selected for the evaluation of academic performance of secondary school students. The present assessment system up to high school level in India is based on the aggregated percentage

---

*Corresponding author: **Satish S. Salunkhe**, Amity Institute of Information Technology, Amity University Mumbai 410206, India, e-mail: satishssalunkhe@gmail.com

Ashok Deshpande: Berkeley Initiative Soft Computing-SIG-EMS, Berkeley, CA 94720-5800, USA; and College of Engineering, Pune 411005, India

Yashwant Joshi: Shri Guru Govind Singh Institute of Engineering and Technology, SRTM University, Nanded 431606, India

of marks. Depending on the “Board of Education”, students' academic performance is expressed as “excellent”, “distinction”, “first class”, “second division”, “pass”, and “fail” [22].

The human brain analyzes information based on perceptions or tacit knowledge while taking any decisions and seldom with numbers. Zadeh [48] stated that:

“As the complexity of a system increases, our ability to make precise and yet significant statements about its behavior diminish until a threshold is reached beyond which precision and significance (or relevance) becomes almost exclusive characteristics”.

Traditional student assessment approaches lack in assigning the experts' “degree of certainty” as reliability to student's performance level. Why not describe the performance of a student straight away in linguistic terms such as “very poor with fair degree of certainty” with more reliability? In the following sections, the authors address the complexity of subjective qualitative assessment by developing a fuzzy criterion-referenced qualitative assessment system. We present the Zadeh-Deshpande (ZD) formalism [16] based on complexity level 1 in Computing with Words (CW) for the evaluation of student's academic performance straight away in linguistic terms with the concept of degree of match (DM) attached to each performance level as shown in Figure 1.

The outline of the paper is as follows: Section 2 is concerned with related works in students' performance evaluation system. Section 3 relates to mathematical preliminaries, whereas an approach for criterion-based evaluation of students' academic performance using the ZD method is described in Section 4. The case study for classifying students' performance is covered in Section 5. The results and discussion of the case study are discussed in Section 6. The conclusion and future scope for research is an integral part of Section 7.

Figure 1: Four-Level Hierarchical Multiexpert FRBS with DM Computations in the Performance Evaluation of Student ID-02.

2 Related Works

Biswas [3] proposed a new evaluation method using the fuzzy set theory in education. More time consumption in the assessment process and lack of sensitivity as multiple fuzzy marks are mapped to same performance letter grade resulted in a new evaluation methodology in a fairer manner as proposed by Chen and Lee [9]. Mendel et al. [32] showed student-centric learning method for better understanding of the assessment criteria to students. Vague sets, instead of fuzzy sets, are used to evaluate students' answer scripts using a measure of optimism of the evaluator [13, 14, 32, 45]. Subjective criteria assessment of students' teamwork in the entire course is considered in academic performance evaluation [33]. Fuzzy logic-based computing methods can be used to develop an intelligent system [41]. Using fuzzy rule-based system (FRBS), a new method is suggested to rank students, considering the characteristics of questions [1, 10–12, 29, 37]. Evaluator's confidence is considered to evaluate students' answer scripts using fuzzy numbers [46]. A hierarchical structure is also suggested for performance evaluation considering vagueness [19]. It is difficult to automate the manual evaluation process of student's performance in subjective test altogether [4]. Centroid index is used in calculating the ranking order of the alternatives [35]. A multicriteria evaluation model having multistage academic personnel evaluation is presented [5]. An optimized multiexpert knowledge aggregation model proposed for multidecision student's evaluation system is suggested [4, 34]. Fuzzy set theory-based methods are recommended to minimize the difference in score in the multiexpert system. The reason behind using this approach is variation in the subjectivity of different evaluators to evaluate the same student [25]. During the course, parameters of fuzzy inference system (FIS) can be adjusted to reflect the pedagogical needs [26]. Gaussian membership function (MFs) can improve the reliability and robustness of systems compared to other MFs [20]. Based on the experimental study, the impact of ethnicity on students' grades is observed [43]. Aggregation operators on discrete fuzzy numbers are suggested in students' evaluation system [44]. A composite operation is used in decision making to rank students based on academic performance [21]. An orientation process is suggested for users of a simulation evaluation instrument to provide a framework for nurse educators to improve the consistency of student performance evaluation [31]. A new fuzzy peer assessment methodology that considers vagueness and imprecision of words used throughout the evaluation process in a cooperative learning environment is proposed [6]. The use of an FIS as an alternative to the conventional addition or weighted addition in criterion-referenced assessment is investigated [23, 24, 27, 38, 42], which possibly relates to level 2 complexity in CW using perceptual computing and fuzzy ranking algorithm.

In [46], Wang and Chen pointed out that the methods presented in [9] have the following drawbacks: (1) it cannot deal with the situation where the evaluating values are represented by fuzzy numbers associated with degrees of confidence between 0 and 1 and (2) it does not consider the extent of optimism of the evaluator in evaluating students' answer scripts.

The methods presented in [9] have two drawbacks for students' answer script evaluation. Hence, Wang and Chen [46] introduced a new approach for evaluating students' answer scripts using fuzzy numbers associated with degrees of confidence of the evaluator. Wang and Chen used the arithmetic operations between the α -cuts of fuzzy numbers to evaluate the total mark of each student, where $\alpha \in [0, 1]$. Although the method presented in [46] is more stable and flexible to assess the students' answer scripts than the methods presented in [9], the limitations of the method are stated below:

1. Invariably, the basis of constructing fuzzy sets is the aggregated tacit KB of the domain experts. It is not very clear from the paper as to whether the triangular fuzzy number (TFN) representation for the degree of satisfaction (%) is based on expert's KB.
2. It may not be possible to express the level of confidence and also the index of optimism by an evaluator in numeric term between 0 and 1 by the examiner.

The above two drawbacks could be the limitations in the computation of students' total score using the Wang and Chen method [46]. ZD formalism is human (expert) centric and therefore helps to resolve the above shortcomings in evaluating students' answer scripts, which have been demonstrated in the paper.

3 Preliminaries and Notations

ZD formalism falls in the province of CW (or CWW). The techniques used in this approach are listed below.

3.1 Bootstrapping

Bootstrapping refers to random sampling with replacement. Bootstrapping allows assigning measures of accuracy to sample estimates. It falls in the broader class of resampling methods. The bootstrap estimate can be calculated as

$$B = \frac{1}{m} \sum_{i=1}^m x_i \quad (1)$$

where B is the sample mean for the collected numerical data x_i , and bootstrap mean $\bar{B}$ (say for 1000 samples) is computed as

$$\bar{B} = \frac{1}{n} \sum_{j=1}^n B_j \quad (2)$$

The standard deviation of the bootstrap mean is

$$S = \sqrt{\frac{1}{n} \sum_{i=1}^n (B_i - \bar{B})^2} \quad (3)$$

According to the central limit theorem, bootstrap distribution is approximately normal as sample size $n \rightarrow \infty$. The “bootstrap resample mean distribution” is shown in Figure 2 for 1000 samples.

Figure 2: Transformation of pdf $p(x)$ to Possibility Distribution, i.e. CNFN “A” for Question 1 of Student ID-02 after 1000 Iterations of Bootstrap Sampling.

3.2 Convex Normalized Fuzzy Number (CNFN)

The governing expressions (4) to (9) used in estimating the DM is explained in this section:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \quad (4)$$

where x is a random variable taking values in real line, μ is the mean, and σ is the standard deviation. The continuous random variable can be discretized using the distribution given in Equation (5):

$$\phi(x) = p(x), \quad x \in X \quad (5)$$

such that the probability of any x_i in the parametric domain, denoted by

$$p(x_i) = \phi\left(x_i + \frac{dx}{2}\right) - \phi\left(x_i - \frac{dx}{2}\right) \quad (6)$$

where dx is a small interval about x_i for $i = 1, 2, 3, 4, \dots, n$.

Randomness in parametric data can be simulated using Gaussian distribution with probability density function (pdf). Thus, if x_i is some point in the parametric domain x for which $p(x_i)$ is maximum, then

$$\mu_A(x) = \frac{p(x)}{p(x_i)} \quad (7)$$

This equation transforms a random variable into a CNFN “A” with membership grade $\mu_A(x)$, as shown in Figure 2, thereby characterizing the dynamic behavior of the governing parameter.

3.3 Degree of Match (DM)

It is true that “probabilistic information” (in our case, student’s marks) and “possibilistic information” (perception of teachers in evaluation) are not derivable (orthogonal) using an existing system of computations other than fuzzy logic. Professor Zadeh has elaborated on this issue in his paper [50]. Because the ZD approach is based on the matching of “possibility concept”, therefore, “probability distribution” obtained from bootstrapping is first transformed into “possibility distribution” and is termed as CNFN. CNFN (A) for field data and fuzzy number (F) characterizing linguistic terms such as student’s academic performance like “very poor” and “good” are matched to arrive at DM is, as shown in Figure 3, defined as

$$DM_{ff}(A, F) = \frac{\int \mu_{(A \cap F)}(x) dx}{\int \mu_A(x) dx} \quad (8)$$

where x denotes the universe and $\int \mu_{(A \cap F)}(x) dx$ is the membership grade for $(A \cap F)$. If A and F are discrete probability distributions, the measure is defined as

$$DM_{ff}(A, F) = \frac{\sum \mu_{(A \cap F)}(x) dx}{\sum \mu_A(x) dx}, \quad x \in X \quad (9)$$

The computed DM relates to uncertainty expressed as “possibility distribution” due to randomness in parametric data (student’s marks as CNFN) and “epistemic uncertainty” in expert’s perception in linguistic terms for the defined parameters.

Figure 3: DM between CNFN "A" (Question 1 Marks) and Fuzzy set "F" (Average Level).

3.4 Fleiss κ

Fleiss κ is a statistical measure of interrater reliability. The κ can be defined as

$$\kappa = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e} \quad (10)$$

The factor $1 - \bar{P}_e$ gives the degree of agreement that is attainable above chance, and $\bar{P} - \bar{P}_e$ provides the level of agreement achieved above chance. If the raters agree strongly, then $\kappa = 1$. If there is no agreement among the raters, then $\kappa < 0$ [17, 18].

3.5 Limitations of Mamdani FIS

In Mamdani FIS, a single value of an input parameter (fact) is used, whereas the concept of DM between the possibility distribution of a parameter based on its probability distribution (fact) and the expert's perception (fuzzy sets) is used in the proposed new formalism. The comparison between Mamdani FIS and the proposed FIS is discussed below.

a. Value of input parameter (fact)

Mamdani FIS: An input parameter is an arithmetic average value (crisp) instead of probability distribution of multiple data points. The statistical parameters (mean, variance) of a defined probability distribution has not been considered in the Mamdani approach.

ZD formalism: An input parameter could be either crisp value or a defined probability distribution of multiple data points. The probability distribution of fact (parameter) is transformed to possibility distribution termed as CNFN.

b. Fuzzification

Mamdani FIS: The crisp value of the input parameter is fuzzified and then used in FIS.

ZD formalism: The crisp value (singleton) of the input parameter is fuzzified like the Mamdani approach and then used in FIS. The expert's perception about the input parameter is represented in various linguistic terms (e.g. good and very good). A DM is the area of overlap between the fact (CNFN) and the perception of the experts for each linguistic variable of defined input parameter.

c. Implication method

Mamdani FIS: In the aggregation of rules, logical AND and OR operations are used. The fuzzy to crisp conversion (defuzzification) is carried out using the center of gravity method.

ZD formalism: However, a new approach min-max convolution is used to arrive at the linguistic description with DM. In the aggregation of rules in the antecedent part of each rule, the minimum value of DM is for each linguistic variable input.

d. Reliability

Mamdani FIS: In this approach, the defuzzification output is a number that is classified in a linguistic class without any reliability (degree of certainty) of classification.

ZD formalism: In the new approach, the reliability of the linguistic class is worked out using the concept of DM.

Therefore, the author believes that the new FIS has a definite advantage over the traditional Mamdani FIS.

4 Proposed ZD Formalism

The steps to follow using the ZD method in student's academic performance evaluation is summarized below.

4.1 Students' Academic Performance

The proposed methodology to classify students' academic performance using the ZD technique is shown in Figure 4 and is explained below.

Step 1: Every expert E_t expresses individually the fuzzy sets representing evaluator's satisfaction level in student's answer script evaluation and hierarchical FRBS, where $1 \leq t \leq n$. A multiexpert KB is constructed using appropriate aggregation function over individual experts' fuzzy sets and hierarchical FRBS.

Step 2: Every expert evaluates each question Q_i of a student S_k to form the vector of score data $V_{ik} = \langle m_{1ik}, m_{2ik}, \dots, m_{mik} \rangle$, where $1 \leq i \leq q$, $1 \leq k \leq p$, and $1 \leq t \leq n$.

Step 3: Perform j number of bootstrapping iterations over score data vector V_{ik} of each question Q_i . Normal distribution is obtained say after $j=1000$ or more bootstrap iterations.

- For every bootstrapping iteration, compute the mean of score data vector $(\bar{x})_{ik} = \frac{1}{n} * (m_{1ik} + m_{2ik} + \dots + m_{mik})$ and standard deviation S_{ik} .
- Using these j numbers of sample means $(\bar{x})_{ik}$ and standard deviations S_{ik} , calculate the bootstrap mean, bootstrap deviation, and bootstrap standard error.

$$(\text{bootstrap\_mean})_{ik} = \frac{1}{j} * (\bar{x}_{1ik} + \bar{x}_{2ik} + \dots + \bar{x}_{jik}) \quad (11)$$

Figure 4: ZD Formalism Framework.

- Compute the lower bound of CNFN

$$\text{Lower\_bound} = \text{bootstrap\_mean} - (2 * \text{bootstrap\_deviation}) \quad (12)$$

- Compute the upper bound of CNFN

$$\text{Upper\_bound} = \text{bootstrap\_mean} + (2 * \text{bootstrap\_deviation}) \quad (13)$$

- Compute the middle bound of CNFN

$$\text{Middle\_bound} = \text{bootstrap\_mean} \quad (14)$$

$\langle \text{lower bound, middle bound, upper bound} \rangle$ represents triangular CNFN for all questions Q_i of student S_k .

Step 4: For all students S_k , compute the $(DM)_{ik}$ between CNFN of each question Q_i with all aggregated fuzzy sets obtained from step 1, i.e. $\langle \text{very poor} \rangle$, $\langle \text{poor} \rangle$, $\langle \text{average} \rangle$, $\langle \text{good} \rangle$, and $\langle \text{very good} \rangle$. The overlapping area of CNFN of every question Q_i with all fuzzy sets k representing satisfaction levels is computed. The value of DM is in the range $[0, 1]$, where 0 indicates no match and 1 indicates perfectly match.

$$DM_{ik} = \frac{\text{Overlapping\_area\_of\_CNFN}(Q_i)\_with\_fuzzy\_set\_K}{\text{Total\_area\_of\_fuzzy\_set\_K}} \quad (15)$$

Every question Q_i will have a vector of DM values with all fuzzy sets representing satisfaction levels.

$$DM(Q_i) = \{DM_{\text{very\_poor}}, DM_{\text{poor}}, DM_{\text{Average}}, DM_{\text{Good}}, DM_{\text{Very\_Good}}\} \quad (16)$$

Step 5: Every rule in the hierarchical rule base system is executed in the following form.

If Question i has a $(DM)_{ik}$ with a fuzzy set (Linguistic Variable k) and Question j has a $(DM)_{jl}$ with a fuzzy set (Linguistic Variable l), then conclusion r has a $(DM)_{rc}$ with a fuzzy set (Linguistic Variable c), where $(DM)_{rc} = \min(DM_{ik}, DM_{jl})$.

After repeating this process for every rule in the current hierarchical level of FRBS, for every conclusion r , select the maximum DM value to assign it as “degree of certainty” value.

$$(DC)_{r_i} = \max(DM_{rc1}, DM_{rc2}, \dots, DM_{rcn}) \quad (17)$$

Repeat Step 5 for all hierarchical levels in FRBS. At the last level of a hierarchical rule-based system, select grade type with a maximum value of the degree of certainty, which represents the overall academic performance level of a student S_k with the “degree of certainty” representing expert's confidence in words.

5 Case study

The case study relates to the evaluation of students' subjective answer scripts using ZD formalism. The examination answer script samples in subject “Marathi” language were obtained from 237 secondary school students from three different institutions in Mumbai, India, during the academic year 2013–2014. Each student wrote a total of 10- to 12-page solution for 12 subjective questions. Answers were evaluated using a 10-point rubric from the Secondary School (SSC) Board, Maharashtra, India. Twenty subject matter experts (teachers) from different schools were identified for the answer script evaluation. All 20 evaluators belonging to the “Marathi” Department of the SSC Board evaluated all the 237 answer scripts. This experiment helps to compute consistency among all the experts in the evaluation process. The experts/evaluators were trained for assessment using rubric scale before the evaluation process starts. Each examiner expressed their judgment based on their perception (mainly their tacit knowledge) for each linguistic variable. Five satisfaction levels to award marks scored for the 12 questions are selected, viz., “very poor”, “poor”, “average”, “good”, and “very good”. Then, all the 20 teachers constructed fuzzy sets for each linguistic variable. The rules for hierarchical FRBS were formulated all 20 teachers.

In the previous research of the authors [40], the approach used was “fuzzy sets as a basis” to identify the degree of similarity among all 20 experts. The computed Fleiss κ coefficient is 0.74, which signifies an excellent agreement between all the 20 experts. When “student's marks as a basis” was used for the same purpose [39], 11 of 20 experts at a possibility of 0.8 (α -level cut) were selected and the computed κ coefficient is 0.41. Therefore, the entire 11 selected experts are given equal weightage while aggregating the MF and fuzzy rules. Score data $\langle V_{2,1}, V_{2,2}, V_{2,3}, \dots, V_{2,12} \rangle$ of Student ID-02 S_{02} for each feature attributes, i.e. 12 questions evaluated by all 11 teachers, are shown in Table 1.

A typical fuzzy set for “very poor”, “poor”, “average”, “good”, and “very good” expressed by Teacher ID-02 is shown in Figure 5. A test score between 0 and 2 reflects definitely very poor performance of students in “very poor” fuzzy set as indicated by $\mu = 1$, whereas, for a score between 2 and 2.5, μ is monotonically decreasing and it approaches 0 at test score 2.5. Based on the Delphi technique, an average aggregated fuzzy set was constructed as shown in Figure 6. A similar procedure was adopted for designing the aggregated type 1 fuzzy rules for creating four-level hierarchical FRBS.

An aggregated hierarchical multiexpert FRBS with four levels has been formulated from individual FRBS of all 11 experts. Table 2 presents a typical aggregated multiexpert fuzzy rule base, whereas the weight of each question is shown in Table 3. The degree of certainty in the evaluation of a student's answer script is expressed as evaluator's confidence as shown in Table 4.

The rule base for Category 14 node as shown in Table 2 is interpreted as follows:

1. If (Q.7 is very poor) and (Q.8 is very poor), then (Category 14 is very poor) {1}.
2. If (Q.7 is very poor) and (Q.8 is poor), then (Category 14 is very poor) {1}.
3. If (Q.7 is very poor) and (Q.8 is average), then (Category 14 is very poor) {1}.
4. If (Q.7 is very poor) and (Q.8 is good), then (Category 14 is poor) {1}.
5. If (Q.7 is very poor) and (Q.8 is very good), then (Category 14 is poor) {1}.
6. If (Q.7 is poor) and (Q.8 is very poor), then (Category 14 is poor) {1}.
7. If (Q.7 is poor) and (Q.8 is poor), then (Category 14 is poor) {1}.
8. If (Q.7 is poor) and (Q.8 is average), then (Category 14 is poor) {1}.

9. If (Q.7 is poor) and (Q.8 is good), then (Category 14 is average) {1}.
10. If (Q.7 is poor) and (Q.8 is very good), then (Category 14 is average) {1}.
11. If (Q.7 is average) and (Q.8 is very poor), then (Category 14 is poor) {1}.
12. If (Q.7 is average) and (Q.8 is poor), then (Category 14 is average) {1}.
13. If (Q.7 is average) and (Q.8 is average), then (Category 14 is average) {1}.
14. If (Q.7 is average) and (Q.8 is good), then (Category 14 is average) {1}.
15. If (Q.7 is average) and (Q.8 is very good), then (Category 14 is good) {1}.
16. If (Q.7 is good) and (Q.8 is very poor), then (Category 14 is average) {1}.
17. If (Q.7 is good) and (Q.8 is poor), then (Category 14 is average) {1}.
18. If (Q.7 is good) and (Q.8 is average), then (Category 14 is good) {1}.
19. If (Q.7 is good) and (Q.8 is good), then (Category 14 is good) {1}.
20. If (Q.7 is good) and (Q.8 is very good), then (Category 14 is good) {1}.
21. If (Q.7 is very good) and (Q.8 is very poor), then (Category 14 is average) {1}.
22. If (Q.7 is very good) and (Q.8 is poor), then (Category 14 is average) {1}.
23. If (Q.7 is very good) and (Q.8 is average), then (Category 14 is good) {1}.
24. If (Q.7 is very good) and (Q.8 is good), then (Category 14 is very good) {1}.
25. If (Q.7 is very good) and (Q.8 is very good), then (Category 14 is very good) {1}.

The numbers in curly brackets for each rule are weights given to rule sentences during aggregation.

Table 1: Score Data of 12 Answers Attempted by Student ID-02 Rated by 11 Experts.

Question ID	(S ₀₂ Marks)										
	Teacher ID										
	T-01	T-02	T-03	T-04	T-05	T-06	T-07	T-08	T-09	T-10	T-11
Q1	6.25	5	6.25	6.25	5	5	5	7.5	5	7.5	5
Q2	5	6.67	6.67	6.67	6.67	6.67	6.67	8.33	6.67	6.67	6.67
Q3	5	6.67	5	5	6.67	6.67	6.67	5	10	3.33	3.33
Q4	4	4	5	7	4	4	4	4	9	4	4
Q5	5	5	5	5	5	7.5	7.5	3.75	7.5	2.5	2.5
Q6	6.67	8.33	10	8.33	10	10	6.67	8.33	8.33	10	10
Q7	3.33	6.67	6.67	3.33	6.67	6.67	6.67	5	6.67	3.33	6.67
Q8	5	6.25	3.75	5	2.5	2.5	5	5	7.5	6.25	7.5
Q9	2.5	5	5	6.25	5	5	2.5	3.75	5	2.5	5
Q10	0	0	0	0	0	0	0	0	0	0	0
Q11	5	4	4	4.5	6	5	5	5	4.5	6	5
Q12	6.25	6.25	7.5	6.25	7.5	7.5	7.5	6.25	5	5	10

Figure 5: Typical Fuzzy Sets Represented by Teacher ID-02 Representing Expert's Satisfaction Level.

Figure 6: Resultant Aggregated Fuzzy Sets Representing Experts' Satisfaction Levels.

Table 2: Multiexpert FRBS for Category 14 Node in Four-Level FRBS as Illustrated in Figure 1.

Category 14					
Question 8	Question 7				
	Very poor (VP)	Poor (P)	Average (A)	Good (G)	Very good (VG)
Very poor (VP)	VP	VP	VP	P	P
Poor (P)	P	P	P	A	A
Average (A)	P	A	A	A	G
Good (G)	A	A	G	G	G
Very good (VG)	A	A	G	VG	VG

Table 3: Weightage of Each Question.

Question number	1	2	3	4	5	6	7	8	9	10	11	12
Marks (W_i)	4	3	3	5	4	3	3	4	4	8	10	4

Table 4: Evaluators' Evaluation Confidence Levels.

Confidence level	Degree of certainty
No	0
Low	>0–0.3
Fair	0.31–0.6
High	0.61–0.9
Very high	>0.91

6 Experimental Results and Discussion

The procedure detailed in Section 4 was followed and the results obtained are discussed below. The transformation of probability density was transformed to possibility distribution (CNFN) for all the questions answered by the students and is depicted in Figure 2. A four-level hierarchical multiexpert FRBS with the flow of DM computations in performance evaluation of Student ID-02 is presented in Figure 1. Hierarchical classification is used in this case study only for computational convenience. The final outcome will remain the same even if we change the design of hierarchical FRBS.

A typical execution of fuzzy rules for the first-level node Category 11 in hierarchical FRBS is shown in Table 5. Input parameters to node Category 11 are Questions 1 and 2 as illustrated in Figure 1. The DM for

Table 5: Fuzzy Inference at Node Category 11 for Student ID-02 S_{02} .

Category 11 rule-ID	Premise A	Degree of match (DM-A) Premise A	Premise B	Degree of match (DM-B) Premise B	Conclusion category 11 is	Degree of match for conclusion "Category 11" DC = MIN(DM-A, DM-B)
1	If Question 1 is "Very Poor"	0	AND Question 2 is "Very Poor"	0	"Very Poor"	0
2	If Question 1 is "Very Poor"	0	AND Question 2 is "Poor"	0	"Very Poor"	0
3	If Question 1 is "Very Poor"	0	AND Question 2 is "Average"	0.604	"Poor"	0
4	If Question 1 is "Very Poor"	0	AND Question 2 is "Good"	0.868	"Average"	0
5	If Question 1 is "Very Poor"	0	AND Question 2 is "Very Good"	0	"Average"	0
6	If Question 1 is "Poor"	0	AND Question 2 is "Very Poor"	0	"Very Poor"	0
7	If Question 1 is "Poor"	0	AND Question 2 is "Poor"	0	"Poor"	0
8	If Question 1 is "Poor"	0	AND Question 2 is "Average"	0.604	"Poor"	0
9	If Question 1 is "Poor"	0	AND Question 2 is "Good"	0.868	"Average"	0
10	If Question 1 is "Poor"	0	AND Question 2 is "Very Good"	0	"Average"	0
11	If Question 1 is "Average"	0.992	AND Question 2 is "Very Poor"	0	"Poor"	0
12	If Question 1 is "Average"	0.992	AND Question 2 is "Poor"	0	"Poor"	0
13	If Question 1 is "Average"	0.992	AND Question 2 is "Average"	0.604	"Average"	0.604
14	If Question 1 is "Average"	0.992	AND Question 2 is "Good"	0.868	"Average"	0.868
15	If Question 1 is "Average"	0.992	AND Question 2 is "Very Good"	0	"Good"	0
16	If Question 1 is "Good"	0.221	AND Question 2 is "Very Poor"	0	"Poor"	0
17	If Question 1 is "Good"	0.221	AND Question 2 is "Poor"	0	"Average"	0
18	If Question 1 is "Good"	0.221	AND Question 2 is "Average"	0.604	"Average"	0.221
19	If Question 1 is "Good"	0.221	AND Question 2 is "Good"	0.868	"Good"	0.221
20	If Question 1 is "Good"	0.221	AND Question 2 is "Very Good"	0	"Good"	0
21	If Question 1 is "Very Good"	0	AND Question 2 is "Very Poor"	0	"Average"	0
22	If Question 1 is "Very Good"	0	AND Question 2 is "Poor"	0	"Average"	0
23	If Question 1 is "Very Good"	0	AND Question 2 is "Average"	0.604	"Good"	0
24	If Question 1 is "Very Good"	0	AND Question 2 is "Good"	0.868	"Very Good"	0
25	If Question 1 is "Very Good"	0	AND Question 2 is "Very Good"	0	"Very Good"	0

conclusion Category 11 in each rule is minimum DM of Questions 1 and 2 with respective linguistic variables. After executing all rules in current node Category 11, the final degree of certainty is computed as shown in Table 6. For Student ID-02 S_{02} using hierarchical FRBS at level 4, select grade type with the maximum value of the degree of certainty. The academic performance of Student ID-02 S_{02} is “average” with evaluator’s degree of certainty of 0.614 having high confidence.

The results obtained for all the students using ZD formalism in a multiexpert system are summarized in Table 7. The overall performance of all 237 students can be inferred as “average”. Of 237 students, 206 (87%) are declared with expert’s high degree of certainty. The performance of 126 students (54%) in answering the paper is “average” with a high “degree of certainty”. Although the overall performance of students is “average”, it tends toward “poor”, as experts/evaluators have attached a high degree of certainty to 36 of 237 students. It is an indication to the administrator/policymaker to initiate suitable corrective action with an objective of improving student’s academic performance. The remedial measures could be improvement of the teaching/learning process, provision of additional facilities, and dialogue with the parents.

Besides, there are some practical situations where student’s evaluation is carried out by a single expert. The following section is devoted to testing the following hypothesis: “Use of multiexpert KB over individual KB enhances agreement among several independent experts in the single-expert evaluation system”.

There are two situations in single-expert evaluation system:

- Scenario 1: Multiexpert KB (i.e. fuzzy sets and FRBS) and individual teacher’s marks.
- Scenario 2: Individual-expert KB and individual teacher’s scores.

In Scenarios 1 and 2, each student’s answer script is evaluated by all 11 experts independently. The performance evaluation of students in a single-expert system in Scenarios 1 and 2 for Teacher ID-01 is summarized in Table 8.

Interrater reliability (Fleiss κ coefficient) among 11 experts in Scenarios 1 and 2 with or without expert’s confidence in students’ performance evaluation using ZD formalism is summarized in Table 9. It is observed that the interrater reliability among experts in Scenario 1 is slightly high (0.41) compared to Scenario 2 (0.37) in single-expert evaluation system. Further studies in this direction are required to confirm the hypothesis

Table 6: Degree of Certainty for Node Category 11 Using DM Values of All Rules.

Conclusion category 11 is	Degree of certainty DC = MAX(DM)
“Very Poor”	0
“Poor”	0
“Average”	0.868
“Good”	0.221
“Very Good”	0

Table 7: Summary of Performance Evaluation of Students in a Multiexpert System.

	Count of students					
Performance levels	Degree of certainty					
	No confidence (NC)	Low confidence (LC)	Fair confidence (FC)	High confidence (HC)	Very high confidence (VHC)	Total students
Very poor (VP)	0	0	2	14	3	19
Poor (P)	0	0	6	36	7	49
Average (A)	0	0	10	126	0	136
Good (G)	0	1	0	29	0	30
Very good (VG)	0	0	0	0	3	3
Total students	0	1	18	205	13	237

Table 8: Summary of Performance Evaluation of Students in a Single-Expert System in Scenarios 1 and 2 for Teacher ID-01.

		Count of students											
Performance levels		Degree of certainty											
		No confidence (NC)		Low confidence (LC)		Fair confidence (FC)		High confidence (HC)		Very high confidence (VHC)		Total students	
		Scenario 1	Scenario 2	Scenario 1	Scenario 2	Scenario 1	Scenario 2	Scenario 1	Scenario 2	Scenario 1	Scenario 2	Scenario 1	Scenario 2
Very poor (VP)	0	0	0	0	1	0	12	0	7	24	20	24	
Poor (P)	0	0	0	0	4	0	59	1	3	61	66	62	
Average (A)	0	0	0	0	20	0	120	1	3	103	143	104	
Good (G)	0	0	0	0	1	0	5	0	0	43	6	43	
Very good (VG)	0	0	0	0	1	0	1	0	0	4	2	4	
Total students	0	0	0	0	27	0	197	2	13	235	237	237	

Table 9: Summary of Interrater Reliability Among 11 Experts in Scenarios 1 and 2 with or without Expert's Confidence in students' Performance Evaluation Using ZD Formalism.

Fleiss κ statistics	κ Coefficient			
	Scenario 1		Scenario 2	
	Multiexpert fuzzy set and individual marks, considering expert's confidence in evaluation	Multiexpert fuzzy set and individual marks, without expert's confidence	Individual fuzzy set and individual marks, considering expert's confidence in evaluation	Individual fuzzy set and individual marks, without expert's confidence
P_BAR	0.429459148	0.610203299	0.286843115	0.563789797
Pe	0.215593319	0.331794589	0.143394146	0.311132622
κ coefficient	0.272646619	0.416651385	0.167462046	0.366771867

stated above. The κ coefficient obtained for 11 experts without considering experts' confidence is almost similar to the results obtained from similar expert's identification process in our previous paper [39]. It confirms that 11 experts identified for evaluation process are having similar perceptions in students' performance appraisal.

Assume that the optimism index [15] of the evaluator is 0.5 (i.e. $\lambda = 0.5$), meaning that an examiner is unbiased. In this study, all 11 examiners used the Wang and Chen method [46] and the proposed ZD method to evaluate the student's answer script. The evaluation results for Student ID-02 by 11 individual experts as well as the multiexpert are shown in Table 10. A typical evaluation sheet of Student ID-02 given by Teacher ID-04 using the Wang and Chen method is provided in Table 11. The DM computed using ZD formalism is considered as "degree of confidence" of an evaluator for satisfaction level awarded to every question. In this evaluation, the overall confidence of an evaluator is 0.66. It is important to note that, in the Wang and Chen method for students' answer script evaluation [46], the degree of certainty is not computed but instead assumed.

The novel contribution of this paper is to describe students' academic performance in linguistic terms straight away with reliability expressed in the form of "degree of certainty" attached to each performance level, which is missing in traditional evaluation approaches. Based on ZD formalism, students whose

Table 10: Comparison of the Evaluating Results of Student ID-02's Answer Script Using the Wang and Chen Method with ZD Formalism.

Teacher ID	Wang and Chen's method				Traditional method	Additional information to traditional score using ZD approach		
	Interval valued total marks [m1, m2]	Total score (55M)	Degree of confidence	Total score (55M)		Overall performance	Experts confidence	Degree of certainty
T01	19.95	22.3	21.13	1	23	Average	High confidence	0.6363636
T02	23.05	26.7	24.88	0.5	25.5	Average	High confidence	0.6363636
T03	24.15	27.15	25.65	0.5	26	Average	High confidence	0.6363636
T04	21.41	24.15	22.78	0.66	26.5	Average	High confidence	0.6363636
T05	22.95	25.3	24.13	1	27	Average	High confidence	0.6363636
T06	26.85	29.2	28.03	1	27	Average	High confidence	0.6363636
T07	22.65	25	23.83	1	26	Average	High confidence	0.6363636
T08	23.77	26.95	25.36	0.5	25.5	Average	High confidence	0.6363636
T09	30.76	33.2	31.98	0.66	30.5	Good	Fair confidence	0.5454546
T10	27.05	29.4	28.23	1	24.5	Average	High confidence	0.6363636
T11	26.45	28.8	27.63	1	27	Average	High confidence	0.6363636
Multiexpert's combined judgement	23.83	27.3117	25.57	0.61	26.226863	Average	High confidence	0.6143192

Table 11: Evaluation Sheet of Student ID-02's Answer Script by Teacher ID-04 Using the Wang and Chen Method.

Wang and Chen method (Multiexpert fuzzy set + individual marks)					
Question id	Weigtage of question	Satisfaction level	degree of confidence (DM using ZD method)	Left bound	Right bound
1	4	Average	1	1.8	2
2	3	Average	1	1.35	1.5
3	3	Average	1	1.35	1.5
4	5	Average	1	2.25	2.5
5	4	Average	1	1.8	2
6	3	Good	1	2.25	2.4
7	3	Poor	0.66666	0.7056	1.25
8	4	Average	1	1.8	2
9	4	Average	1	1.8	2
10	8	Very Poor	1	0	0
11	10	Average	1	4.5	5
12	4	Average	1	1.8	2
Total [m1 m2]				21.406	24.15
Overall confidence = minimum (Degree of confidence)					0.66
Total Score out of 55					22.777778

academic performance is “very poor” with varying degree of certainty can be classified in one group (in this case, 19 students), so that the management can initiate necessary steps to improve the performance of these students. In the ZD approach, there is no need to assume a degree of certainty, which can be termed as a “level of confidence” of the evaluator, and this is the strength of the method suggested in [46]. The aim of the proposed method is not to replace the traditional method of student's performance evaluation; instead, it will strengthen/enrich the present system by providing reliability of information to be used in decision making by the policymakers.

In summary, the ZD method can evaluate student's answer scripts in a more flexible and more intelligent manner considering uncertainty in an evaluation process. Besides, it can be stated that, to verify the hypothesis: “Use of multiexpert KB over individual KB enhances agreement among several independent experts in the single-expert evaluation system”, a few more studies are needed. The software implementation of ZD formalism in student's subjective answer script evaluation is done in MATLAB R2008a and Microsoft Access 2010.

7 Concluding Remarks

Traditional practice in the estimation of quality (such as air, water, and student performance) is either by indices or at best with aggregated percentages of ratings. Test score is used in traditional evaluation system to classify the students' academic performance (very poor, poor, etc.) without attaching any reliability (i.e. degree of certainty) to each linguistic term representing different performance levels. The proposed FIS (ZD formalism) could be more reliable for all the stakeholders, as it computes the degree of certainty for each linguistic term to classify students' academic performance. Although efforts are made to study the robustness of the ZD method in air/water quality classification, few more case studies are needed to establish the validity of the concept of degree of certainty in students' academic performance evaluation. In this quest, the authors have initiated the application of evidence theory and fuzzy relation calculus in students' performance evaluation.

Acknowledgments: We are grateful to the learned reviewers for their valuable comments.

Bibliography

- [1] S.-M. Bai and S.-M. Chen, Evaluating students' learning achievement using fuzzy membership functions and fuzzy rules, *Expert Syst. Appl.* **34** (2008), 399–410.
- [2] J. Biggs, *Teaching for Quality Learning at University*, SRHE & Open University Press, Buckingham, 2001.
- [3] R. Biswas, An application of fuzzy sets in students' evaluation, *Fuzzy Sets Syst.* **74** (1995), 187–194.
- [4] J. Casanovas and J. Vicente Riera, Weighted means of subjective evaluations, in: *Soft Computing in Humanity and Social Science*, *STUDFUZZ* 273, pp. 323–345, 2012.
- [5] M. Celik and A. Kandakoglu, Structuring fuzzy integrated multi-stages evaluation model on academic personnel recruitment in MET instruction, *Expert Syst. Appl.* **36** (2009), 6918–6927.
- [6] K. C. Chai, K. M. Tay and C. P. Lim, A new fuzzy peer assessment methodology for cooperative learning of students, *Appl. Soft Comput.* **32** (2015), 468–480.
- [7] D. F. Chang and C. M. Sun, Fuzzy assessment of learning performance of junior high school students, in: *First National Symposium on Fuzzy Theory and Applications*, June 25–26, 1993, Hsinchu, Taiwan, 1993.
- [8] S. M. Chen, A new approach to handling fuzzy decision-making problems. *IEEE Trans. Syst. Man Cybernet.* **18** (1998), 1012–1016.
- [9] S. M. Chen and C. H. Lee, New methods for students' evaluating using fuzzy sets, *Fuzzy Sets Syst.* **104** (1999), 209–218.
- [10] S.-M. Chen and T.-K. Li, Evaluating students' learning achievement based on fuzzy rules with fuzzy reasoning capability, *Expert Syst. Appl.* **38** (2011), 4368–4381.
- [11] S.-M. Chen and T.-K. Li, Evaluating students' learning achievement based on the eigenvector method, *Expert Syst. Appl.* **38** (2011), 8240–8250.
- [12] S.-M. Chen and T.-K. Li, Evaluating students' learning achievement by automatically generating the importance degrees of attributes of questions, *Expert Syst. Appl.* **38** (2011), 8240–8250.
- [13] S.-M. Chen and T.-S. Li, Evaluating students' answer scripts based on interval-valued intuitionistic fuzzy sets, *Inf. Sci.* **235** (2013), 308–322.
- [14] S.-M. Chen and H.-Y. Wang, Evaluating students' answer scripts based on interval-valued fuzzy grade sheets, *Expert Syst. Appl.* **36** (2009), 9839–9846.
- [15] C. H. Cheng and K. L. Yang, Using fuzzy sets in education grading system, *J. Chin. Fuzzy Syst. Assoc.* **4** (1998), 81–89.
- [16] A. Deshpande and J. Yadav, Zadeh-Deshpande (ZD) fuzzy logic-based formalism for the linguistic description of air quality: a case study, in: *3rd International Conference on Informatics, Electronics & Vision*, Dhaka, 2014.
- [17] J. L. Fleiss, Statistical inference for a single proportion, Chapter 2, in: *Statistical Methods for Rates and Proportions*, 2nd ed., pp. 38–46, John Wiley, New York, 1981.
- [18] J. L. Fleiss and J. Cohen, The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability, *Educ. Psychol. Meas.* **33** (1973), 613–619.
- [19] Y.-Y. Guh and M.-S. Yang, Establishing performance evaluation structures by fuzzy relation-based cluster analysis, *Comput. Math. Appl.* **56** (2008), 572–582.
- [20] I. A. Hameed, Using Gaussian membership functions for improving the reliability and robustness of students' evaluation systems, *Expert Syst. Appl.* **38** (2011), 7135–7142.
- [21] C.-S. Hsieh and Y.-W. Chen, Characteristics of fuzzy synthetic decision methods for measuring student achievement, *Qual. Quant.* **46** (2012), 523–543.
- [22] https://en.wikipedia.org/wiki/Grading_systems_by_country#India, Accessed November, 2016.
- [23] T. L. Jee and K. M. Tay, Enhancing a fuzzy inference system-based criterion-referenced assessment with an analogical reasoning schema, *Procedia-Soc. Behav. Sci.* **56** (2012), 12–21.
- [24] T. L. Jee, K. M. Tay and C. K. Ng, A new fuzzy criterion-referenced assessment with a fuzzy rule selection technique and a monotonicity-preserving similarity reasoning scheme, *J. Intell. Fuzzy Syst.* **24** (2013), 261–279.
- [25] Z. C. Johanyak, Survey on five fuzzy inference-based student evaluation methods, in: *Computational Intelligence in Engineering, SCI 313*, pp. 219–228, Springer-Verlag, Berlin/Heidelberg, 2010.
- [26] D. Kseibat, Student model based on flexible fuzzy inference, in: *Innovations in Computing Sciences and Software Engineering*, pp. 39–43, 2010.
- [27] S. H. Lau, K. M. Tay and C. K. Ng, An application of fuzzy adaptive resonance theory to engineering education, in: *International Conference on Neural Information Processing*, Springer International Publishing, 2014.
- [28] C. K. Law, Using fuzzy numbers in education grading system, *Fuzzy Sets Syst.* **83** (1996), 313–323.
- [29] T.-K. Li and S.-M. Chen, A new method for students' learning achievement evaluation by automatically generating the weights of attributes with fuzzy reasoning capability, in: *Proceedings of the 8th International Conference on Machine Learning and Cybernetics, Baoding*, July 12–15, 2009.
- [30] J. Ma and D. Zhou, Fuzzy set approach to the assessment of student-centered learning, *IEEE Trans. Educ.* **43** (2000), 237–241.
- [31] J. A. Manz, M. Hercinger, M. Todd, K. S. Hawkins and M. E. Parsons, Improving consistency of assessment of student performance during simulated experiences, *Clin. Simul. Nurs.* **9** (2012), e229–e233.

- [32] J. M. Mendel, J. Lawry and L. A. Zadeh, Foreword to the special section on Computing with Words, *IEEE Trans. Fuzzy Syst.* **18** (2010), 437–440.
- [33] J. A. Montero and F. Alias, Assessing student's teamwork performance by means of fuzzy logic, in: *IWANN 2007, LNCS-4507*, pp. 383–390, 2007.
- [34] E. N. Nasibov and A. Ovgu Kinay, An iterative approach for estimation of student's performance based on linguistic evaluations, *Inf. Sci.* **179** (2009), 688–698.
- [35] N. Ramli and D. Mohamad, A centroid-based performance evaluation using aggregated fuzzy numbers, *Appl. Math. Sci.* **3** (2009), 2369–2381.
- [36] K. A. Rasmani and Q. Shen, A data-driven fuzzy rule induction and its application to student performance evaluation, *Appl. Intell.* **25** (2006), 305–319.
- [37] I. Saleh and S.-I. Kim, A fuzzy system for evaluating students' learning achievement, *Expert Syst. Appl.* **36** (2009), 6236–6243.
- [38] S. Saliu, Constrained subjective assessment of student learning, *J. Sci. Educ. Technol.* **14** (2005), 271–284.
- [39] S. Salunkhe and A. Deshpande, Can fuzzy relational calculus bring complex issues in selection of examiners into focus?, *J. Intell. Syst.* **25** (2016), 85–97.
- [40] S. Salunkhe, A. Deshpande and Y. Joshi, Fuzzy similarity measures as a basis for the selection of examiners, *J. Fuzzy Set Valued Anal.* **2015** (2015), 245–258.
- [41] A. M. G. Solo and M. Gupta, Uncertainty in computational perception and cognition, in: *Forging New Frontiers: Fuzzy Pioneers I.*, Springer-Verlag Berlin Heidelberg, 2007.
- [42] K. M. Tay and C. P. Lim, A fuzzy inference system-based criterion-referenced assessment model, *Expert Syst. Appl.* **38** (2011), 11129–11136.
- [43] R. van Ewijk, Same work, lower grade? Student ethnicity and teachers' subjective assessments, *Econ. Educ. Rev.* **30** (2011), 1045–1058.
- [44] J. Vicente Riera and J. Torrens, Aggregation of subjective evaluations based on discrete fuzzy numbers, *Fuzzy Sets Syst.* **191** (2012), 21–40.
- [45] H.-Y. Wang and S. M. Chen, Artificial intelligence approach to evaluate student's answer scripts based on the similarity measure between vague sets, *Educ. Technol. Soc.* **10** (2007), 224–241.
- [46] H.-Y. Wang and S.-M. Chen, Evaluating students' answer scripts using fuzzy numbers associated with degree of confidence, *IEEE Trans. Fuzzy Syst.* **16** (2008), 403–415.
- [47] L. A. Zadeh, Fuzzy sets, *Inf. Control* **8** (1965), 338–353.
- [48] L. A. Zadeh, Outline of a new approach to the analysis of complex systems and decision processes, *IEEE Trans. Syst. Man Cybernet.* **SMC-3** (1973), 28–44.
- [49] L. Zadeh, Toward a perception-based theory of probabilistic reasoning with imprecise probabilities, *J. Stat. Plann. Inference* **105** (2002), 233–264.
- [50] L. A. Zadeh, The concept of a Z-number – a new direction in uncertain computation, in: *IEEE International Conference on Information Reuse & Integration*, 2011.