Asia-Pacific Edu Res (2014) 23(3):717–726
DOI 10.1007/s40299-013-0145-2

REGULAR ARTICLE

Profiling Instructional Effectiveness to Reveal Its Relationship to Learning

Ilker Kalender

Published online: 23 November 2013

Abstract The purpose of the present study is to define instructional profiles and investigate the relationship between these profiles and learning indicators such as end-of-semester grades and self-reported amount of learning. Instructional profiles were obtained using a segmentation method. Student ratings were used as indicators of instructional effectiveness. Results revealed that instructors who receive higher scores from students seem to be effective instructors in learning. However, instructors with high ratings from students did not receive high scores for all measures of instructional effectiveness. Effective instructors seem to have varying scores due to the imperfect relationship between instructional effectiveness and learning. It can be concluded that the definition of an effective instructor can vary across subgroups. For an instructor to be defined as effective, it is not necessary for them to receive higher scores for all measures. Low-rated aspects of effectiveness can be compensated for by showing high performance in other areas. Based on the results of the present study, instructional profiles or any other related traits should be investigated under subgroups that show differences.

Keywords Instructional effectiveness · Student learning · Student ratings · Clustering · CHAID analysis

Introduction

Instructional effectiveness does not have a precise definition that is applicable for all circumstances; it has several components that may be included in the definition (Abrami and d'Apollonia 1991; Cashin and Downey 1992; Feldman 1997; Marsh and Roche 1993). Among reported components are sensitivity, clarity, enthusiasm of instructor, and quality of assignments (Feldman 1988; Marsh and Bailey 1993). The weighting of each component in a potential definition may vary for different contexts (McKeachie 1997). For example, Young and Shaw (1999) suggest that an instructor who receives low scores for motivation may still be determined to be effective by compensating with high scores for communication.

There are different means for assessing instructional effectiveness, such as colleague or expert ratings, self-ratings (Feldman 1989), and the predominant mechanism of using student evaluation forms for assessment.

Studies by Ellett et al. (1997) statistically validated student ratings. Additional studies included a comparison of student ratings to different measures of instructional effectiveness have revealed correlation coefficients that favor student ratings (Abrami et al. 1990; Teven and McCroskey 1997). Several researchers state evidence for the reliability of student ratings (Cashin et al. 1994; Marsh 1984). Some researchers have found supporting evidence for unbiased judgments of students for instructional effectiveness (Benz and Blatt 1995; Johannessen 1997; Rodabaugh and Kravitz 1994).

Student ratings are affected by several controllable and uncontrollable factors (class size, contact hour, grade level of courses, etc.) associated with course and instructor characteristics. Among them, the most controversial relationship that student ratings have is probably with grading. Studies provided inconsistent results in regards to the grading of instructors. For example, Gigliotti and Buchtel (1990) reported that there is no correlation between grades received by students and their ratings of the instructors,

I. Kalender (✉)
Faculty of Education, Bilkent University, Ankara, Turkey
e-mail: kalenderi@bilkent.edu.tr

whereas Greenwald and Gillmore (1997a) and Wilson (1998) found positive relationship between them. However, the relationship or mutual interaction between grades and ratings is somewhat complex due to grading leniency (Greenwald and Gillmore 1997b). Supporting findings on the complicated relationship between grading policy and student ratings were provided by Rodabaugh and Kravitz (1994) and Sailor et al. (1997), who also found significant correlations between student ratings and the grades that the student received. A weak relationship between class size and student ratings was reported (Aleamoni and Hexner 1980), whereas other studies (Lin 1992; VanArsdale and Hammons 1995) did not find any relationship between the two variables. Rayder (1968) and Feldman (1983) showed that there is a negative relationship between instructional experience and student ratings. The number of sections of courses is related to student ratings as shown by Kalender (2011), who stated that instructors who teach courses with only one section tend to receive lower ratings from students compared with those instructors teaching courses with more than one section. Other factors reported to be correlated with student ratings include grade level of the course (Braskamp and Ory 1994; Donaldson et al. 1993), gender of instructor (Atamian and Ganguli 1993; Tatro 1995), contact hour (Dawson 1986), and course credit (Kockelman 2001), etc.

Although student ratings were shown to be influenced by many factors, they are still widely used for assessment of instructional effectiveness. Cashin (1995) found a positive correlation between student ratings and learning measures, supporting the use of student ratings as an indicator of in instruction effectiveness.

Previous studies have attempted to show correlations between several variables and student ratings, but identification of factors related to student ratings does not provide any insight into subgroups of the student body that may differ in several aspects such as learning style and achievement level (Trivedi et al. 2011). Marsh and Hocevar (1991) have shown that the various characteristics of student bodies are not homogenous. They identified 21 subgroups based on such academic characteristics such as instructor rank and course level. The present study aims to identify the relationship between the amount of learning and instructional profiles in different student subgroups. By investigating student subgroups, several definitions of instructional effectiveness are expected to be obtained. Instructional profiles providing maximization in differences with respect to student learning are sought to be defined under student subgroups, instead of generalizing results to whole population.

The main approaches for identifying subgroups and revealing hidden relationships among them are clustering or segmentation (Aldenderfer and Blashfield 1984). Rather than obtaining relational variables on a population, use of segmentation procedures provides information about the existence and relationships of variables under homogenous subgroups. Borden (1995) investigated two hierarchical clustering methods and showed promising results regarding the use of clustering methods. Thomas and Galambos (2004) used a segmentation technique to investigate the relationship among several student characteristics, experiences, perceptions, and general satisfaction. Regarding instructional profiles, Marsh and Bailey (1993) showed that there are distinct instructional profiles among instructors as identified by systematic differences in terms of instructional measures. A similar study conducted by Young and Shaw (1999) revealed different instructor profiles that showed variation among measures. The researchers concluded that instructors who are defined as "*effective*" may not be receiving high scores for all measures. Some of the scores they receive, for example, for organization or communication may be low; however, the instructors may still be called "*effective*". It can be surmised from these studies that investigation of relationships between instructional profiles and external criteria such as cognitive or affective variables may yield significant information in terms of student learning. Identification of instructional profiles and their relationships to students' learning may also provide additional information as to instructional practices that need to be improved and that could not be identified from correlational studies on entire student bodies.

Method

The present study sought to find significant indicators of learning by dividing the student body into subgroups using selected variables related to course and instructor. For segmentation, two leading indicators of amount of learning were used: (a) end-of-semester grades that students received for their courses and (b) students' self-reported amount of learning. After the student subgroups were defined, differences among them were investigated from the perspective of instructional effectiveness measures.

Data

Data for the present study were drawn from responses to evaluation forms completed anonymously by students to rate instructors in a university setting. Rating forms were distributed at the end of the semester, before final examinations were given. A group of 20,694 students who were registered for 628 different courses in four-year undergraduate programs from social sciences, natural sciences,

and engineering were randomly selected. Distribution of the courses with respect to grade levels was as follows: freshman (34.1 %), sophomore (24.5 %), junior (20.1 %), and senior (21.3 %). Gender information of student raters was not available; which is consistent with many findings in the literature that did not identify the gender of students as an influencing factor on student ratings (Fernandez and Mateo 1997; Freeman 1994; Ludwig and Meacham 1997). Student grades had a distribution with a mean of 2.39 (out of 4.00) and a standard deviation of 0.59. Mean and standard deviation of class size were 30.99 and 19.37, respectively. Course credit changed between two and five with a mean of 3.18 and a standard deviation of 0.66, whereas minimum and maximum contact hours of the courses were 1–6 h per week, respectively ($M = 2.95$, $SD = 0.67$). Fifty-one percent of the courses had one section, 19 % had two sections, 11 % had three sections, and the rest had four or more sections.

The experience of the instructors was defined by the number of years they had taught after receiving their Ph.D. Courses were selected, if the instructors teaching those courses had started teaching immediately after receiving their Ph.D.s at the university where the data were collected. Because instructors of the courses from which the data were collected had started teaching at the university after they had received their Ph.D.s, they had gained their instructional experience at the university in the study.

A two-stage selection and placement procedure is used for admittance to higher education programs in Turkey. The procedure involves multiple-choice testing of students in higher order cognitive skills in science, mathematics, social sciences, and the Turkish language. Therefore, sample was considered to be qualified to have the skills that are expected to be gained before university level.

Instructional Effectiveness Measures

In the student rating forms, there were several five-point Likert-type items (1: strongly disagree to 5: strongly agree) to measure different aspects related to course and instructor. Among them, seven items related to instructional effectiveness were selected for the present study (Cronbach's Alpha is 0.97). Instructional effectiveness items and abbreviations with their means and standard deviations are as follows: (a) *Promoting student participation of students* (*active*) ($M = 3.37$, $SD = 0.63$), (b) *Evaluation of assessment material* (*exams*) ($M = 3.35$, $SD = 0.57$), (c) *Stimulating interest in the subject* (*interest*) ($M = 3.38$, $SD = 0.62$), (d) *Amount of learning* (*learned*) ($M = 3.37$, $SD = 0.60$), (e) *Overall rating for instructor* (*overall*) ($M = 3.45$, $SD = 0.56$), (f) *Behavior of instructor toward student* (*respect*) ($M = 3.75$, $SD = 0.39$), and (g) *Developing critical thinking skills* (*think*) ($M = 3.40$, $SD = 0.59$). Prior to analyses, student ratings were transformed to have a unit standard distribution ($M = 0$, $SD = 1$) to ensure that all items were on the same scale.

Confirmatory analysis with Lisrel (Jöreskog and Sörbom 1999) was employed to investigate unidimensionality of the seven items. Root mean square errors of approximation (RMSEA), comparative fit index (CFI), non-normed fit index (NNFI), and standardized root mean square residual (SRMR) indices, which had values of 0.04, 0.98, 0.97, and 0.015, respectively, were checked. All indices had acceptable values (Kelloway 1998), indicating unidimensionality of the trait. Goodness-of-fit indices produced by LISREL were also checked for model fit, and values of the indices were found as follows: goodness-of-fit index (GFI) = 0.92 and adjusted goodness-of-fit index (AGFI) = 0.85. Although values greater than 0.90 indicate a good fit, as Kline (2005) stated a threshold of 0.85 is acceptable. As a result, analyses provided supporting evidence for grouping of the items under a latent trait which is named "instructional effectiveness" in the present study.

Procedure

For segmentation, a Chi squared automatic interaction detector (CHAID), one of the decision tree analysis methods (Sonquist and Morgan 1964), was employed. As an exploratory method, CHAID is used for identification of determinants of subgroups or segments. CHAID uses a dependent or target variable on which classification is made and clusters data by determining independent or predictor variables that differentiate the target variable. Although CHAID is similar to regression in identification of factors related to a target variable, a unique advantage of CHAID offers is the opportunity of determination of significant factors maximizing differences between subgroups. The predictor variable explaining the largest portion of variable on target variable defines the first level in a classification tree and different values of that variable constitutes the first-level clusters. Another predictor variable providing the second largest contribution to explain the differences on target variable is used for the second-level classification by dividing the clusters formed in the first classification into subsequent clusters. There are several options that can be set in CHAID analysis, for example, depth of tree can be defined prior to the CHAID analysis. Likewise, a minimum number of cases for nodes can be changed to obtain different tree structures. The values should be accordingly set especially when the sample size is not large enough for clusters to have an adequate number of cases. The CHAID procedure uses χ^2 tests to determine the significant differences between

clusters with respect to mean of target variable for whole body. A study comparing two clustering methods by Borden (1995) suggested CHAID analysis as a useful way to identify patterns on data.

In the present study, segmentation was made based on two indicators of students' learning: end-of-semester grades of students and self-reported amount of learning. Correlation between the two indicators is moderate: r (626) = 0.43, $p < 0.05$. The degree of the relationship between the two led to investigation of differentiating patterns with respect to instructional profiles for both indicators. Therefore, it was decided to conduct two separate CHAID analyses with two measures of learning as target variables. Predictor variables that have been shown to have relationships with student ratings include class size, credit, grade level, contact hour, number of section of the courses, and instruction experience of instructors. The variables used in the present study were limited to availability.

To run the CHAID procedure, a classification tree module of SPSS 13.0 was used (Norusis 2005). Both the CHAID analyses included 628 different courses from which 20,694 students were selected. The number of maximum levels was set to two for minimizing the number of clusters and keeping the number of courses in the clusters higher.

After obtaining student subgroups, differences between means of clusters and the whole group in target variable were checked using one-sample t-test to remove the clusters which were not different from the whole group. In this way, only clusters significantly different from the mean were kept. An additional analysis was conducted on the remaining clusters. The remaining clusters were further analyzed using one-sample t test to again demonstrate a difference in the means such that only clusters representing significantly different profiles of effectiveness measures were determined. After those analyses, instructional profiles and the differences among them were investigated.

Results

The decision tree produced by the CHAID procedure using end-of-semester grades as the target variable is presented in Fig. 1. Of the independent variables entered into the CHAID analysis, grade level was found to be the predictor factor most associated with the target variable. Courses at grade levels 1 and 4 were grouped separately, whereas those at grade levels 2 and 3 were included in one cluster. Courses at grade level 1 were split into two subgroups (Clusters 1 and 2) with respect to course credit, which is the next predictor variable for those clusters. Courses at grade levels 2 and 3 were divided into two subgroups with respect to contact hour, forming Cluster 3 and 4. For Cluster 5, 6, and 7, the determinant variable was class size after grade level of courses.

To find the clusters that were significantly different than the whole body in terms of end-of-semester grades, further analysis conducted on clusters showed that Cluster 7, $t(20) = 0.14$, $p < 0.05$, and Cluster 4, $t(243) = -1.37$, $p < 0.05$, were not statistically significantly different than the mean of the whole student body ($M = 2.39$). Based on that finding, those clusters were excluded from the rest of the analyses. To investigate the differences in terms of instructional effectiveness measures, an additional analysis was conducted with the remaining five clusters. One-sample t-tests were conducted to determine if effectiveness measures in the clusters were significantly different than the mean of the whole body, which was equal to 0. The results revealed that none of the items were different than 0 for Cluster 1, indicating that there was no distinct profile for that cluster with respect to instructional effectiveness measures. For this reason, the cluster was also removed.

Of the remaining clusters, Cluster 6, courses at grade level 4 with between 24 and 47 students had the highest mean for end-of-semester grade ($M = 2.88$). Similarly, Cluster 3, courses having two credits or less at grade levels 2 and 3 and Cluster 5, courses at grade level 4 including less than 24 students, had higher grades: 2.82 and 2.77, respectively. Alternately, Cluster 2 included courses with a lower grade mean ($M = 2.00$). At the end of the analyses, three clusters (3, 5, and 6), courses with higher end-of-semester grades, and one cluster (2), courses with means below the whole body, were left.

After obtaining four statistically significant clusters, investigation of instructional effectiveness revealed that Clusters 3, 5, and 6 included instructors received scores above the mean for effectiveness measures, and all items were rated below the mean for Cluster 2. Means of instructional effectiveness measures for each cluster are given in Figs. 2, 3, 4, and 5.

Profile 1 ($n = 36$) can be characterized by higher ratings with the exception of low ratings given for the measure *Respect*. The mean score of measures in that profile was 0.57. Fifty-eight instructors in Profile 2 have a mean score of 0.32. The measures *Exams* and *Respect* have relatively higher scores. Profile 3 includes 55 courses that were rated with lower scores with a mean of 0.26. Measures *Active* and *Interest* were rated relatively higher than other items. Profile 4 ($n = 92$) had courses with the lowest rating for all measures (mean score is −0.34). Measures *Active* and *Learned* received the lowest ratings. It was also observed that three profiles (1, 2, and 3) with higher end-of-semester grades had higher instructional effectiveness scores. Profile 3 had the highest end-of-semester grade, and instructional effectiveness scores were the lowest (but positive) compared with those scores of Profiles 1 and 2.

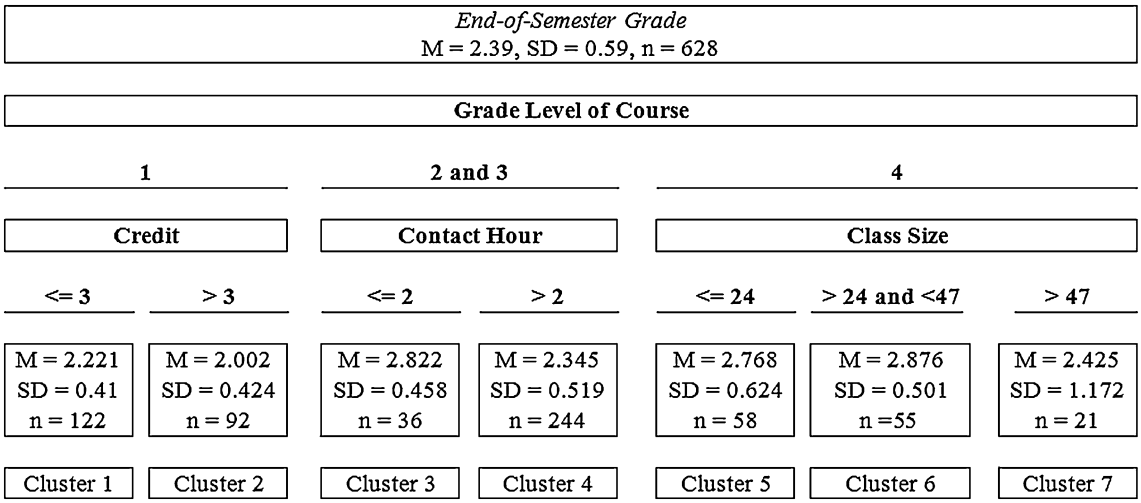

Fig. 1 Tree structure explaining predictors of "end-of-semester grade"

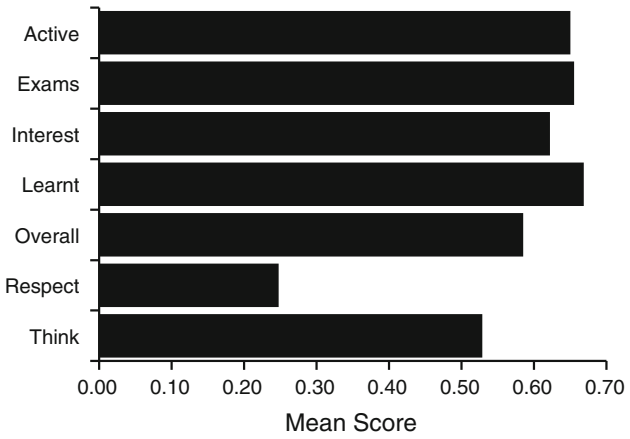

Fig. 2 Profile 1 (cluster 3) for end-of-semester grades

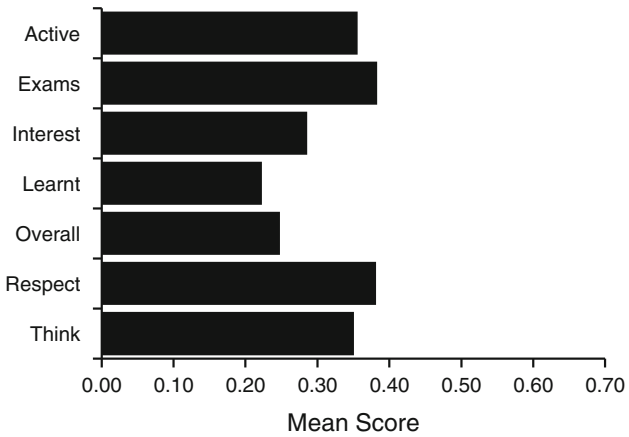

Fig. 3 Profile 2 (cluster 5) for end-of-semester grades

For Profile 1, all instructional measures were five or more times greater than the mean, except the *Respect* measure, which had no significant difference in the mean ($p < 0.05$). The measure *Learned* in Profile 2 and the measures *Think* and *Exams* in Profile 3 were not significantly different from the whole body ($p < 0.05$). For Profile 4, all measures were significant.

A second CHAID analysis was conducted using the other learning indicator: *amount of learning. The d*ecision tree produced is shown in Fig. 6. Similar to the first tree, grade level is the primary determinant of the target variable. Courses at the grade level 1 were split into two subgroups with respect to course credit, forming Clusters 1 and 2. Similarly, Clusters 3 and 4 included both courses at grade level 2 and 3 and were separated by contact hours. Clusters 5 and 6 were split based on course credit for grade level 4 courses. The first four clusters were the same as those in the first tree in their predictors and number of courses.

Investigation of mean differences among clusters with respect to target variable revealed that Cluster 6, $t(119) = 1.82$, $p < 0.05$, Cluster 1, $t(121) = -0.12$, $p < 0.05$, and Cluster 4, $t(243) = -0.64$, $p < 0.05$, were not significantly different from the mean of the whole body, and these three clusters were removed from the study. Additional analysis revealed that there are no clusters that have mean values of instructional effectiveness that are significantly different. Three clusters (1, 4, and 6) were removed as a result of the second CHAID analysis.

For the three remaining clusters: Cluster 2 (grade level 1 courses with more than three credits) had a mean below 0 ($M = -0.45$) for self-reported learning; Cluster 3 (courses at grade level 2 and 3 with less than three contact hours) had the highest mean of amount of learning reported by students ($M = 0.67$); Cluster 5 (grade level 4 courses having two or less credits) also had a high a mean of 0.66.

At the end of the analyses, three separate clusters, corresponding to different learning levels and distinct profiles

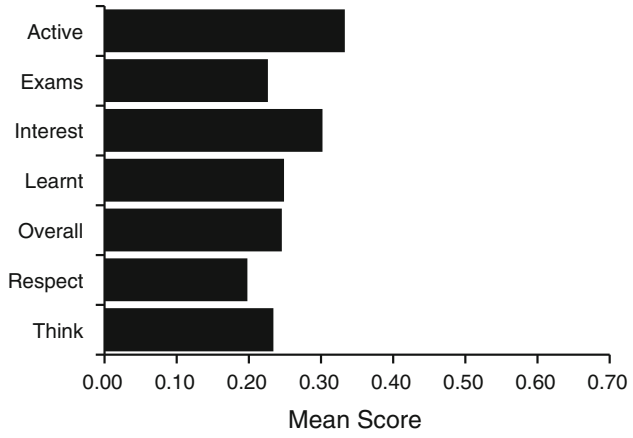

Fig. 4 Profile 3 (cluster 6) for end-of-semester grades

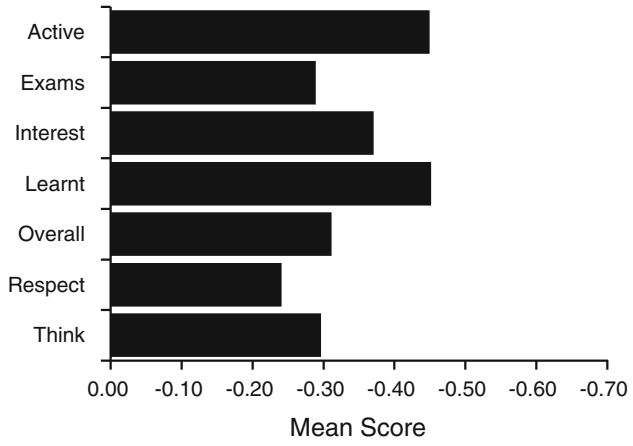

Fig. 5 Profile 4 (Cluster 2) for end-of-semester grades (with *x* axis reversed)

in terms of instructional effectiveness measures, were obtained. Investigation of ratings scores given by students for instructors revealed that Clusters 3 and 5 had positive scores for all items, whereas all items rated below the mean for the Cluster 2. Figures 7, 8, and 9 show the means of effectiveness measures for each cluster.

Profile 1, with 36 instructors, had relatively higher ratings except for the measure *Respect*. The mean score was 0.52. Profile 2 ($n = 14$) included instructors who received the highest ratings for the measure *Overall*. The mean score for that profile was 0.55, a similar value to that of Profile 1. Ninety-two instructors in Profile 2 had a mean score of −0.33, the lowest ratings for all measures. The measure *Active* had the lowest ratings in the profile.

Courses in Profiles 1 and 2 had similar pattern of scores across effectiveness measures. The measure *Respect* was lower for both of them, with no significant difference for Profile 1 ($p < 0.05$). Similarly the measure *Think* was not different from the whole body for Profile 2 ($p < 0.05$) that included the scores of the instructors of courses with two and less credits at the grade level 4. The lowest scores were given for Profile 3 across all measures. As expected, subgroups obtained using the dependent variable "*I learned a lot in this course*" provided a better discrimination among effectiveness measures compared with the dependent variable end-of-semester grade because it was rated by students along with effectiveness measure (Borden 1995).

Discussion

Instructional effectiveness is probably one of the most controversial topics in educational literature. Although a large body of literature on the topic that includes definition, dimensions, and assessment of instructional effectiveness, the results reported are inconclusive, especially in studies on the factors related to student ratings that focus on subgroups.

In the present study, the focus was on clusters of students rather than whole student body. By using a segmentation method, CHAID, relatively homogenous clusters were obtained and differences among them were investigated with respect to instructional effectiveness measures. Findings of the present study may provide additional insight into the issue of instructional profiles under subgroups.

Two indicators of student learning were included to form student clusters: (i) end-of-semester grades and (ii) self-reported amount of learning. After student subgroups were defined using CHAID analyses, additional analyses on clusters showed that nearly half of them were found not to be different from the homogenous whole body, which may be an expected outcome because many instructors behave in parallel to the majority and receive similar scores.

The grade level of courses was found to be the principle predictor variable on both target variables. Credit, contact hour, and class size of courses were identified as other predictor variables as course-related factors. Although it is not within the scope of the present study to discuss the relationship between these variables and learning, results indicated that course-related factors provided a good differentiation among learning segments. The relationship between these variables and learning deserves a separate study.

One of the results of the present study is that students' learning levels, as defined by end-of-semester grades, are the lowest for the grade level 1. A similar finding for Turkish students was also reported by Kalender (2011), who conducted a discriminant analysis to find out the factors differentiating between high- and low-rated instructors. The results of that study revealed that grade level of the course is the most discriminating factor for

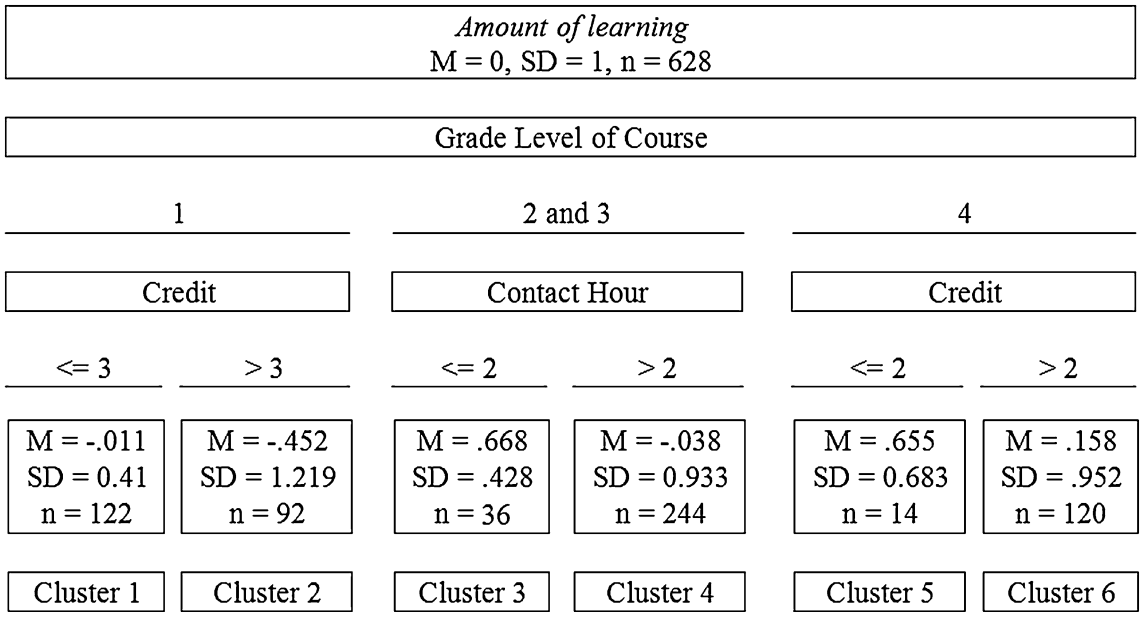

Fig. 6 Tree structure explaining predictors of "amount of learning"

ratings of instructors. This may be due to lower maturity level of students at lower grade levels who may not be competent to make a sound assessment of instructors. As they advance in grade level, students may become familiar with instructional experiences and make more qualified evaluations on effectiveness. A similar result was revealed from the second CHAID analysis conducted on the amount of self-reported learning as a target variable. Students in higher grade levels report a higher amount of learning compared with students at the grade level 1.

For the first decision tree, four clusters were produced, and three of them had higher means of grades than the whole body. A common result for the clusters above the mean was that instructors whose students receive higher end-of-semester grades obtained higher ratings by students. This finding supports the conclusion of Cashin (1995), who stated that students learn better in classes in which effective instructors teach. Another potential explanation for the relationship between grades and students' ratings reported in the literature is grading leniency, which can be described as receiving higher ratings from students in return for students receiving higher grades than they deserve. Greenwald and Gillmore (1997a, b) consider it, as a potential explanation for the relationship between learning measures and achievement indicators. However, as Marsh and Roche (1997) stated, grading leniency does not produce an effective contamination component for this relationship. Similar findings were obtained for clusters formed using amount of learning as the dependent variable. Two significant clusters above the mean included instructors who received higher ratings, whereas instructors of courses in the cluster had a lower mean given lower scores.

The first four clusters produced by CHAID are exactly the same in terms of number of courses for both dependent variables. The difference was observed in the clusters defined by grade level 4. The moderate correlation (0.43) between end-of-semester grades and amount of learning may provide an explanation for this finding. Because the dependent variables have a relationship, it is expected that the CHAID procedure determined similar or identical predictors when clusters were formed. For higher correlations, a higher degree of similarity between the two trees would be expected.

The predictor class size existed in the first tree, and it was not included in the second tree. Class size is related to student ratings as shown in eight studies cited by Aleamoni and Hexner (1980). Average correlation between class size and student ratings was weak. For example, Sixbury and Cashin (1995) found the average correlation between them as −0.14 between several instructional effectiveness measures and ratings. Furthermore, in the decision tree, the absence of class size does not necessarily mean it is not related to student ratings. Because the present study used tree with two-levels, class size might not be selected as a predictor variable by CHAID procedure. If decision trees with more levels were investigated, class size may be observed as a significant factor.

Another result that should be noted is that there was no perfect relationship between grades and student ratings among profiles. Profiles 1 and 3 obtained in the first CHAID analysis had the highest and closest values in terms of end-of-semester grades; however, means of those profiles in terms of instructional effectiveness measures were 0.57 and 0.26, respectively. The reason for lack of a perfect

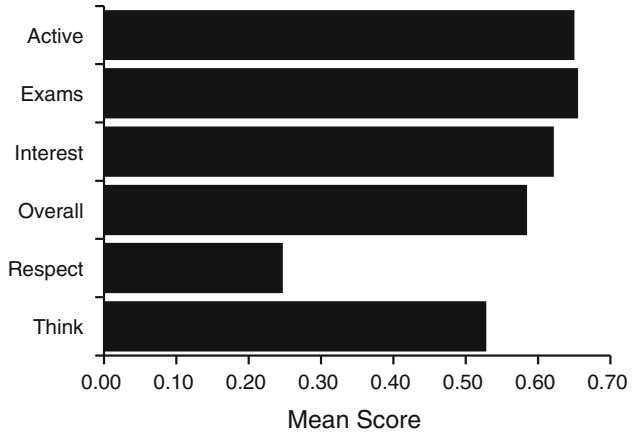

Fig. 7 Profile 1 (Cluster 3) for amount of learning

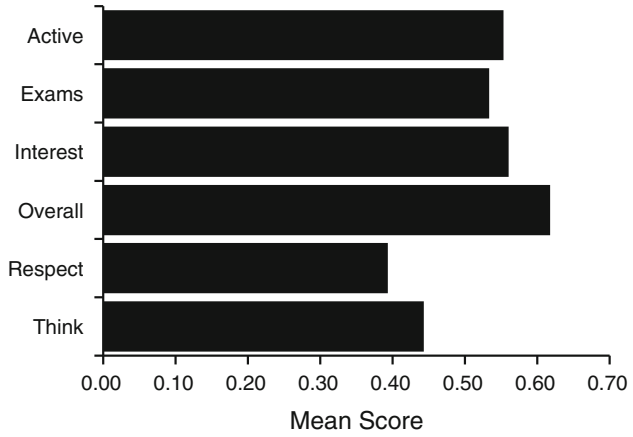

Fig. 8 Profile 2 (Cluster 5) for amount of learning

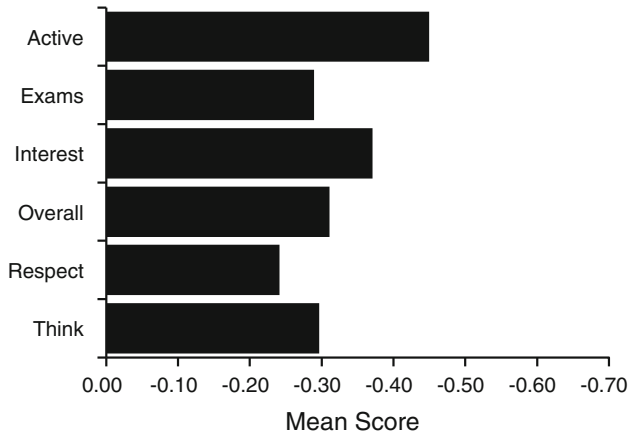

Fig. 9 Profile 3 (Cluster 2) for amount of learning (with *x* axis reversed)

relationship between learning and instructional effectiveness can be attributed to other contaminating factors or biases such as faculty rank, student motivation, and workload. As Cashin (1995) stated, if a factor has an influential effect on student ratings, it should be statistically controlled to obtain unbiased information about instructional effectiveness.

It is noteworthy that the effectiveness measures did not receive equally high scores for instructors who taught higher level classes in learning. Quality of examinations, respect shown by instructors toward students, and the development of thinking skills are measures not related to achievement indicators. Young and Shaw (1999) suggested that a teacher characterized as effective with higher learning may not necessarily receive equally high ratings for all measures. Instructors may still be effective even when some aspects of instruction received lower scores from students. The results of the present study provide supporting evidence from a different perspective. Although instructional skills are important to students' achievement levels, an instructor may not have to possess all skills to be successful. For example, in the present study, instructors in the cluster with the highest achievement level received ratings that did not differ from the whole group for the items "evaluation of assessment material" and "developing critical thinking skills". Similarly, for the cluster having the highest amount of learning, the item "developing critical thinking skills" was not rated as high as the other measures. This is also supported by McKeachie (1997), who suggested that "effective teachers come in all shapes and sizes" (p. 1218).

Instructors were rated lower in respectful behavior compared with other measures. Moreover, for some clusters, that measure does not have a difference from the whole body. Coladarci and Kornfield (2007) stated that respectful behavior is highly correlated to external criteria of student ratings, but the findings of the present study revealed that the respect variable was generally rated by lower scores. When compared with other measures related to active participation, assessment material, and assessment of respectful behavior of instructors, a more abstract variable may be difficult for students to evaluate.

Based on the results of the present study, the following conclusions can be drawn: (i) It does not seem possible to provide a general definition of instructional effectiveness that is valid for all instructors whose classes exhibit differences in the amount of learning. For different clusters including courses with higher amount of learning, instructors may be labeled "effective" although they receive different scores for different facets unless they have higher means in general. (ii) The relationship between learning and instructional effectiveness exists; however, it is not perfect. (iii) Investigation of student ratings or any other related issue should be made by grouping data into meaningful subgroups that can provide more informative results. (Marsh and Hocevar 1991; Trivedi et al. 2011).

References

Abrami, P. C., d'Appollonia, S., & Cohen, P. A. (1990). The validity of student ratings on instruction: What we know and what we do not. *Journal of Educational Psychology, 82*(2), 219–231.

Abrami, E. C., & d'Apollonia, S. (1991). Multidimensional students' evaluations of teaching effectiveness: generalizability of "n = 1" research: Comment on Marsh (1991). *Journal of Educational Psychology, 83*(3), 415–441.

Aldenderfer, M. S., & Blashfield, R. K. (1984). *Cluster analysis*. Beverly Hills: Sage Publications.

Aleamoni, L. M., & Hexner, P. Z. (1980). A review of the research on student evaluation and a report on the effect of different sets of instructions on student course and instructor evaluation. *Instructional Science, 9*, 67–84.

Atamian, R., & Ganguli, G. (1993). Teacher popularity and teaching effectiveness: Viewpoint of accounting students. *Journal of Education for Business, 68*(3), 163–169.

Benz, C., & Blatt, S. J. (1995). Factors underlying effective college teaching: What students tell us. *Mid-Western Educational Researcher, 8*(1), 27–31.

Borden, V. M. H. (1995). Segmenting student markets with a student satisfaction and priorities survey. *Research in Higher Education, 30*(1), 73–88.

Braskamp, L. A., & Ory, J. C. (1994). *Assessing faculty work: Enhancing individual and institutional performance*. San Francisco: Jossey-Bass.

Cashin, W. E. (1995). Student ratings of teaching: the research revisited (idea paper No. 32). Manhattan: Kansas State University, Center for Faculty Evaluation and Development.

Cashin, W. E., & Downey, R. G. (1992). Using global student rating items for summative evaluation. *Journal of Educational Psychology, 84*, 563–572.

Cashin, W. E., Downey, R. G., & Sixbury, G. R. (1994). Global and specific ratings of teaching effectiveness and their relation to course objectives: Reply to Marsh (1994). *Journal of Educational Psychology, 86*, 649–657.

Coladarci, T., & Kornfield, I. (2007). RateMyProfessors.com versus formal in-class student evaluations of teaching. *Practical Assessment, Research and Evaluation*, *12*(6). Retrieved 20 Oct 2012 from pareonline.net/getvn.asp?v=12andn=6.

Dawson, N. (1986). Hours of contact and their relationship to students' evaluations of teaching effectiveness. *The Journal of Nursing Education, 25*(6), 236–239.

Donaldson, J. F., Flannery, D., & Ross-Gordon, J. (1993). A triangulated study comparing adult college students' perceptions of effective teaching with those of traditional students. *Continuing Higher Education Review, 57*(3), 147–165.

Ellett, C. D., Loup, K. S., Culross, R. R., McMullen, J. H., & Rugutt, J. K. (1997). Assessing enhancement of learning, and student efficacy: Alternatives to traditional faculty evaluation in higher education. *Journal of Personnel Evaluation in Education, 11*(2), 167–192.

Feldman, K. A. (1983). Seniority and experience of college teachers as related to evaluations they receive from students. *Research in Higher Education, 18*, 3–124.

Feldman, K. A. (1988). Effective college teaching from the students' and faculty's view: matched or mismatched priorities? *Research in Higher Education, 28*(4), 291–329.

Feldman, K. A. (1989). Instructional effectiveness of college teachers as judged by teachers themselves, current and former students, colleagues, administrators, and external (neutral) observers. *Research in Higher Education, 30*(2), 137–194.

Feldman, K. A. (1997). Identifying exemplary teachers and teaching: Evidence from student ratings. In R. P. Perry & J. C. Smart (Eds.), *Effective teaching in higher education: research and practice* (pp. 368–395). New York: Agathon Press.

Fernandez, J., & Mateo, M. A. (1997). Student and faculty gender in rating of university teaching quality. *Sex roles: A Journal of Research, 37*(11–12), 997–1003.

Freeman, H. R. (1994). Student evaluations of college instructors: Effects of type of course taught, instructor gender and gender role, and student gender. *Journal of Educational Psychology, 86*(4), 627–630.

Gigliotti, R. J., & Buchtel, F. S. (1990). Attributional bias and course evaluations. *Journal of Educational Psychology, 82*, 341–351.

Greenwald, A. G., & Gillmore, G. M. (1997a). Grading leniency is a removable contaminant of student ratings. *American Psychologist, 52*, 1209–1217.

Greenwald, A. G., & Gillmore, J. M. (1997b). No pain, no gain? The importance of measuring course workload in student ratings of instruction. *Journal of Educational Psychology, 89*, 743–751.

Johannessen, T. A. (1997). What is important to students? Exploring dimensions in their evaluations of teachers. *Scandinavian Journal of Educational Research, 41*(2), 165–177.

Jöreskog, K.G. & Sörbom, D. (1999). LISREL 8.30 for Windows [Computer software]. Skokie, IL: Scientific Software International, Inc.

Kalender, I. (2011). Contaminating factors in university students' evaluation of instructors. *Education and Science, 36*(162), 56–65.

Kelloway, E. K. (1998). *Using lisrel for structural equation modeling*. New York NY: Guildford Press.

Kline, R. B. (2005). *Principles and practices of structural equation modeling* (2nd ed.). New York: Guilford Press.

Kockelman, K. M. (2001). Student grades and course evaluations in engineering: What makes a difference. *ASEE Annual Conference Proceedings*, 9085–9110.

Lin, W. Y. (1992). Is class size a bias to student ratings of university faculty? A review. *Chinese University of Education Journal, 20*(1), 49–53.

Ludwig, J. M., & Meacham, J. A. (1997). Teaching controversial courses: Student evaluations of instructor and content. *Educational Research Quarterly, 21*(1), 27–38.

Marsh, H. W. (1984). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases, and utility. *Journal of Educational Psychology, 76*(5), 707–754.

Marsh, H. W., & Bailey, M. (1993). Multidimensional students' evaluations of teaching effectiveness: a profile analysis. *The Journal of Higher Education, 64*(1), 1–18.

Marsh, H. W., & Hocevar, D. (1991). The multidimensionality of students' evaluations of teaching effectiveness: the generality of factor structures across academic discipline, instructor level, and course level. *Teaching and Teacher Education, 7*, 9–18.

Marsh, H. W., & Roche, L. A. (1993). The use of students' evaluations and an individually structured intervention to enhance university teaching effectiveness. *American Educational Research Journal, 30*, 217–251.

Marsh, H. W., & Roche, L. A. (1997). Making students' evaluations of teaching effectiveness effective. *American Psychologist, 52*(11), 1187–1197.

McKeachie, W. J. (1997). Student ratings. *American Psychologist, 52*, 1218–1225.

Norusis, M. (2005). *SPSS 13.0 Guide to Data Analysis*. Upper Saddle River, NJ: Prentice Hall.

Rayder, N. F. (1968). College student ratings of instructors. *Journal of Experimental Education, 37*, 76–81.

Rodabaugh, R. C., & Kravitz, D. A. (1994). Effects of procedural fairness on student judgments of professors. *Journal on Excellence in College Teaching, 5*(2), 67–83.

Sailor, P., Worthen, B., & Shin, E. H. (1997). Class level as a possible mediator of the relationship between grades and student ratings of teaching. *Assessment & Evaluation in Higher Education, 22*(3), 261–269.

Sixbury, G. R., & Cashin, W. E. (1995). Comparative data by academic field. In Idea technical report, vol. 10: Center for Faculty Evaluation and Development, Manhattan, Kansas: Kansas State University.

Sonquist, J.A., & Morgan, J. N. (1964). *The detection of interaction effects: a report on a computer program for the selection of optimal combinations of explanatory variables*. Monograph no: 35, Survey Research Center, Institute for Social Research, University of Michigan.

Tatro, C. N. (1995). Gender effects on student evaluations of faculty. *Journal of Research and Development in Education, 28*(3), 169–173.

Teven, J. J., & McCroskey, J. C. (1997). The relationship of perceived teacher caring with student learning and teacher evaluation. *Communication Education, 46*(1), 1–9.

Thomas, E. H., & Galambos, N. (2004). What satisfies students? Mining student-opinion data with regression and decision-tree analysis. *Research in Higher Education, 45*(3), 251–269.

Trivedi, S., Pardos, Z. A., & Heffernan N. T. (2011).Clustering students to generate an ensemble to improve standard test score predictions. In *Proceedings of the 15th International Conference on Artificial Intelligence in Education*, Auckland, New Zealand.

VanArsdale, S. K., & Hammons, J. O. (1995). Myths and misconceptions about student ratings of college faculty: Separating fact from fiction. *Nursing Outlook, 43*(1), 33–36.

Wilson, R. (1998). New research casts doubt on value of comparing adult college students perceptions of effective teaching with those of traditional students. *Chronicle of Higher Education, 44*(19), A12–A14.

Young, S. M., & Shaw, D. G. (1999). Profiles of effective college and university teachers. *The Journal of Higher Education, 70*(6), 670–686.