RASCH ANALYSIS OF THE NBC 461 INSTRUMENT FOR FACULTY EVALUATION

Maria Azucena B. Lubrica[1] and Joel V. Lubrica[2]

ABSTRACT

Responses of university students to a 20-item Student Evaluation of Faculty Instrument having a 5-point Likert-type scale were analyzed through Rasch Measurement Theory. The primary aims were to determine: a) the reliability of the Instrument in measuring faculty performance; and, if each item in the Instrument can be considered as an indicator of performance. Participants were from 882 students of various degree programs in Benguet State University, who rated 7 teachers. Results revealed that, as a whole, the Instrument was reliable and that seventeen of the items can be considered as independent indicators of performance. The other three items might need to be re-phrased so that they can also become indicators.

KEYWORDS: Faculty evaluation; Rasch analysis

INTRODUCTION

In the State Universities and Colleges in the Philippines, the evaluation of faculty by students continues to be an important component in monitoring and providing feedback regarding the performance of faculty members. At Benguet State University (BSU), the student evaluation comprises 60% of the Performance Evaluation Scheme (PES) of a full-time (i.e., on a Teacher's Leave basis) faculty member, with the other 40% being composed of the Chairperson's Evaluation (25%) and Peer Evaluation (15%). Although teaching effectiveness evaluation is a high-stakes activity because it is used as basis for retention, promotion, tenure and pay raises (SUNT, 1997), the fundamental reason for this type of evaluation should be the improvement of teaching (MPS, 2000).

Starting on the 2nd semester of school year 2007-2008, through Administrative Order No. 2008-04 dated February 20, 2008, a new instrument for the student evaluation of faculty was utilized at BSU, replacing the existing one. This new instrument was based on the Philippine Association of State Universities and Colleges (PASUC) guidelines on the evaluation of teaching effectiveness for the uniform implementation of National Budget Circular (NBC) 461 Qualitative Contribution Effectiveness (QCE), in the area of instruction. It was developed in order to satisfy the requirements of the CHED (Commission on Higher Education) Zonal Center for the uniform implementation of NBC 461 in the zone that covers the Cordillera Administrative Region, Region 1 and Region 2. The instrument is divided into four parts of equal weight, with each part being composed of 5 items that are to be rated from 1 (poor) to 5 (outstanding). Thus, the maximum 'score' that a teacher can have for this instrument is 100, if all of the 20 indicators of teaching performance are given the rating of 5. The instrument has some form of validity, having been considered fully by an appropriate committee before it was finalized and recommended for use. However, these considerations do not preclude the conduct of studies to examine the instrument. Possible questions that can be answered are, "Can each item in the instrument

[1] *Associate Professor IV, is a member of the Department of Mathematics- Physics- Statistics of the College of Arts and Sciences.;*

[2] *Professor V, is a member of the Department of Mathematics- Physics- Statistics of the College of Arts and Sciences.*

be considered as an indicator of the performance of faculty members?", and "Are the items working coherently to reveal the performance of a faculty member?"

At BSU, there were already studies involving faculty evaluation by students. One was conducted by Lubrica and Lubrica (2009), who used Rasch analysis to investigate the Student Evaluation of Faculty Instrument utilized in 2005. This instrument was a 24-item questionnaire involving a 5-point Likert-type scale (i.e, Excellent, Very Good, Good, Fair, Poor). The researchers discovered that only seventeen of the items in the Instrument could be considered as indicators of teaching performance. Moreover, they claimed that there were interactions between students' responses and gender of teacher and/or subject matter taught. The authors recommended that the seven items that did not fit the Rasch model had to be re-phrased so that their meanings would not differ from one student to the next.

Another study was done by the Department of Mathematics-Physics-Statistics in 2000 (MPS, 2000). The primary aim was to produce a smaller set of items through factor analysis. Still another study was done by the same Department (MPS, 1995), making use of the results of another version of the Student Evaluation of Faculty Instrument and relating these to selected teacher variables such as department, age, faculty rank, and length of service. The statistical techniques used were analysis-of-variance and Pearson-product moment correlation analysis. Results indicated that the performance of teachers differed significantly when grouped according to subject.

Overseas, Bond (2005) made use of Rasch analysis to develop the Student Feedback About Teaching Instrument at James Cook University in Australia and to estimate how difficult or easy it was, on the average, for students to endorse each item in this Instrument. The study revealed interactions between class size and endorsement of teaching practices. On the other hand, the study of Onwuegbuzie, *et al*. (2007) was a systematic inquiry into students' perceptions regarding characteristics of effective college teachers. The researchers discovered that there were interactions between students' backgrounds and perceptions.

As a whole, the present study has relations to the above-cited researches. These are along the aspects of investigating an evaluation instrument and/or the use of Rasch measurement theory. Specifically, the present study attempted to determine if: a) the Instrument was reliable in providing a measure of faculty performance; and, b) each item could be considered as an indicator of performance of a faculty member.

MATERIALS AND METHODS

Data from the evaluation of 882 students for 7 faculty members of the Department of Mathematics-Physics-Statistics for 2nd semester 2007-2008 were used. Access to data and subsequent analysis were done from July 2009 to May 2010.

Rasch analysis, also called as Rasch Measurement Theory (Wright & Masters, 1982), was used in the treatment of data. It is an item response measurement model, often equated with one-parameter model (Karabatsos, 1999), in the form of a conjoint measurement where two quantities (e.g., ability of a person and difficulty of a test item) can be measured separately because of their interaction (Bond & Fox, 2001). In short, it uses the idea of independence by separating person (or student) and item parameters. This independence is analogous to a situation in physics whereby relative masses (analogous to 'student perceptions') of, say, five objects are measured through their responses (i.e., acceleration) to various forces (the 'items'). In the end, although there was an interaction between masses and forces, these masses and forces can be separated from each other, leading to the ascertainment of the relative values of masses, irrespective of the forces

(Wright & Stone, 1979). Symmetrically, the relative values of the forces can be obtained regardless of the masses, because it was the acceleration (the 'response') that was considered. Thus, both student and item parameters (e.g., person or item fit) can be measured at the same.

What this means is that, Rasch Analysis used the responses of students to the 20 items in the Faculty Evaluation Instrument in order to provide measurements regarding the items themselves. These measurements, specifically, were 1) item reliability index, which connotes the coherence of the items in indicating faculty performance, and 2) item fit, which indicates whether or not an item can be included in the instrument.

WINSTEPS, a software for Rasch analysis that was developed by Linacre (2005), was used in the computations.

RESULTS AND DISCUSSION

In relation to reliability of the Instrument, Rasch modeling produced an item reliability index of 0.96. This is above the cut-off value of 0.7 (Wright & Masters, 1982), and indicates that the Instrument has a high reliability in measur-ing faculty performance. That is, the items are working coherently in providing a measure of the performance of a faculty member, as per-ceived by students.

Since the item reliability index, which is equivalent to Cronbach's alpha (Bond & Fox, 2001) , indicates the replicability of responses to the items, it can be deduced further that the item characteristics are stable and would not change if these same items were administered to another sample of students of comparable circumstances. In short, there can be confidence in the consistency of characteristics, such as item fit and item order, if one wanted to delve deeper into these.

In relation to each item being an indicator of performance, the fit of items to the Rasch measurement model was considered (Table 1). Seventeen of the 20 items had infit values that were less than the cut-off of 1.20, applicable to samples between 500 and 1000 (Smith, Schumacker, & Bush, 1998, cited by Bond & Fox, 2001). This result shows that the 17 items had fit to the Rasch model and implies that they can be considered as indicators of teacher performance. These items are:

- Demonstrates sensitivity to students' ability to attend and absorb content information

- Integrates sensitively his/her learning objectives with those of the students in a collaborative process

- Makes self available to students beyond official time

- Keeps accurate records of students' performance and prompt submission of same

- Demonstrates mastery of subject matter (explains the subject matter without relying solely on the prescribed textbook)

- Draws and shares information on the state of the art of theory and practice in his/her discipline

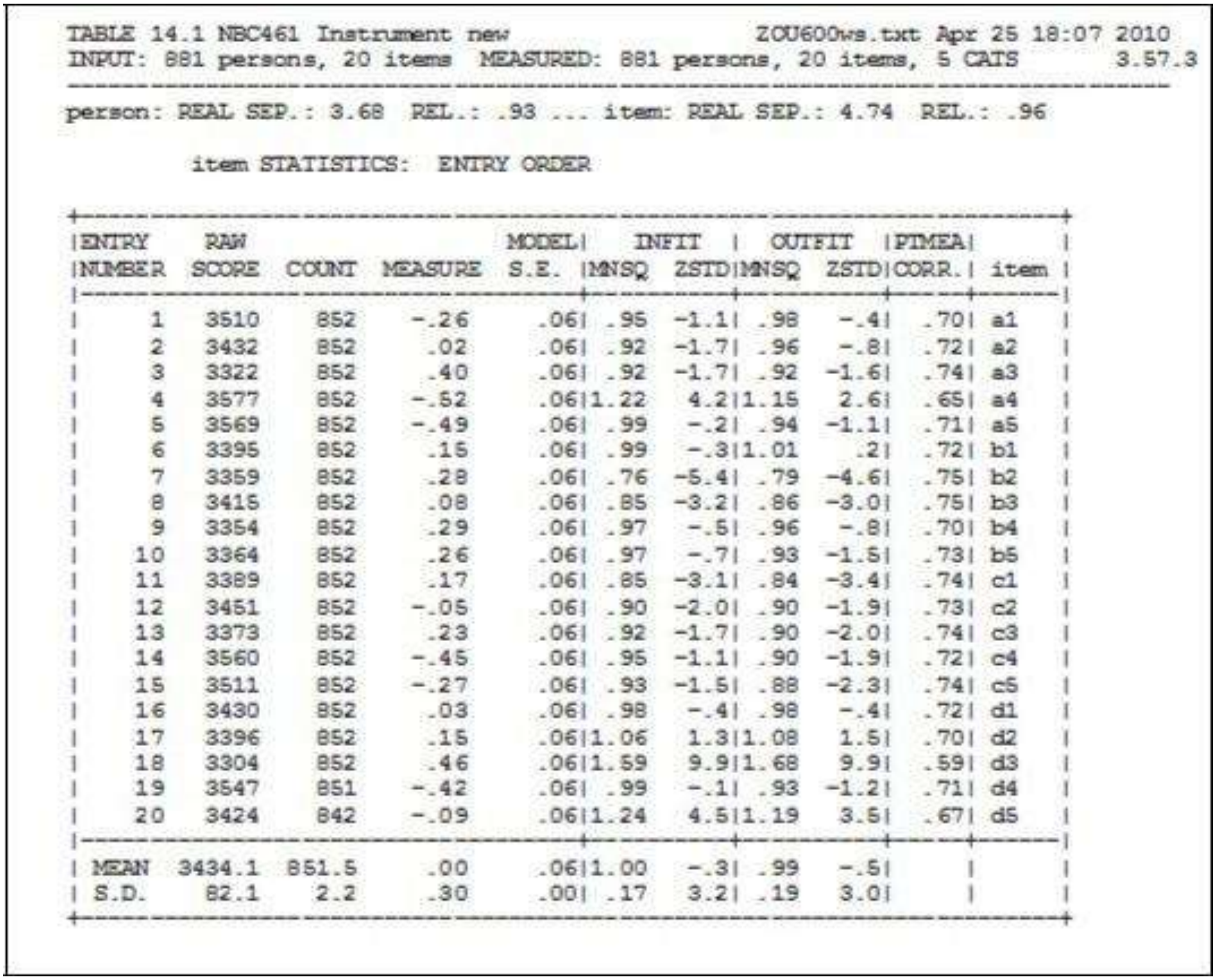
- Integrates subject matter to practical circumstances and learning intents/ purposes of students

- Explains the relevance of present topics to the previous lessons, and relates the subject matter to relevant current issues and/ or daily life activities

- Demonstrates up-to-date knowledge and/or awareness on current trends and issues of the subject

- Creates teaching strategies that allow students to practice using concepts they need

Table 1. Item Fit



```
TABLE 14.1 NBC461 Instrument new                  ZOU600ws.txt Apr 25 18:07 2010
INPUT: 881 persons, 20 items  MEASURED: 881 persons, 20 items, 5 CATS      3.57.3
--------------------------------------------------------------------------------
person: REAL SEP.: 3.68  REL.: .93 ... item: REAL SEP.: 4.74  REL.: .96

        item STATISTICS:  ENTRY ORDER

+------------------------------------------------------------------------------+
|ENTRY    RAW                       MODEL|   INFIT  |  OUTFIT  |PTMEA|         |
|NUMBER  SCORE  COUNT  MEASURE  S.E. |MNSQ  ZSTD|MNSQ  ZSTD|CORR.| item    |
|------------------------------------+----------+----------+-----+---------|
|     1   3510    852     -.26     .06| .95  -1.1| .98   -.4| .70| a1      |
|     2   3432    852      .02     .06| .92  -1.7| .96   -.8| .72| a2      |
|     3   3322    852      .40     .06| .92  -1.7| .92  -1.6| .74| a3      |
|     4   3577    852     -.52     .06|1.22   4.2|1.15   2.6| .65| a4      |
|     5   3569    852     -.49     .06| .99   -.2| .94  -1.1| .71| a5      |
|     6   3395    852      .15     .06| .99   -.3|1.01    .2| .72| b1      |
|     7   3359    852      .28     .06| .76  -5.4| .79  -4.6| .75| b2      |
|     8   3415    852      .08     .06| .85  -3.2| .86  -3.0| .75| b3      |
|     9   3354    852      .29     .06| .97   -.5| .96   -.8| .70| b4      |
|    10   3364    852      .26     .06| .97   -.7| .93  -1.5| .73| b5      |
|    11   3389    852      .17     .06| .85  -3.1| .84  -3.4| .74| c1      |
|    12   3451    852     -.05     .06| .90  -2.0| .90  -1.9| .73| c2      |
|    13   3373    852      .23     .06| .92  -1.7| .90  -2.0| .74| c3      |
|    14   3560    852     -.45     .06| .95  -1.1| .90  -1.9| .72| c4      |
|    15   3511    852     -.27     .06| .93  -1.5| .88  -2.3| .74| c5      |
|    16   3430    852      .03     .06| .98   -.4| .98   -.4| .72| d1      |
|    17   3396    852      .15     .06|1.06   1.3|1.08   1.5| .70| d2      |
|    18   3304    852      .46     .06|1.59   9.9|1.68   9.9| .59| d3      |
|    19   3547    851     -.42     .06| .99   -.1| .93  -1.2| .71| d4      |
|    20   3424    842     -.09     .06|1.24   4.5|1.19   3.5| .67| d5      |
|------------------------------------+----------+----------+-----+---------|
| MEAN  3434.1  851.5      .00     .06|1.00   -.3| .99   -.5|     |         |
| S.D.    82.1    2.2      .30     .00| .17   3.2| .19   3.0|     |         |
+------------------------------------------------------------------------------+
```

to understand (interactive discussion)

• Enhances student self-esteem and/ or gives due recognition to students performance/ potentials

• Allows students to create their own course with objectives and realistically defend student-professor rules and make them account-able for their performance

• Allows students to think independently and make their own decisions and holding them accountable for their performance based largely on their success in executing decisions

• Encourages students to learn beyond what is required and help/ guide the students how to apply the concepts learned

• Creates opportunities for intensive and/ or extensive contribution of students in the class activities (e.g., breaks the class into dyads, tri-ads or buzz/ task groups)

• Assumes roles as facilitator, resource person, coach, inquisitor, integrator, referee in drawing student to contribute knowledge and understanding of the concepts at hand

• Structures/ re-structures learning and teaching-learning context to enhance attainment of collective learning objectives

On the other hand, there were three mis-fitting items -- items number 4, 18 and 20 – because they had infit values greater than 1.2. What this result implies is that there were inconsistencies in responses of students to each of these three items (Bond & Fox, 2001). In short, students were interpreting them in various ways.

The three mis-fitting items are shown in the first column of Table 2, vis-à-vis seven items in the old Instrument that were also interpreted inconsistently, as revealed by Lubrica and Lubrica's (2009) study. It is notable that there is agreement between the results of the two studies. For example, both involve regularity in attendance, timeliness in coming to class, being in proper attire, and so on. Both also involve learning conditions that promote freedom of expression or reinforce learning.

The inconsistencies in interpretation could be due to the various backgrounds of stu-dents (Onwuegbuzie *et al.*, 2007). For instance, the Item #4 "Regularly comes to class on time, well-groomed and well-prepared to complete assigned responsibilities", aside from being composed of four different ideas, can be interpreted with emphasis on the aspect of regularity. That is, a student who had much experience regarding absentee teachers might think that "regular" meant two or three total absences in an entire semester, while another (who was, perhaps, exposed to a school atmosphere where teachers were always present) might think that it meant no absences at all. Furthermore, for one student, the term "on time" might mean on time according to his/her timepiece, while for another, it is might be according to the time of a favorite FM station. In addition, the term "well-groomed" could have different connotations to different students, and a teacher can be considered as well-groomed or not, depending on the 'standards' of a student. Finally, the term "well-prepared" can be assumed to be true by some students, while others might need to see evidence of the preparedness of a teacher, such as the bringing of lecture notes or similar materials.

For Item #18 "Designs and implements learning conditions and experience that promote healthy exchange and/ or confrontations", it is possible that there are inconsistencies in interpretation because the terms "healthy exchange" and "confrontations" can be considered as contradictory. That is, a student might give a high rating to a teacher because "healthy exchange" is promoted; on the other hand, another student might give a low rating to the same teacher because the teacher promotes "confrontations", which has a negative connotation. Moreover, the

Table 2. Item Misfits

PRESENT STUDY	LUBRICA and LUBRICA'S (2009) STUDY ON THE OLD INSTRUMENT
1 (Item #4) Regularly comes to class on time, well-groomed and well-prepared to complete assigned responsibilities	1 Regular in attendance 2 Comes to class on time 3 Starts and dismisses classes on time 4 Comes to class in proper attire 5 Returns corrected papers promptly
2 (Item #18) Designs and implements learning conditions and experience that promote healthy exchange and/ or confrontations	6 Relates subject matter to real life situations
3 (Item #20) Uses instructional materials (audio/ video materials, fieldtrips, film showing, computer aided instruction, etc.) to reinforce learning processes	7 Makes students feel free to inquire, express ideas, or disagree

phrase “designs and implements” is composed of two related but different attributes. “Designs” can be considered as a hidden attribute and cannot readily be given a rating by students. On the other hand, “implements” can readily be rated by students because whatever is implemented is what they experience in class. So, perhaps some students might give a high rating to a teacher because they assumed that the teacher actually designed whatever was being implemented, while other students might give a different rating because they might not be sure whether the teacher really designed the implemented learning environment or not.

For Item #20 “Uses instructional materials (audio/ video materials, fieldtrips, film showing, computer aided instruction, etc.) to reinforce learning processes”, there could be inconsistencies in interpretation because the word “fieldtrips” might, to some students, be a stronger basis for giving a rating than the others mentioned. That is, one student might say that no fieldtrips were done at all, thus a low rating should be given. On the other hand, another student might consid-er not having experienced a fieldtrip under the same teacher as a minor matter, compared to the use, perhaps, of the teacher of audio/video materials. Thus, the latter student will give the teacher a high rating.

In general, the inconsistencies in interpreting the three mis-fitting items appear to arise from many meanings that the wordings or phrasing of the items imply. The implication is that these items have to be made more specific or concrete.

Should these three items then be deleted from the Instrument, because they were being interpreted inconsistently? The advice is that, “there are no hard and fast rules” in the interpretation of fit statistics (Bond & Fox, 2001), and item mis-fit should make the researcher “Think again!”, and not “Throw out!”” mis-fitting items (Bond & Fox, 2001). The implication is that these three items can still be included in the Instru-ment. What is needed perhaps is to re-phrase them so that they can have only one meaning. For example, coming to class on time can per-haps be specified in terms of number of minutes from the scheduled start of class, so that stu-dents can have a consistent measure of punctu-ality. Being well-groomed should be given some sort of measure, such as wearing of the Univer-sity uniform when required. Using instructional materials should be quantified, possibly in terms of the number of times these are used in one week, and so on.

This re-phrasing has to be done because it has to be recognized that results can only give a reliable and indispensable perspective on teachers’ performance provided that evalua-tion instruments are properly constructed (CTL, 1994). It is also aligned with the idea that these kinds of instruments should be ‘data-driven’, in the sense that they should be based on feedback from students (whether directly or indirectly – for instance, as inconsistent responses in this case), and not purely based on perspectives of faculty or administrators (Onwuegbuzie, et al., 2007).

The result that some items were be-ing interpreted in various ways has conformity with the findings of Onwuegbuzie *et al.* (2007) that individual differences existed with respect to students’ perceptions of the characteristics of effective college teachers. For the specific item involving punctuality, Bond’s (2005) study also revealed that an item about punctuality did not fit the Rasch model, indicating some form of cross-cultural similarity of interpretation of time.

Nevertheless, if the mis-fit of the three items were weighed against the fit of the other 17, it can still be said that the Instrument, as a whole, had reliability as indicated by the item reliability index value of 0.96.

SUMMARY, CONCLUSION, AND RECOMMENDATION

This study investigated the Student Faculty Evaluation Instrument being utilized at BSU for the implementation of NBC 461 Qualitative Contribution Effectiveness in the area of instruction. Rasch analysis was done for data that involved 882 student evaluations for 7 faculty members. It was found out that the Instrument, as a whole, was reliable. Also, 17 out of the 20 items that comprised it were indicators of faculty performance in the classroom; the other 3 can be re-phrased so that they can have better fit to the Rasch model, and can then be considered as indicators of performance.

Due to the high item reliability index and the general fit of the items to the Rasch model, it can be concluded that the items comprising the Instrument were working together coherently, or functioning in unison, in order to reveal faculty performance. Or, from the student side, it can be said that the response to each item is affected by the same process and in the same form (Bejar, 1983, cited by Bond & Fox, 2001). That is, it can be assumed that only one trait was affecting the response patterns, in consonance with the important uni-dimensionality requirement of the Rasch model (Andrich, 1999). All of these points to the same idea: that one can accept the Instrument as being able to provide a measure of faculty performance, from the point of view of students.

Based on the findings, the following are recommended:

1. The three mis-fitting items have to be re-phrased so that the performance they are measuring can be quantified. a) For example, Item #4 "Regularly comes to class on time, well-groomed and well-prepared to complete assigned responsibilities" may be re-phrased as "Has no absences, comes to class within five minutes of the scheduled start of class, wears the University uniform on designated days, and brings instructional aids and lesson plan in every class". Even better would be to divide it into four, possibly: "Has no absences", "Comes to class within five minutes of the scheduled start of class", "Wears the University uniform on designated days", and "Brings instructional aids and lesson plan in every class". b) Item #18 "Designs and implements learning conditions and experience that promote healthy exchange and/ or confrontations" might be re-phrased as "Implements learning conditions that promote healthy exchange of ideas." c) Item #20 "Uses instructional materials (audio/ video materials, fieldtrips, film showing, computer aided instruction, etc.) to reinforce learning processes" might be re-phrased as "Uses various aids (such as utilizing audio/video materials, fieldtrips, film showing, computer aided instruction, etc.) to reinforce learning processes".

2. It is noteworthy that the BSU Administration has promulgated a university-wide memorandum for an official time that follows the Philippine Standard Time. However, this official time has to be given a more tangible implementation, possibly through the ringing of a bell like in other institutions, so that the punctuality of teachers will be given the same interpretation by students.

3. Further research involving qualitative data, probably obtained through interviews with faculty members and students, may be done to validate the findings of this study, because these were derived mainly from numerical responses of students.

LITERATURE CITED

Andrich, D. 1999, ‘Rating scale analysis’, in Advances in Measurement in Educational Research and Assessment, eds G. Masters & J. Keeves, Pergamon, Oxford, UK, pp. 110-121.

Bond, T. G. 2005. Accountability in the Academe: Rasch Measurement of Student Feedback Surveys. In Frontiers in Educational Psychology. Ed: R. Waugh. Nova Science Publishers, Inc. pp. 119-129.

Bond, T. G. and C. M. Fox, 2001. Applying the Rasch Model: Fundamental Measurement in the Human Sciences. Laurence Erlbaum Associates, Inc., USA: New Jersey. p. 104, 179

Bejar, I. I., 1983, Achievement Testing: Recent Advances. Beverly Hills. CA: Sage.

CTL (Center for Teaching and Learn-ing). 1994, Student Evaluation of Teach-ing. Center for Teaching and Learn-ing, University of North Carolina at Chapel Hill. P.16.

Karabatsos, G. 1999, Rasch vs. Two- and Three-parameter Logistic Models from the Perspective of Conjoint Measurement Theory. Paper presented at the 32nd Annual Meeting of the American Education Research Association, Montreal, Canada, 1999.

Linacre, J. M. 2005. WINSTEPS Rasch Measurement Computer Program, Chicago: Winsteps.com.

Lubrica, J.V. and M.A.B.Lubrica (2009), A Rasch Analysis of the Student Evaluation of Physics, Mathematics and Statistics Faculty Members, presented in the 2009 International Conference on Physics Education held in Bangkok, Thailand, on October 18-24, 2009.

MPS (Math-Physics-Statistics Department) 1995. Teaching Performance of CAS Faculty as Perceived by Students. CAS Research Digest 1995. Benguet State University.

MPS (Math-Physics-Statistics Department) 2000. Factor Analysis of BSU Students’ Evaluation of Faculty. College of Arts and Sciences, Benguet State University.

Onwuegbuzie, A. J., A. E. Witcher, K. M. T. Collins, J. D. Filer, C. D. Wied maier and C. W. Moore, 2007. Stu-dents’ Perceptions of Characteristics of Effective College Teachers: A Validity Study of a Teaching Evaluation Form Using a Mixed-Methods Analysis. Ameri-can Educational Research Journal, 44: 113-169.

Smith, R. M, R.E. Schumacker, and M. J. Bush, 1998, Using Item Mean Squares to Evaluate Fit to the Rasch Model. Journal of Outcome Measurement 2(1), 66-78.

SUNT (Stanford University Newslet ter on Teaching) 1997. Using Stu-dent Evaluations to Improve Teaching. Stanford University Newsletter on Teaching, 9(1).

Wright, B. and G. Masters, 1982. Rating Scale Analysis. Chicago: MESA Press.

Wright, B. D. and M. Stone, 1979. Best Test Design. Chicago: MESA Press.