Applying the Item Response Theory to English Classroom Examinations for Ethnic Students

Lanfen Ji[1], Xiaoqin Zhang[2], Dianjun Lu[3], Dianxiang Lu[4*]

[1]Foreign Language Department of Qinghai Normal University, Xining 810008, China
[2]Financial and Economic College of Qinghai University, Xining 810008, China
[3]Department of Mathematics of Qinghai Normal University, Xining 810008, China
[4]Medical College of Qinghai University, Xining 810001, China

Abstract

Tests and assessment are necessary for the presentation of the language learners' competence in the process of learning. This paper discussed and is trying to find out the particularity and characteristics of English teaching at minority regions in consideration of the influencing factors: trilingual teaching, small classes, fewer students. The measurement of constructs of English classroom examination at minority regions is put forward. The techniques used in this paper are trying to keep pace with the advances in psychometric theory and methods. This paper presents an example of applying item response theory to English classroom examinations at minority regions and demonstrates graded response models.

Keywords: English Classroom Examination, Item Response Theory, Graded Response Models;

1. Introduction

Language learners need to be tested or assessed and will be presented with a set of the competence that the learners have in the language study. When students are assessed, both their competence and performance in the use of language are focused. Just like in other subjects, language tests are given with a variety of purposes in mind. However in a second language situation, two major purposes can be identified and these are (a) to determine how much the learner has learnt from the planned syllabus. This is called Achievement testing. (b) to determine the strength and weakness found in the language used by the students. This is known as diagnostic testing.

The purpose of writing this paper is trying to find out the similarities and dissimilarities of teaching English in regular college classrooms and in classrooms at minority regions. English learning characteristics of college students are: trilingual teaching, small classes, fewer students. How to test English classroom teaching effect is particularly challenging for faculty. The analyses of most classroom examinations suffer because of small sample sizes. In contrast to majority students, students at minority regions are multi-faceted, having more diverse educational experiences. Some groups of students are better prepared academically than other groups. Other students are not as prepared owing to not being able to graduate on time or they just were not sure what to do with their lives. Teachers who sample classroom examinations are challenged because students' knowledge, skills, and attitudes vary. It is important for teachers to find out whether the variance is caused by differences in exami-

nation difficulty or student ability. The diversity of student abilities decides test equating using standard statistical methods.

The most notable development in the last forty years is item response theory (IRT). IRT is a family of techniques for understanding the psychometric properties of measures and relationships between properties of the measures and the individuals completing those measures. These techniques have been used to advance our understanding in a variety of researches. Many of the constructs of interest in English classroom examination research are assessed using self-report or other-report measures and the conceptualization of these constructs is becoming more complex (e.g., Marion & Uhl-Bien, 2001; Osborn et al., 2002). As such, gaining a better understanding of the measures used to support English classroom examination theories and constructs is a prudent course for English classroom examination research to take (Schriesheim et al., 1993).

2. Graded response models

In most studies on language learning, authors usually aim at attitudinal, and personality research, attitudinal, and personality research, the measures utilize response options that are ordered (e.g., Likert-type response options) and the data are used to make conclusions concerning the level of the construct that is assessed by the measure. In these cases, graded response models can be used. In these models, the IRT analysis examines the relationships between item or option parameters, person parameters, and the selection of a particular option. For cumulative graded response models, it is assumed that the value of the latent trait is smaller for individuals who choose the first response option than for individuals who choose the second response option in an ordered response set. This assumption exemplifies the dominance response process that underlies these models.

The relationship between the estimate of the latent trait, the response option characteristics, and the probability of selecting a particular option is presented graphically with an option characteristic curve (OCC). This item is rated on a six-point scale with the anchors of 'strongly disapprove' and 'strongly approve'. The estimate of the latent trait is along the x-axis and the probability of selecting a particular response option at a given level of the latent trait is along the y-axis. In IRT, the latent trait is referred to as theta (i.e., θ). The values of theta are expressed as standardized scores (i.e., z-scores). Thus, an individual with θ =1.0 has a value on the latent trait that is one standard deviation above the mean. Using the first response option (i.e., strongly disapprove) as an example, the probability of selecting this option is greatest for individuals with low levels of exchange quality ($\theta < 0.0$) and the probability becomes smaller as the exchange quality increases. The sixth response option (i.e., strongly approve), on the other hand, demonstrates the reverse pattern. The probability of selecting the option is small for individuals with low levels of exchange quality and the probability increases as the level of exchange quality increases.

Samejima's (1969) graded response model is a particular cumulative graded model that is often used in organizational research (see van der Linden & Hambleton, 1997, for a discussion of other graded response models). In Samejima's (1969) graded response model (GRM), two parameters associated with the items are estimated. The first is an option difficulty parameter. The difficulty parameter is referred to as the 'threshold' parameter. This refers to the probability of an individual with a given level of the latent trait selecting a given option (e.g., disap-

prove) or any of the subsequent higher ordered options (e.g., neutral, approve, and strongly approve). Specifically, this parameter is the point on the theta scale where there is a 50% chance that a given option or a higher ordered option will be selected (i.e., P(θ)=0.50). In other words, this parameter represents the thresholds between the response options. The second parameter is the discrimination parameter. This parameter represents how well an option discriminates between individuals at different levels of the latent trait. The larger the value, the better the option is at discriminating between individuals at different levels of the latent trait.

To estimate the OCCs, one first estimates the boundary response functions. Boundary response functions are the cumulative probability of selecting a response option equal to or higher than the current response option. Option difficulty parameters are estimated for $m_i - 1$ boundary response functions where mi equals the number of response options. Each boundary response function has a difficulty parameter, but only one discrimination parameter is estimated for each item in Samejima's (1969) GRM. Therefore, on a leadership measure with six response options, five difficulty parameters and only one discrimination parameter are estimated. The boundary response functions are used to estimate the OCC. Mathematically, the boundary response function is expressed as, $P_{ik}^{*}(\theta) = e^{Da_i(\theta - b_{ik})} / 1 + e^{Da_i(\theta - b_{ik})}$, where $P_{ik}^{*}(\theta)$ is the probability of a respondent at a particular level of theta responding to option k or any of the other higher ordered options on item i, b_{ik} is the option difficulty parameter, a_i is the discrimination parameter, D is a scaling constant equal to 1.702, and e represents an exponential function. Thus, the probability of selecting option k is a function of the level of the latent trait, the difficulty of the option, and the discrimination. In essence, the boundary response functions are estimated by utilizing a two-parameter logistic IRT model on the response option data (see Hambleton et al., 1991 for a discussion of logistic models).

From the boundary response functions, the probability of selecting a particular option and the OCCs are estimated. The probability of selecting a particular option $P_{ik}^{*}(\theta)$, is determined by subtracting the boundary response functions for each option. Mathematically, this is represented as $P_{ik}(\theta) = P_{i(k-1)}^{*}(\theta) - P_{ik}^{*}(\theta)$, where the probability of selecting an option is a function of the conditional probability of responding above the threshold parameter (i.e., b_{ik}) for option $k-1$ minus the conditional probability of responding above the threshold parameter for option k. For example, on a four option item $P_{ik}(\theta)$ is as follows,

$P_{i0} = 1 - P_{i1}^{*}(\theta)$ for option 1,

$P_{i1} = P_{i1}^{*}(\theta) - P_{i2}^{*}(\theta)$ for option 2,

$P_{i2} = P_{i2}^{*}(\theta) - P_{i3}^{*}(\theta)$ for option 3,

$P_{i3} = P_{i3}^{*}(\theta) - 0$ for option 4.

The probabilities that are computed for each response option at each level of θ serve as the basis for the OCC.

3. ACKNOWLEDGMENTS

This paper was undertaken while the author was a visiting scholar at the University of California, Riverside. It was sponsored by Western Light Talent Culture Project supported by Chinese Academy of Sciences (Grant No.2009(236)).

4. REFERENCES

[1]. T. Doxey, R. Phillips. A comparison of entrance requirements for health care

professions. J Manipulative Physiol Ther1997, 20: 86-91.
[2]. D. Lawson, C. Violato, A. Marini, M. McEwen. Differential performance on the Canadian Chiropractic Examining Board Examinations: an eight year longitudinal study. J Can Chiropr Assoc 1998, 39: 11-7.
[3]. F. Lord. Applications of item response theory to practical testing problems. 1st ed. Hillsdayle (NJ)7 Lawrence Erlbaum Associates, Inc; 1980.
[4]. R. Hambleton. Applications of item response theory. Vancouver (Canada)7 Educational Research Institute of British Columbia; 1983.
[5]. J. Linacre. Sample size and item calibration stability. Rasch Meas Trans 1994, 7: 328.
[6]. P. Harasym. The use of the Rasch model to test the equivalence of two methods of standard setting. Annual Conference on Research in Medical Education 1980, 3-8.
[7]. E. Howard. Applying the Rasch model to test administration. J Nurs Educ 1985, 24:340-3.
[8]. M. Lunz, J. Stahl. The effect of rater severity on person ability measure: a Rasch model analysis. Am J Occup Ther1993;47:311-7.
[9]. M. Lunz, J. Stahl. Impact of examiners on candidate scores: an introduction to the use of multifacet Rasch model analysis for oral examinations. Teach Learn Med 1993, 5:174-81.
[10]. B. Clauser, L. Ross, R. Nungester. An evaluation of the Rasch model for equating multiple forms of a performance assessment of physicians' patient-management skills. Acad Med 1997;72(10 Suppl 1):S76-8.
[11]. G. Ingebo. Probability in the measurement of achievement: Rasch measurement. Chicago7 Mesa Press; 1997.
[12]. W. Wang. Rasch analysis of distractors in multiple-choiceitems. J Outcome Meas 1998, 2:43-65.
[13]. S. Chae. Controlling the judge variable in grading essay-type items: an application of Rasch analyses to the recruitment exam forKorean public school teachers. J Outcome Meas 1998, 2:123-41.
[14]. L. Ryser, B. Wright, A. Aeschlimann. A new look at the Western Ontario and McMaster Universities Osteoarthritis Index using Rasch analysis. Arthritis Care Res 1999, 12:331-5.
[15]. F. Wolfe, S. Kong. Rasch analysis of the Western Ontario McMaster Questionnaire (WOMAC) in 2205 patients with osteoarthritis, rheumatoid arthritis, and fibromyalgia. Ann Rheum Dis 1999, 58:563-8.
[16]. T. Sheehan, L. DeChello, R. Garcia, J. Fifield. Measuring disability: application of the Rasch model to activities of daily living (ADL/IADL). J OutcomeMeas4 2000;681-705.
[17]. M. Banerji. Construct validity of scores/measures from a developmental assessment in mathematics using classical andmany-facet Rasch measurement. J Appl Meas 2000, 1: 177-198.
[18]. E. Smith. Metric development and score reporting in Rasch measurement. J Appl Meas 2000, 1:303-306.
[19]. C. Luquet, N. Chau, F. Guillemin, et al. A method for shortening instruments using the Rasch model. Validation on a hand functional measure. Rev Epidemiol Sante Publique 2001, 49:273-86.
[20]. G. Karabatsos. The Rasch model, additive conjoint measurement, and new models of probabilistic measurement theory. J Appl Meas 2001, 2:389-423.
[21]. E. Smith. Evidence for the reliability of measures and validity of measure interpretation: a Rasch measurement perspective. J Appl Meas 2001, 2:281-311.
[22]. T. Bond, C. Fox. Applying the Rasch model: fundamental measurement in the human sciences. 1st ed. Mahwah (NJ) 7 Lawrence Erlbaum Associates, Inc; 2001.
[23]. C. Myford, E. Wolfe. Detecting and measuring rater effects using many-facet Rasch measurement: Part I. J Appl Meas2003, 4:386-422.
[24]. M. Stone. The effect of sample size for estimating Rasch/IRT parameters with dichotomous items. J Appl Meas2004, 5:48-61.
[25]. R. Gershon. Understanding Rasch measurement: computer adaptive testing. J Appl Meas 2005, 6:109-27.