Application of Logistic Regression Model in Physics Education

Shobha Kanta Lamichhane
Tribhuvan University, Prithwi Narayan Campus, Pokhara, Nepal
sklamichhane@hotmail.com

Abstract
This paper introduces a logistic regression model used to analyze multiple-choice test data of physics test. It does not involve decision making. It is a predictive model and is more akin to nonlinearity, such as fitting a polynomial to a set of data values. Brief description of the goals and algorithms of such a model is provided, together with examples illustrating their applications in physics. Priority haven been given for data interpretation rather than mathematical complexities.

Key words: logistic regression, multiple choice test items, classical test theory, item response theory, probability, evaluation.

Introduction

The logistic model computes the probability of the selected response as a function of the values of the predictor variables. If a predictor variable is categorical with two values, then one of the values is assigned the value 1 and the other is assigned the value 0. If a predictor variable is a categorical with more than two categories, then a separate dummy variable is generated to represent each of the categories except for one which is excluded. The value of the dummy variable is 1 if the variable has that category, and the value is 0 if the variable has any other category. If the variable has the value of excluded category, the entire dummy variable generated for the variables are zero.

In connection with physics education research (PER), its prime goal is to develop pedagogical techniques and strategies that will help students learning physics more effectively. Teaching and evaluation are two inseparable parts of education. Evaluation is considered to be an integral part of learning, which consists of multiple choice test items, frequently used in competitive exams. Keeping this in mind, multiple-choice tests are increasingly used in physics education to assess students understanding/learning. Appropriate and effective approaches to data analysis of multiple-choice tests thus become an important research topic. To facilitate data analysis and interpretation, physics education researchers have adopted various testing techniques from educational and psychological studies. These techniques benefited many studies published in international journal and other forms of publication.

Despite the volume of the literatures on mathematical theories of diverse testing techniques, a concise introduction to frequently encountered approaches of data analysis suitable for PER is much needed. In this paper, attempts have been made to introduce approaches to analyzing multiple-choice test data, viz: classical test theory (CTT). Specifically, the goals and basic algorithms of said approach offering examples to demonstrate data interpretation. Emphasis is placed on applications of said approach in the context of PER studies. Since it is not my intention to present comprehensive theories of statistics so that I have minimize mathematical details and avoid derivations that can be found in the listed references [1,2,3,4,5,6,7]. I also do not intend to pursue highly technical issues that are controversial even among statisticians and psycho-metricians; therefore results and discussions presented in this paper are in compliance with conventional norms that are commonly recognized in education. Other related issues not covered herein include the pros and cons of multiple-choice tests [1] various types of test validity [2] and pre/post use of multiple-choice tests to gauge the effectiveness of traditional courses and hence taken as a feedback to reform them [3]. Keeping these in mind, we choose the use of multiple choice test items for students understanding. Other logical reasons behind it are easy for strict time framing, interpretation of results in numerical form and wide range of course coverage.

Methodology

Several multiple choice test items were designed for the entrance examination of grade eleven

(immediately after the result of School Leaving Certificate, SLC, board examination) for SOS Herman Gmeiner Higher Secondary School Gandaki (one among the best in Nepal) Rambazar, Pokhara, in July, 2008. Out of those only 33 items were selected fall under general mechanics, hydrostatic and elasticity based on SLC course. Those students, who have good background of Mathematics, English, Science and have secured 70 percent or above in their SLC are considered to be eligible for entrance examination. Three hundred and eighty nine students were appeared in the examination. The exam was under full control and its duration was for two hours. In brief, overall management of the examination was quite satisfactory. The seat plan were arranged as (2×2) in each bench-desk in a row with two coulombs having the capacity of forty students invigilated by two faculty members in each room, supervised by the principal. The distribution of the examinee is according to their registration number. Sealing of the papers was made after verifying and recounting according to their attendance slip. The checking of the papers was started just after the examination till their end under the close supervision of school authorities. Listings of the marks were done immediately after the checking and sample cross checking and then entry of marks was made with the help of the computer programming. The tabulation of the raw score distribution is in tabular form given in appendix.

Results and discussion

Based on statistics, item response theory (IRT) is a modern test theory, used to estimate item characteristic parameters and examinees' latent abilities [8]. Here, item characteristic parameters include item difficulty and discrimination index, which may seem the same as those in CTT but have different meaning. Examinees' latent abilities are referred to as examinees' general knowledge, capabilities, and skills in a specific domain. IRT assumes one uni-dimensional skill or ability that underlines examinees' responses to all items. This skill or ability is considered as latent because it is a nonphysical entity and is not directly measured. For example, a student's score on mechanics multiple-choice test is only an outcome of his/her understanding of it but is not his/her understanding in physics itself. Simply, general mechanics test can be taken as the subset of the universal set (understanding SLC physics course). IRT, however, intends to provide an estimate of such un-measurable entities.

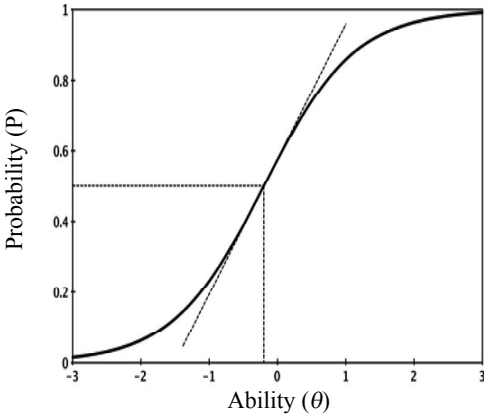

Fig.1. characteristic curve.

The basic task of IRT is to use logistic regression to formulate observed binary data. A graphical representation of this logistic regression (also known as the item characteristic curve) [9] is depicted in Fig. 1. Here, the horizontal axis represents latent ability (θ) and the vertical axis shows the probability $P(\theta)$ of answering an item correctly. Two parameters are useful in describing the shape of the curve. One is the location of the curve's middle point; the other is the slope of the curve at the middle point. The middle point is at $P(\theta) = 0.5$, and its corresponding value along the ability scale θ is defined as item difficulty. In other words, item difficulty is the ability value at a 50% probability of correct response. So, the greater the difficulty value of a particular test item, the higher an ability level is required to have a 50% probability of correct response. This is different from the difficulty measure in classical test theory. As for the slope of the curve, it has a maximum at the middle point. If the slope is large, the curve is steeper, indicating that students of high abilities have a greater probability of correct response than those of low abilities. Conversely, if the middle point slope is small, the curve is flatter; students of high abilities have nearly the same probability of correct response as those of low abilities. In this sense, the slope at the middle point is a measure of an item's discrimination. In IRT, let the item difficulty and

discrimination parameters are denoted by *b* and *a*, respectively. Using these notions, the mathematical expression of this logistic regression model is given by

$$P(\theta) = \frac{1}{1+e^{-z}}$$

Where, P(θ) is estimated probablity, $z = a + b\,\theta$, z is predictive valiable. Further, item characteristic parameters *a* and *b* belongs to the students' hidden abilities.

Now, one of the core issues in IRT is to determine the item characteristic parameters 'a' and 'b'. This is two-parameter logistic regression model. So as to evaluate the said constants, let us further discuss more about this model. The extension of this is three-parameter Birnbaum model [10] considers a guessing effect and introduces a new parameter *c*. This parameter '*c'* represents the probability of guessing a correct response for those who do not possess the necessary ability to answer it correctly. Thus, the observed probability of correct response now becomes 'c[1-P(θ)] + P(θ)'. In this model, three parameters need to be determined. To demonstrate how IRT can be a useful tool in evaluating multiple-choice items, we provide the following example using the three-parameter Birnbaum model as calculated by multilog [11]. Results are based on binary data collected from three hundred and eight students' responses to thirty three items in the SOSHGS entrance examination test. Recall that our goal is to estimate the item characteristic parameters *a*, *b*, and *c* for each of the individual items. For illustration purposes, as shown in Fig. 2, item characteristic curves of two items: 27 and 33.

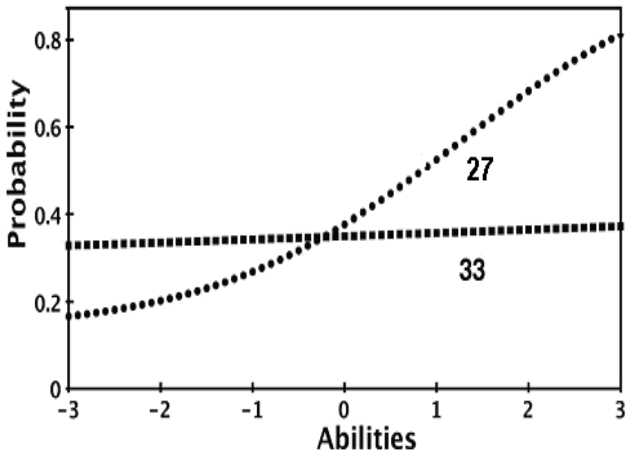

Fig .2. Item characteristic curves for two items (say, 27 and 33).

As seen, item 27 has a positive discrimination value *a*=0.74, displaying a monotonically increasing "sigmoidal shape" curve in the range of θ ∈ [–3, +3]. The value along the θscale for the curve middle point is 1.25, meaning is its item difficulty is *b*=1.25. Simply put, students whose ability value is 1.25 have a 50% probability of correctly answering this item. The lower left part of the curve shows an asymptote of 0.13, indicating that students of low abilities have a 13% probability of guessing this item correctly. As opposed to item 27, item 33 displays nearly a flat line (*a*=0.04) in the θ ∈ [–3, +3] range, indicating the item fails to distinguish students of high abilities from those of low abilities. The reason for this may be due to the high difficulty level of this item (*b* =22.9).

In the above example, the ability scale can be generally described as students' knowledge of energy topics in mechanics. The reason is twofold. First, IRT assumes a uni-dimensional scale for the entire test. Second, the exam solely focuses on topics in mechanics that are covered in the SLC mechanics course. Of course, here is that IRT estimated abilities may be correlated with, but are not identical to, test total scores. A total score may be dependent on the specific questions used in a test, whereas IRT-estimated abilities are independent of the questions used in a test. For an elaborated proof, refer to Ref [11]. Similarly, the item difficulty and discrimination parameters in IRT are also independent of examinees who take the test.

In addition to the above application, IRT can be used to evaluate the functions of distracters in each item. The basic idea is to examine trace lines for alternative choices. As an example, we plot in Fig. 3 alternative-choice trace lines for one item (say, 30) in the SOSHGS entrance test using some other model [12,13]. In this example, the correct choice (choice b) displays a monotonically increasing 'S' curve in the θ ∈ [–3, +3] range. Therefore, students of high abilities are more likely to choose the correct answer than those of low abilities. As for choices 'a' and 'c', the trace lines have a reverse trend. So, students of low abilities are more likely to select choice 'a' or 'c' than those of high abilities. Take choice 'c' for

example; the probability of choosing this answer is less than one or two percent for students of an ability value of +3, but it is as more than fourty percent for those of an ability value of −3. As for choice'd', the trace line is relatively flat and low, suggesting that not many students choose this answer at any ability level. Therefore, alternative choices 'a' and 'c' seem to function better than choice'd' in distracting students of low abilities.

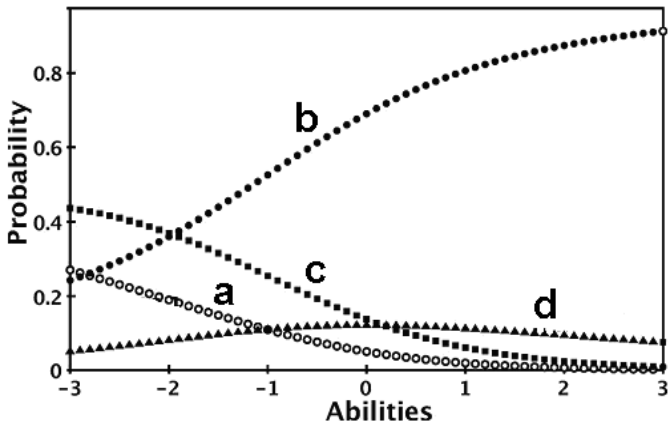

Fig.3. Alternative choices of a single item.

In Fig. 3, we once again plot probabilities against latent abilities, not test total scores. In fact, it is much easier to use total scores as a substitute for abilities than to perform IRT. This is particularly true when one lacks adequate knowledge of IRT. As a rudimentary step toward IRT using total scores as a substitute for θ can provide a glimpse of what real IRT curves may look like. A recent study by Morris *et al.* used this approach to evaluate force concept inventory items [14]. Conversely, a paper by Lee *et al.* employed two-parameter IRT to measure student latent abilities in physics [15].

Finally, some practical issues of IRT are worth noting. First, a large sample size generally is recommended for a good model fit. Since the Rasch model estimates fewest parameters, a data set of as few as 100 may be needed for stable results [16]. (Linacre [17] suggested 50 for the simplest Rasch model.). For other models, a sample size of several hundred often is required. Also, different study purposes may call for different sample sizes. For example, calibration of high-stake test items may require sample sizes over 500 to ensure accuracy. But for low-stake tests, "one does not need large sample sizes." [18]. Secondly, there have been great controversies on IRT model selections. Though the three-parameter model seems to be the most complicated and hence the most stringent model, arguments have been made that it in fact is the most general model and that the other two models (the Rasch and two-parameter models) are just special cases of the three parameter model [19]. Therefore, it is recommended that the three-parameter should be used [20]. On the other hand, the seemingly simple expression of the Rasch model continues to attract many researchers. Thorough discussions on these issues, interested readers can refer to ref [21] for more information.

Conclusion

In this paper, discussions have been made to the goals, basic algorithms, and applications of analyzing multiple-choice test items in physics. Using logistic regression principle, one can describe the propensity of correct responses to the individual items. As a result, it is evident from above discussion that estimated item measure and examinees' abilities are mutually independent. Moreover, IRT can also use to examine how effective the choices of a particular item/question. Because of its emphasis on individual items, theory is better named as "item response theory." Our ultimate goal is to make sense of raw data; therefore, we choose four items, encountering two (or more) equally sound choices. While analyzing the obtained raw data, one should always prefer the one that better facilitates the graphical data interpretation. In our case choosing an item have equal probability. For the purposeful meaning of data interpretation, fitting a curve should be probabilistic, and hence logistic regression model is chosen which follows probability distribution.

Appendix

Table: Indics for item analysis and studens' score distribution

Item	Difficulty index	Discrimination index	Point biserial coefficient
1	0.24	0.4	0.39
2	0.65	0.57	0.46
3	0.39	0.64	0.52
4	0.88	0.22	0.24
5	0.76	0.25	0.25
6	0.79	0.14	0.18
7	0.31	0.57	0.50
8	0.26	0.41	0.37
9	0.82	0.28	0.26
10	0.52	0.45	0.37
11	0.83	0.27	0.26
12	0.72	0.5	0.39
13	0.65	0.59	0.42
14	0.79	0.42	0.32
15	0.43	0.62	0.47
16	0.32	0.47	0.38
17	0.73	0.29	0.23
18	0.34	0.39	0.35
19	0.65	0.30	0.30
20	0.66	0.27	0.25
21	0.33	0.3	0.26
22	0.37	0.39	0.27
23	0.43	0.41	0.31
24	0.13	0.10	0.15
25	0.80	0.37	0.30
26	0.59	0.26	0.16
27	0.39	0.56	0.47
28	0.36	0.18	0.18
29	0.67	0.41	0.34
30	0.67	0.45	0.36
31	0.33	0.32	0.28
32	0.29	0.54	0.43
33	0.32	0.33	0.29

Total score	No. of students	Total score	No. of students
0	0	17	39
1	0	18	27
2	0	19	25
3	0	20	17
4	1	21	25
5	0	22	14
6	1	23	14
7	2	24	15
8	2	25	10
9	4	26	3
10	13	27	7
11	17	28	6
12	17	29	4
13	25	30	3
14	39	31	0
15	29	32	1
16	29	33	0

References

1. R. Lissitz and K. Samuelsen, 2007. Suggested changes in terminology and emphasis regarding validity and education, *Educ. Res*. **36:** 437.
2. S. Ramlo, 2008.Validity and reliability of the force and motion conceptual evaluation, *Am. J. Phys*. **76:**882.
3. http://www.physics.indiana.edu/~hake/MeasChangeS.pdf
4. T. Kline, 2005. *Psychological Testing: A Practical Approach to Design and Evaluation, SAGE*, Thousand Oaks, CA, pp. 91.
5. http://eric.ed.gov/ ERICDocs/data/ericdocs2sql/content_storage_01/0000019b/80/ 38/90/26.pdf.
6. R. Doran, 1980. *Basic Measurement and Evaluation of Science Instruction,* NSTA, Washington, DC, p. 97.
7. A. Oosterhof, 2001. *Classroom Applications of Educational Measurement*, Merrill, Saddle River, NJ pp. 176.
8. F. Baker, 2001. *The Basics of Item Response Theory*, 2nd ed. _ERIC, College Park, MD,
9. F. Baker and S. Kim, 2004. *Item Response Theory: Parameter Estimation Techniques*, 2nd ed., Dekker, NY.

10. F. Baker, 2001. *The Basics of Item Response Theory*, 2nd ed. _ERIC, College Park, MD, pp. 21–45.
11. http://eric.ed.gov/ERICDocs/data/ ericdocs2sql/ content_storage_01/0000019b/80/19/5b/97.pdf.
12. R. Bock, 1972. Estimating item parameters and latent ability when responses are scored in two or more nominal categories, *Psychometrika* **37:** 29.
13. F. Samejima, 1979. University of Tennessee Research Report No. 79.
15. G. Morris, L. Branum-Martin, N. Harshman, S. Baker, E. Mazur, S. Dutta, T. Mzoughi, and V. McCauley, 2006. Testing the test: Item response curves and test quality, *Am. J. Phys*. **74:** 449.
16. Y. Lee, D. Palazzo, R. Warnakulasooriya, and D. Pritchard, 2008. Measuring student learning with item response theory, *Phys. Rev. ST Phys. Educ. Res*. **4:** 010102.
17. B. Reeve and P. Fayers, 2005. *Assessing Quality of Life in Clinical Trials: Methods and Practice*, 2nd ed., edited by P. Fayers and R. Hays, Oxford University Press, Oxford, pp. 55.
18. M. Linacre, 1994. Sample size and item calibration stability, *Rasch Measurement Transactions* **7:** 328.
19. L. McLeod, K. Swygert, and D. Thissen, 2001. *Test Scoring*, edited by D. Thissen and H. Wainer, Erlbaum, Mahwah, NJ, p. 189.
20. X. Fan, 1998. Item response theory and classical test theory: An empirical comparison of their item/person statistics, *Educ. Psychol. Meas*. **58:** 357.
21. D. Harris, 1989.Comparison of 1-, 2-, and 3-parameter IRT modes, *J. Educ. Meas*. **8:** 35.

❑❑❑