Proc. of the Third Intl. Conf. on Advances in Management, Economics and Social Science - MES 2015.

ISBN: 978-1-63248-081-1 doi: 10.15224/ 978-1-63248-081-1-81

The Construct Comparability Approach: An Empiral Study in Spain

[Alejandro Veas, Raquel Gilar, Pablo Miñano, Juan Luis Castejón]

***Abstract*—The present study aims to perform a comparative analysis of 6 courses using the Rasch model with a sample of 1398 students (*M* = 12.5, *SD* = 0.67) from 8 schools in the province of Alicante (Spain). The fist data analysis showed a misfit of the model. After recoding the categories (from 9 to 4), a good fit was observed in most of the courses. The differential item functioning (DIF) by gender was significant in the visual arts education course. The measurement of a single latent construct is confirmed in this study.**

Keywords—educational evaluation, construct comparability approach, Rasch model.

I. Introduction

The evaluation processes in Spain are based on conducting non-standaridized written test, as well as the assessment of attitudinal variables observed in the classroom. The curriculum employeed let teacher to know the evaluation criteria in each of the academic courses, taking a final grade that serves as an indicator of the academic achievement process [1], [2]. This kind of assessment has been improved with studies at the international level, such as the Trends in International Mathematics and Science Study (TIMSS), International Assessment of Literacy Survey (IALS) and the Programme for International Student Assessment (PISA), in where standardized tests are used to measure academic performance [3], [4].

Standardized achievement tests provide objective, reliable measures, with greater use in the field of educational evaluation on a large scale. Despite considering academic grades and test performance as complementary [5], it is possible that test performance is not an appropriate measure for current student performance [6]. However, there are no studies in Spain which demonstrate the adequacy or inadequacy of the use of academic grades for analyzing performance at the local or national level, although traditionally grades have been conceived as good indicators. Furthermore, academic grades are used in University Entrance Examination to access to higher education. In this sense, it is necessary to ask the following question: Are grades for 2 different courses comparable? Moreover, are the grades obtained in various courses actually good indicators of academic performance? In the present study, the comparison of academic grades is observed under the construct comparability approach, which is based on the application of the Rasch model.

The construct comparability approach is a new conceptualization of the term "comparison", based on the integration of the performance approach and the statistical approach [7]. This model indicates that when comparing 2 elements, whatever they may be, they must have something in common that serves as the basis for this comparison. Just as 2 tests can be compared based on their measurement using the same scale, in the context of comparing academic grades, we can only compare those that measure a shared construct, in our case, academic performance. The premise of this approach would be as follows [8]: Two grades from 2 students are comparable if the performance of both students, which corresponds to the same level of the latent construct that they share, leads to the same grade.

According to this postulate, the difficulty of a course will correspond to a specific level established in the latent variable; that is, a course will be more difficult than another to the extent that, to achieve a higher grade, a higher level of performance or ability is needed. If the latent construct is changed, this relationship may easily be the inverse [9].

The theoretical approach presented has its functionality through the Rasch model [10] [11]. This model establishes that the difficulty of the items and the ability of the subjects can be measured on the same scale, and the likelihood that a subject responds correctly to an item is based on the difference between the ability of the subject and the difficulty of the item. Both measures (ability and difficulty) are estimated using logit units because the scale used by the model is logarithmic. Using the same measurement scale establishes homogeneous intervals, which means that the same difference between the difficulty parameter of an item and the ability of a subject involves the same probability of success along the entire scale [12]. It is important to highlight that, in this case, data must fit the model to be accepted. This adjustment can be conducted using residual measures, and can be standardized for a particular item or subject in 2 ways [13]:

- Outfit: This is the root mean square of the residuals, divided by the degrees of freedom. This measure can be interpreted as an overall measure that expresses whether the answers given to a particular item will fit the model.
- Infit: This measure eliminates the extreme scores that influence the outfit in such a way that uses the residuals of the individuals whose ability levels are in the closest range to the particular item.

Statistical *infit* and *outfit* are calculated based on root mean squares, depending on the statistical value of

Alejandro Veas, Raquel Gilar, Pablo Miñano, Juan Luis Castejón
University of Alicante
Spain

Raquel Gilar
University of Alicante. Dep. of developmental psychology and didactic
Spain

Pablo Miñano
University of Alicante. Dep. of developmental psychology and didactic
Spain

Juan Luis Castejón
University of Alicante. Dep.of developmental psychology and didactic
Spain

Proc. of the Third Intl. Conf. on Advances in Management, Economics and Social Science - MES 2015.

ISBN: 978-1-63248-081-1 doi: 10.15224/ 978-1-63248-081-1-81

Pearson's chi-squared divided by the degrees of freedom, thus forming a scale with values that can range from 0 to infinity. Values below 1 indicate a higher than expected fit of the model, while values greater than 1 indicate a poor fit of the model. Thus, if we have an *infit* value of 1.40, then we can assert that there is 40% more data variability compared to the model's prediction; while an *outfit* of 0.80 indicates that 20% less data variability is observed with respect to the model's prediction.

We start from the consideration of each of the courses as a specific item, with the range of grades from 1 to 10, which implies various degrees of success. The partial credit model [14] enables an analysis of the difficulty of achieving a specific score for each of the courses separately, following the Rasch methodology. The formula of the model is as follows:

$$\text{Ln}\,(P_{nij} / P_{ni(j-1)}) = B_n - D_i F_{ij} = B_n - D_{ij} \quad (1)$$

where:

P_{nij} is the probability of subject n responding correctly to item i observed in category j;

B_n is the measured ability of subject n;

D_i is the measured difficulty of item i; and

F_{ij} is the calibration measured for item i in category j compared to category $j-1$, the point at which categories $j-1$ and j are equally likely compared to the measurement of the item.

According to this model, the present study proposes to analyze whether different courses are able to measure the same latent construct.

II. Method

A. *Sample*

The cluster sampling technique was used, with the school as the sampling unit. A total of 8 schools in the province of Alicante were included; 2 schools were private, while the reminder were public. A total of 1456 students in the first and second year of Compulsory Secondary Education participated in the study. Of these, 58 were excluded due to coding errors or the lack of qualifications, leading to a total of 1398 subjects (n = 1398). A total of 53% of the students were mal, and 47 % were women, with an average age of 12.5 years and a standard deviation of 0.67. A total of 1137 participants (81.4) were enrolled in a public school, while 261 (18.6%) were enrolled in a private school. A total of 52.4% of students were from the first grade of E.S.O., and 47.6% were from the second grade of E.S.O.

B. *Measures*

For the present analysis, General Points Average (GPA´s) from 6 courses, which teachers provided at the end of the school year, were considered. The courses recorded were Spanish language and literature, natural sciences, social science, mathematics, English and physical education. Students scores showed high reliability, with a Cronbach´s alpha of 0.93 for the first grade and 0.94 for the second grade.

C. *Procedure*

In the first place, permissions were requested from the educational administration and school boards of the various schools. After obtaining the permits, the parents or legal guardians of the students had to provide the corresponding informed consent. Data collection was performed in the schools themselves during the second trimester of the 2011-2012 school year and during normal school hours. Collaborating researchers were previously trained in the standards and guidelines for data collection. Students and parents participated voluntarily, and the parents signed and informed consent form that ensured data confidentiality at all times.

D. *Data Analysis*

The partial credit model was used with WINSTEPS version 3.81 statistical software [15], whose estimates were based on the joint maximum likelihood [16] [17].

III. Results

Table 1 shows the courses analysed, the indices of fit, and the item-scale correlation. We used an approximate range of 0.8-1.2 for infit and outfit [13], in addition to the observation of each of the ítem characteristic curves (ICCs). In Table 1, it is observed a lack of fit in a number of courses, which assumes a lack of fit for the subjects´ pattern of responses with respect to the model. Furthermore, in the ICCs we see categories whose highest response probabilities are exceeded by adjacent categories, especially the lowest categories. This situation implies that the pattern of responses does not adequately fit the model and that reconversion of the performance categories for all courses may be appropriate.

TABLE 1: STATISTICS OF FIT FOR THE COURSES IN THE FIRST AND SECOND GRADES IN ESO.

Courses	**Count**	***Infit***	***Outfit***
Spanish language and literature	1396	.62	.63
Natural sciences	1398	.60	.60
Social sciences	1394	.88	.86
Mathematics	1391	.94	.93
English	1397	1.16	1.13
Physical Education	1392	1.53	1.87

Based on the qualitative scores of schools, recoding was performed using the following values: 1 for categories 1, 2, 3, and 4 (“poor”); 2 for categories 5 and 6 (“sufficient” and

Proc. of the Third Intl. Conf. on Advances in Management, Economics and Social Science - MES 2015.

ISBN: 978-1-63248-081-1 doi: 10.15224/ 978-1-63248-081-1-81

"good"); 3 for categories 7 and 8 ("notable"); and 4 for categories 9 and 10 ("outstanding").

The new calibration of the courses provided a better fit for the data (see Table 2), except for physical education (*Infit* = 1.43; *Outfit* = 1.52). The analysis of Differential Item Functioning (*DIF*) estimated the distribution of the difficulty parameter in the sample of boys and girls in each course. The differences found in the courses were not statistically significant, with $p > 0.001$. Thus, physical education was removed to estimate the new model. As shown in Table 2, the difficulty indexes for the courses are in the 0.8-1.2 range.

For the analysis of unidimensionality, a principal components analysis of the residual scores was conducted [18]. The results showed a principal factor that was able to explain 69.3% of the variance of the latent trait, with a wide difference between the weight of the first factor and the next (*Eigenvalue* = 1.4), which favours the unidimensionality of the model.

Although not shown, the ICCs fit the model, which means that the greater the ability of the subjects is, the greater the probability of obtaining a high grade in the course.

TABLE 2. INDICES OF FIT FOR FIRST AND SECOND GRADES OF ESO WITH RECODED VALUES

Courses	Count	Infit	Outfit
Spanish Language and Literature	1396	0.75	0.79
Natural Sciences	1398	0.75	0.75
Social Sciences	1394	0.83	0.83
Mathematics	1391	0.94	0.99
English	1397	1.03	1.04

IV. Discussion

In the educational contexts, academic grades have been used frequently in a lot of empirical studies. In this article, based on the theoretical aspects of the concept of construct comparability, and empirical study is conducted that uses academic grades and the Rasch model.

In light of the results, we may assert that the hypothesis is partially confirmed. It was necessary to reduce the number of categories for all courses and eliminate the physical education course to obtain adequate levels of fit [19], [20]. Furthermore, no DIF was observed in the analyzed courses.

With the adjustment of categories, it can be assumed that all of the courses aim at measuring overall academic performance, showing optimal values of factor loadings in the principal component analysis, confirming the unidimensionality of the construct.

The partial credit model is employed in the calculation of the difficulty indices for each course, which allows us to know the ability level required by the subject to achieve a certain grade. Family Rasch measures has been widely used in educational studies, as it is an effective tool of analysis. However, it is important to consider the numerous cognitive and non-cognitive variables that could affect academic results [21], [22].

According to the results, Spanish category grades are strongly recommended to be reduced , especially in the lowest categories. In the present study, it was found that in all schools analysed, the grades 1,2 and 3 are assigned in a very low proportion in all courses. In addition, a wider range of grades leads to a more heterogeneous distribution of evaluation criteria than the standards indicate. In this regard, schools in countries such as the United Kingdom used smaller grade ranges [23].

The completion of this study at the provincial level has enabled us to approximate construct comparability analysis, confirming the possibility of conducting future studies analogous to the English studies [8], i.e., using the scores of participants in the PAU exams or the future final assessment tests that will be taken in Primary Education and Compulsory Secondary Education. Thus, national data may be obtained that would enable a better comparison of the courses and their relative weight on each of the exams and conclusions concerning the implications of the use of grades in the selection process for students in higher education.

Acknowledgment

The present work was supported by the Vice Chancellor of Research of the University of Alicante under Grant GRE 11-15, and by the Spanish Ministry of Economy and Competitiveness under Grant EDU2012-32156.

First author thanks to the Ministry of Economy and Competitiveness due to a lecturer grant BES-2013-064331.

References

[1] P. Miñano, J.L. Castejón, "Variables cognitivas y motivacionales en el rendimiento académico en lengua y matemáticas: Un modelo estructural", Revista de Psicodidáctica, vol.16(2), pp. 203-230, 2011.

[2] M. Gutiérrez, E. López, "Motivation, students´ behaviour and academic achievement", Infancia y aprendizaje: Journal for the study of education and development, vol.35(1), pp. 61-72.

[3] J. Calero, A. Choi, S. Waisgrais, "Determinantes del riesgo del fracaso escolar en España: Una aproximación a través de un análisis logístico multinivel aplicado a PISA[Determinants for the risk of school failure in Spain: An approach using a multilevel logistic analysis applied to PISA]", Revista de Educación, vol. Extra (1), pp. 225-256, 2010.

[4] J.M.C. Ferrera, C.M. López, R.S.Rodríguez, "Análisis de los condicionantes del rendimiento educativo de los alumnos españoles en PISA 2009 mediante técnicas multinivel[Analysis of the constraints of educational performance of Spanish students in PISA 2009 using multilevel techniques]", Presupuesto y Gasto Público, vol.67,pp.71-96, 2012.

[5] H. Marreno, M. Orlando, "Evaluación comparativa del poder predictor de las aptitudes sobre notas escolares y pruebas objetivas [Comparative evaluation of the predictive power of skills on school grades and objective tests.]", Revista de Educación, vol. 287, pp.97-112, 1988.

[6] B.D.McCoach, D.DelSiegle, "Underachievers",in Encyclopedia of Adolescence, R.J.Levesque (Ed.),New York:Springuer Science & Business Media, 2014, pp.3025-3032.

[7] P.E.Newton,"Examination standards and the limits of linking", Asessment in Education,vol.12(2),pp.105-123, 2005.

ISBN: 978-1-63248-081-1 doi: 10.15224/ 978-1-63248-081-1-81

[8] R. Coe, "Comparability of GCSE examinations in different subjects: An application of the Rasch model", Oxford Review of Education, vol.34(5), pp.609-636,2008.

[9] R. Coe, "Understading comparability of examination standards", Research Papers, vol.25(3), pp.271-284, 2010.

[10] G. Rasch, "Probabilistic models for intelligence and attainment tests", Copenhagen, Danish Institute for Educational Research, Chicago: The university of Chicago Press, expanded edition, 1980.

[11] B.D. Wright, G.N. Masters, "The measurement of knowledge and attitude", Research memorandum No.30, Chicago: MESA Psychometric Laboratory.

[12] P. Preece, "Equal-interval measurement: The foundation of quantitative educational research", Research Papers in Education, vol.17(4), pp.363-372, 2010.

[13] T.Bond,C.M.Fox, "Applying the Rasch model: Fundamental measurement in the human sciences", New York: Psychology Press, 2007.

[14] B.D. Wrigh, G.N. Masters, "Rating scale analysis. Rasch measurement", Chicago: MESA Press, 1982.

[15] J. Linacre, "WINSTEPS Rasch measurement computer program [computer software]", Chicago: Winsteps, 2011.

[16] T. Bond, "Validity and assessment: A Rasch measureement perspective", Metodología de las Ciencias del Comportamiento, vol.5, pp.179-194, 2004.

[17] J. Linacre, "A user´s guide to WINSTEPS & MINISTEP Rasch-Model Computer Programs", Program Manual 3.74.0.2012. Access in October 2014, from http://www.winsteps.com/winman

[18] J. Linacre, "Structure in Rasch residuals: Why principal component analysis? Rasch Measurement Transactions, vol12, p.636, 1998.

[19] B:D.Wright, "Despair and hope for educational measurement", Contemporary Education Review", vol.3(1), pp. 281-288, 1984.

[20] B.D.Wright, J.M.Linacre,J.Gustafson, P.Martin-Lof, "Reasonable mean-square fit values", Rasch Measurement Transactions, vol.8(3), p. 370, 1994.

[21] W.Cavendish, "Student perceptions of school efforts to facilitate student involvement, school commitment, self-determination, and high school graduation", Social Psychology of Education, vol.16(2), pp.257-275, 2013.

[22] J. Green, G.A.D. Liem, A.J.Martin,S.Colmar,H.W.Marsh,D.McInerney, "Academic motivation, self-concept, engagement, and performance in high school: Key processes from a longitudinal perspective",Journal of Adolescence, vol.35(5),pp.1111-1122, 2012.

[23] Department for Education. National Curriculum and Assessment: information for schools, "Curriculum and qualifications and schools, colleges and children´s services. Government of the United Kingdom". Access in April 2015 from http://www.gov.uk/goverment/uploads/system/uploads/attachment_data/file/358070/NC_assessment_quals_factsheet_Sept_update.pdf

About Author (s):

Alejandro Veas was born on June 27, 1987 in Murcia, Spain. He entered university in 2005 and received his bachelor degree in Psychology from University of Murcia (UMU) in 2010, majored in clinical psychology. In 2011 he obtained a master of secondary education teacher in the same institution. He is currently a PhD student and research trainee at the University of Alicante, Spain.

He previously worked as a teacher in two high schools and as a researcher assistant at the Meta-analysis Unit at the Department of Basic Psychology and Methodology and at the Department of Didactics and School Organization, University of Murcia. His research interests are focused on academic achievement, underachievement, high abilities, emotional intelligence and research methodology.