



Innovation and Society 2013 Conference, IES 2013

Objective measurements of student satisfaction by comparing the effects of different factors

Ida Camminatiello^{a,*}, Tullio Menini^b, Michele Gallo^b

^aSecond University of Naples, Capua, Italia

^bUniversity of Napoli L'Orientale, Napoli, Italia

Abstract

Item Response Theory (IRT) is proposed in literature for evaluating the customer satisfaction in private and public service. In this work a particular version of IRT, known as Partial Credit Model (PCM), is proposed to analyse the satisfaction of university students of a middle university of Southern Italy. Moreover, to consider the effects of different factors (degree course, departments, presence of integrative teaching activities, etc) and their interactions, factorial differential item functioning proposed for the Rasch model is extended to PCM.

© 2014 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/3.0/>).

Selection and peer-review under responsibility of the Organizing Committee of IES 2013.

Keywords: Item response theory; differential item functioning; student satisfaction.

1. Introduction

In order to improve Italian educational services, ANVUR (National Agency for the Evaluation of Universities and Research Institutes) has implemented a large program for evaluating the quality of teaching. In this program, seven kinds of questionnaires are used to assess the several aspects of educational services. All the questionnaires are composed of a set of items where the responses to the items are scored for ordered categories to indicate increasing

* Corresponding author. Tel.: +393492516073; fax: +390823274042.

E-mail address: camminat@unina.it

levels of a response (Likert scale). The principal approach to analyze these assessments is to transform ordinal data in pseudo-interval by scaling procedures, so traditional statistical methods can be used to analyze them. Other methods regard the difficulty of each item as information to be incorporated in scaling items, just like Item Response Theory - IRT or the Rasch Model - RM (Rasch, 1960; Bond and Fox, 2001; Baker and Kim, 2004). The latter is more realistic when a set of items is used to evaluate the student satisfaction, thus IRT is proposed in this paper.

Sometimes, the results given by IRT are not reliable, for example when an inadequately calibrated questionnaire is used or a mixture of individuals only apparently belonging to the same population are considered (Camminatiello et al., 2010). The latter case can be a symptom of a different functioning of the items corresponding to distinct groups of individuals: this phenomenon is called Differential Item Functioning (DIF). In literature there are several DIF diagnostics (Glas and Verhelst, 1995), but the most used is based on the comparison of the item difficulties among the subgroups identified by one or more aggregation variables (Wu et al. 1998). Other authors (Wang, 2000) treat the DIF as a parameter to be estimated rather than simply a sign to be detected. By adopting the formulation of factorial analysis of variance, DIF is parameterized and partitioned into main and interaction effects. DIF parameters and the item difficulties for all items are simultaneously estimated.

To analyze the satisfaction of university students of a middle university of Southern Italy we propose to extend the formulation of factorial analysis of variance to Partial Credit Model (PCM). Moreover, the nature of problem necessitates of generalizing the approach to more than two levels.

2. Item Response Theory.

The aim of the IRT is to establish the position of the individual along a not directly observable dimension called latent trait. Because of the many educational applications the latent trait is often called ability. The IRT derives the probability from each response as a function of the latent trait and some item parameters. The same model is then used to obtain the likelihood of ability as a function of the actually observed responses and, again, the item parameters. The ability value that has the highest likelihood becomes the ability estimate.

The simplest IRT model is the RM defined for dichotomous items. When the test consists of polytomous items, which present the same set of alternatives, but the relative difficulties of the steps between categories change from item to item, the PCM, proposed by Wright and Masters (1997), is the more appropriate version of the Rasch models.

Let there be one factor with G levels, indexed as $g = 1, 2, \dots, G$. Through the PCM the measure of satisfaction for the group g is achieved by

$$\text{logit}_g = \ln \left(\frac{P_{nim}}{1 - P_{nim}} \right)_g = \theta_n - \delta_{img} - \gamma_m \quad (1)$$

where P_{nim} is the probability that the student n scoring in category m of the item i ; $(1 - P_{nim})$ is the probability that the student n having chosen the category $m - 1$ for the item i , does not reach the category; θ_n is the ability of student n , δ_{img} is the difficulty of item i for the threshold m and group g , γ_m is the difficulty of category threshold m .

Let there be two factors: factor A , indexed as $g = 1, 2, \dots, G$ and B , indexed as $h = 1, 2, \dots, H$. The measure of satisfaction for the group gh is

$$\text{logit}_{gh} = \ln \left(\frac{P_{nim}}{1 - P_{nim}} \right)_{gh} = \theta_n - \delta_{img} - \gamma_m \quad (2)$$

where δ_{img} is the difficulty of item i for the threshold m and the group gh .

2.1. Detection of Differential Item Functioning

Methods for detection of differential item functioning define DIF as the presence of differences in the probability of a correct response for two or more groups, conditional on the latent trait (Maij-de Meij, 2010). In literature, a number of methods have been proposed to examine DIF. The conventional one-factor DIF analysis estimates item difficulties one group at a time and then tests their differences as follows

$$t_i = \frac{\hat{\delta}_{ig} - \hat{\delta}_{ig'}}{\sqrt{\text{var}(\hat{\delta}_{ig}) + \text{var}(\hat{\delta}_{ig'})}} \tag{3}$$

where $\hat{\delta}_{ig}$ and $\hat{\delta}_{ig'}$ are maximum likelihood estimates of item i 's difficulty for the group g and g' respectively; $\text{var}(\hat{\delta}_{ig})$ and $\text{var}(\hat{\delta}_{ig'})$ are their estimated error variances.

The squared t-statistic values for each group can be summarized as chi-square statistics for each item, indicating whether the observed DIF within each item is due to chance alone. The null hypothesis is that the DIF is statistically zero across the classes.

Conventional DIF analysis is usually based on one factor at a time ignoring potential interactions between the factors. To detect DIF when there is more than one factor, one might both form a new factor by crossing the original factors and do conventional one-factor DIF analysis. With this procedure, the meaning of the new factor is vague since it is artificial. For this reason Wang (2000) proposes the factorial DIF analysis.

In presence of one factor, the item parameters can be reparameterized as $\delta_{ig} = \delta_i + \alpha_{ig}$ given the usual restriction in ANOVA: $\sum_g \alpha_{ig} = 0$, where δ_i is the difficulty of item i , α_{ig} is DIF parameter which depicts the effect of group g on item i . The PCM becomes:

$$\text{logit}_g = \theta_n - (\delta_i + \alpha_{ig}) - \gamma_m \tag{4}$$

With two factors, the item parameters can be reparameterized as $\delta_{igh} = \delta_i + \alpha_{ig} + \beta_{ih} + \alpha\beta_{igh}$ given the usual restriction in ANOVA: $\sum_g \alpha_{ig} = \sum_h \beta_{ih} = \sum_g \alpha\beta_{igh} = \sum_h \alpha\beta_{igh} = 0$, where β_{ih} is the effect of group h on item i and $\alpha\beta_{igh}$ is the DIF interaction effect of Factor A_g by Factor B_h , on item i . The model changes into:

$$\text{logit}_g = \theta_n - (\delta_i + \alpha_{ig} + \beta_{ih} + \alpha\beta_{igh}) - \gamma_m \tag{5}$$

where a likelihood ratio test can be used to study which parameter is significantly different from zero. In this way, it is possible to recognize items that yield both the DIF main effects and the interaction effects, items that yield only the interaction effects, and items that yield factor A_g or B_h main effects.

3. Data analysis

Our database includes information on nearly 12,000 questionnaires collected on three departments from twenty-seven degree courses of the University of Naples - "L'Orientale". The questionnaire implemented by ANVUR is divided into 15 items (labelling from D01 to D15), which concerns 5 dimensions: 'organization of degree course', 'learning organization', 'teaching activities', 'facility and learning interest'. All the items have four response categories (1 - negative, 4 - positive).

When all items present the same set of alternatives, it seems reasonable to expect that the relative difficulties of the steps between categories will not vary from item to item (Gallo, 2011), and, for this reason, initially, we applied the

Rating Scale Model (Andrich, 1978; Wright and Masters, 1982) by using the statistical software ConQuest (Wu et al., 1998). After 40 iterations, the model has a deviance of 85,641.49 with 18 parameters. The deviance is a statistic that indicates how well the item response model fits the data. Unfortunately, the step parameter is outside the range proposed by Linacre, from 0.6 to 1.4, for INFIT and OUTFIT values (see Fig. 1), therefore we decided to use the PCM. This last one has a deviance of 83,334.32 with 46 parameters. A formal statistical test of the relative fit of RSM and PCM can be undertaken by comparing the deviance of the two models. Examining the deviance, note that RSM deviance is 2,307.18 greater than the deviance for the PCM. If this value is compared to a chi-squared distribution with 28 degrees of freedom, it can be concluded that the fit of the RSM is significantly worse than the fit of the PCM.

The principal results of the PCM are given in figure 2, where it is possible to observe that the separation reliability, equal to 0.996, indicates how well the item parameters are separated: it has a maximum of one and a minimum of zero, for more details see Wright and Stone (1979). Only the item 12, in bold in figure 2, does not fit well. This item concerns the adequacy of the classrooms where lessons take place. The null hypothesis for the chi-square test is equality of the set of parameters. In this case equality of all of the parameters is rejected because the chi-square is significant.

VARIABLES			UNWEIGHTED FIT			WEIGHTED FIT		
category	ESTIMATE	ERROR	MNSQ	CI	T	MNSQ	CI	T
0			2.96 (0.95, 1.05)		49.1	1.64 (0.92, 1.08)		13.2
1	-1.039	0.021	1.39 (0.95, 1.05)		13.0	1.49 (0.94, 1.06)		13.8
2	-0.605	0.017	1.93 (0.95, 1.05)		27.7	1.98 (0.95, 1.05)		29.4
3	1.644*		1.64 (0.95, 1.05)		20.2	1.59 (0.95, 1.05)		18.1

An asterisk next to a parameter estimate indicates that it is constrained

Fig. 1. Rating scale model: step

VARIABLES		UNWEIGHTED FIT			WEIGHTED FIT				
item		ESTIMATE	ERROR	MNSQ	CI	T	MNSQ	CI	T
1	D01	0.404	0.030	1.01 (0.95, 1.05)		0.2	1.00 (0.95, 1.05)		0.0
2	D02	0.693	0.027	1.13 (0.95, 1.05)		4.6	1.11 (0.95, 1.05)		4.2
3	D03	-0.021	0.031	0.96 (0.95, 1.05)		-1.4	0.97 (0.95, 1.05)		-1.3
4	D04	-0.794	0.045	1.07 (0.95, 1.05)		2.7	1.04 (0.94, 1.06)		1.1
5	D05	-0.896	0.052	0.83 (0.95, 1.05)		-6.7	0.89 (0.94, 1.06)		-3.7
6	D06	0.219	0.032	1.21 (0.95, 1.05)		7.4	1.20 (0.95, 1.05)		7.2
7	D07	-0.183	0.033	0.75 (0.95, 1.05)		-10.4	0.78 (0.94, 1.06)		-8.4
8	D08	-0.405	0.036	0.75 (0.95, 1.05)		-10.4	0.79 (0.94, 1.06)		-7.6
9	D09	0.118	0.031	0.97 (0.95, 1.05)		-1.2	0.97 (0.95, 1.05)		-1.1
10	D10	-0.250	0.037	0.85 (0.95, 1.05)		-6.1	0.86 (0.94, 1.06)		-5.2
11	D11	0.086	0.042	1.04 (0.95, 1.05)		1.6	1.05 (0.92, 1.08)		1.3
12	D12	0.907	0.026	1.62 (0.95, 1.05)		19.6	1.51 (0.95, 1.05)		18.5
13	D13	0.973	0.035	1.31 (0.95, 1.05)		10.5	1.24 (0.94, 1.06)		7.3
14	D14	-0.654	0.042	0.97 (0.95, 1.05)		-1.1	1.06 (0.94, 1.06)		1.9
15	D14	-0.196*	0.035	0.68 (0.95, 1.05)		-13.5	0.71 (0.95, 1.05)		-11.7

An asterisk next to a parameter estimate indicates that it is constrained
 Separation Reliability = 0.996
 Chi-square test of parameter equality = 3952.82, df = 14, Sig Level = 0.000

Fig. 2. Partial Credit Model: response model parameter estimates

In order to investigate the reasons of the misfit of item D12 we proceeded to calculate the DIF by the statistical software Winstep (Linacre, 2004) In our case it is interesting to examine the functioning of items among the various departments (Dep) with three categories (Dep1, Dep2, Dep3) and the degree courses (DC) with two categories (T: three-year university degree; B: master degree). The conventional DIF allows us to test the null hypothesis that the items have the same functioning among the degree courses. Figure 3 shows the results: four items have a statistically different functioning between the degree courses at a significance level of 5%. These items are in bold in figure 3. Among them there is the item D12. The eight items in bold in figure 4 have a statistically different functioning among the departments at a significance level of 5%.

PERSON CLASSES	SUMMARY DIF CHI-SQUARE	D.F.	PROB.	ITEM Number	Name
2	.3707	1	.5426	1	2
2	25.7533	1	.0000	2	3
2	20.1291	1	.0000	3	3
2	1.1040	1	.2934	4	4
2	2.9463	1	.0861	5	4
2	1.0126	1	.3143	6	3
2	.0000	1	1.0000	7	3
2	2.5510	1	.1102	8	3
2	.0000	1	1.0000	9	2
2	2.8357	1	.0922	10	4
2	.0000	1	1.0000	11	3
2	10.9528	1	.0009	12	3
2	.9567	1	.3280	13	3
2	4.3622	1	.0367	14	3
2	3.4811	1	.0621	15	4

Fig. 3. Conventional DIF, factor: degree course

PERSON CLASSES	SUMMARY DIF CHI-SQUARE	D.F.	PROB.	ITEM Number	Name
3	.6989	2	.7038	1	2
3	7.9369	2	.0184	2	3
3	3.0302	2	.2167	3	3
3	1.3166	2	.5146	4	4
3	7.9856	2	.0180	5	4
3	44.5203	2	.0000	6	3
3	16.7022	2	.0002	7	3
3	4.0174	2	.1319	8	3
3	4.0686	2	.1285	9	2
3	.3000	2	.8618	10	4
3	5.1305	2	.0754	11	3
3	144.544	2	.0000	12	3
3	19.8891	2	.0000	13	3
3	114.618	2	.0000	14	3
3	12.5010	2	.0019	15	4

Fig. 4. Conventional DIF, factor: department

The factorial DIF is applied to consider the effects of the degree course, the department and their interaction. Several models are formed. Model 1 is a full model with 121 parameters and, after 135 iterations, it presents a deviance of 82,698.55. It explores if main and interaction effects are significant. In fact, as it is possible to read in figure 5, the model 1 includes the main effects: degree course (DC), department (Dep) and item; interaction effects: degree course by department (DC*Dep), degree course by item (DC*item), department by item (Dep*item), degree course by department by item (DC*Dep*item) and item by step (item*step). Eliminating the interaction effects: DC*Dep and DC*Dep*item from the model 1 we obtain the model 2. This last one tests degree course and department DIF simultaneously. It has a deviance of 82,803.16 with 91 parameters. The likelihood ratio (LR) test is used to compare the two models. The test is statistically significant. Thus, at least one item shows DIF interaction effect. Eliminating the terms: Dep and Dep*item from model 2 we carry out the model 3.1, eliminating the terms DC and DC*item we have the model 3.2. The likelihood ratio test between model 2 and 3.1 and between model 2 and 3.2 are both significant. Therefore, degree course and department DIF are significant. Finally, the model 4 is a model without DIF and has a deviance of 83,334.32 with 46 parameters. By comparing model 3.1 versus 4 and model 3.1 versus 4, we find both the model 3.1 and 3.2 are significant. Figure 5 shows ConQuest commands for the five models and the likelihood ratio tests applied for comparing them.

With respect to DIF effects, as the power of DIF detection depends on sample sizes, we need a classification guideline to distinguish substantial from trivial DIF effects, in addition to hypothesis testing. Draba (1977) provided a general rule to classify the model preferred: an item is identified as exhibiting substantial DIF if the difference of item difficulty estimates for two groups is significantly different from zero and is more than 0.50 logit. For instance, the DIF main effect of factor department for item D12 is significant, indicating that the difference of item difficulty between Dep1 and Dep2 is 0.686 (the students who attend the Dep1 are less satisfied with the adequacy of the classrooms). The DIF main effect of factor department is also significant for the item D06 (importance of prior knowledge), D13 (adequacy of the classrooms where supplementary lessons take place) and D14 (interest for the discipline). For sake of simplicity, in figure 6 there are the items that present a significant DIF main effect of factor department.

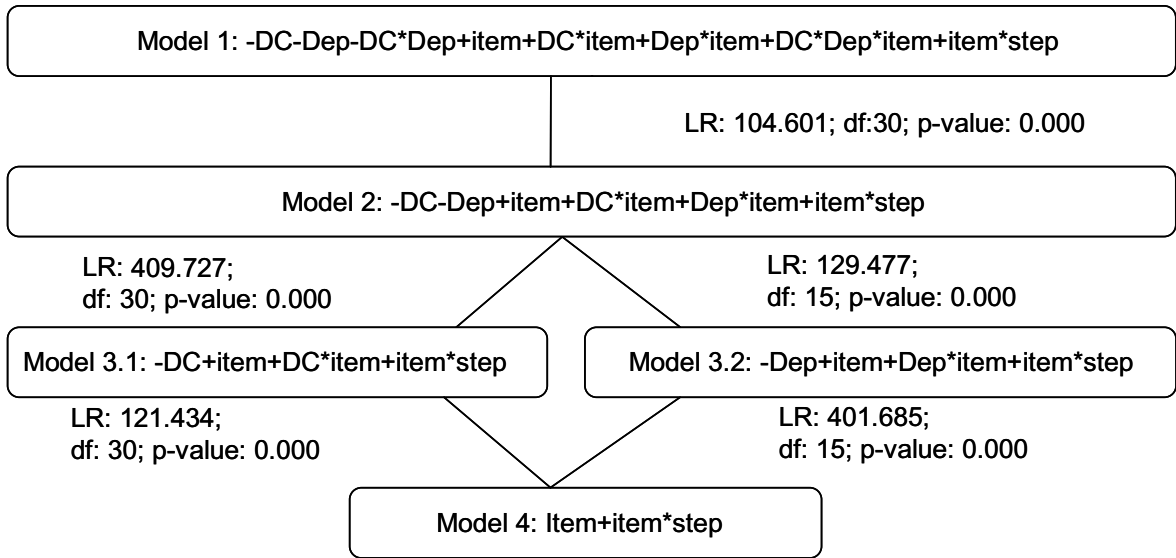


Fig. 5. Likelihood ratio tests for the five nested models, main effects: degree courses (DC), Departments (Dep), item; interaction effects: degree courses by Departments (DC*Dep), degree courses by items (DC*IT), Departments by items (Dep*item), degree courses by Departments by items (DC*Dep*item), item*step; likelihood ratio test (LR); degree of freedom (df).

VARIABLES				UNWEIGHTED FIT			WEIGHTED FIT				
fac	item			ESTIMATE	ERROR	MNSQ	CI	T	MNSQ	CI	T
1	Dip1	6	D06	0.174	0.050	1.19	(0.89, 1.11)	3.3	1.17	(0.89, 1.11)	3.0
2	Dip2	6	D06	0.146	0.052	1.25	(0.88, 1.12)	3.8	1.25	(0.88, 1.12)	3.8
3	Dip3	6	D06	-0.321*	0.044	1.21	(0.93, 1.07)	5.5	1.21	(0.93, 1.07)	5.6
1	Dip1	12	D12	0.500	0.044	1.79	(0.89, 1.11)	11.7	1.57	(0.90, 1.10)	9.7
2	Dip2	12	D12	-0.186	0.046	1.54	(0.88, 1.12)	7.6	1.49	(0.89, 1.11)	7.5
3	Dip3	12	D12	-0.314*	0.039	1.53	(0.93, 1.07)	13.0	1.45	(0.94, 1.06)	12.3
1	Dip1	13	D13	0.280	0.056	1.58	(0.89, 1.11)	9.1	1.46	(0.88, 1.12)	6.5
2	Dip2	13	D13	-0.140	0.061	1.13	(0.88, 1.12)	2.1	1.15	(0.85, 1.15)	1.8
3	Dip3	13	D13	-0.140*	0.050	1.27	(0.93, 1.07)	7.2	1.22	(0.92, 1.08)	5.2
1	Dip1	14	D14	-0.513	0.061	1.09	(0.89, 1.11)	1.7	1.21	(0.85, 1.15)	2.6
2	Dip2	14	D14	0.276	0.059	1.00	(0.88, 1.12)	-0.0	1.08	(0.87, 1.13)	1.2
3	Dip3	14	D14	0.237*	0.048	0.92	(0.93, 1.07)	-2.5	0.96	(0.92, 1.08)	-1.0

An asterisk next to a parameter estimate indicates that it is constrained

Fig. 6. Factorial DIF, factor: department

There is a substantial DIF interaction effect in item D12 (Fig. 7). In fact, it is 0.528 logits more difficult for the students who attend the Dep1 and master degree course to be satisfied with the classroom adequacy than it is for the students who attend the other two departments and three-year degree courses.

VARIABLES					UNWEIGHTED FIT			WEIGHTED FIT					
cdstb	fac	item			ESTIMATE	ERROR	MNSQ	CI	T	MNSQ	CI	T	
1	B	1	Dip1	12	D12	-0.073	0.044	1.91	(0.80, 1.20)	7.2	1.83	(0.82, 1.18)	7.1
2	T	1	Dip1	12	D12	0.073*	0.044	1.73	(0.87, 1.13)	9.3	1.46	(0.88, 1.12)	6.9
1	B	2	Dip2	12	D12	0.264	0.046	1.35	(0.78, 1.22)	2.9	1.26	(0.80, 1.20)	2.4
2	T	2	Dip2	12	D12	-0.264*	0.046	1.63	(0.85, 1.15)	7.1	1.59	(0.87, 1.13)	7.3
1	B	3	Dip3	12	D12	-0.191*	0.039	1.35	(0.84, 1.16)	3.9	1.31	(0.85, 1.15)	3.6
2	T	3	Dip3	12	D12	0.191*	0.039	1.57	(0.92, 1.08)	12.6	1.48	(0.93, 1.07)	11.8

An asterisk next to a parameter estimate indicates that it is constrained

Fig. 7. Factorial DIF: interaction effect

The measures, the abilities of the students, the difficulties of the items and the DIF parameters can be displayed graphically through the person-item map (Fig. 8). The vertical line represents the student satisfaction and each X represents 17 students. It is noted that the students are quite well approximated by a normal distribution, where the most satisfied are upward. The items are aligned to the right, and the most difficult to generate satisfaction are upward. The distribution of students is displayed in a higher position than that of item distribution. Therefore, students have a high probability to get satisfaction from university service. The items D12 and D13 are the most difficult to generate satisfaction, while the items D04, D14 are the simplest. It is possible to see that it turns out more difficult for the students who attend the Dep2 and master's degree courses to be satisfied with the classroom adequacy than it is for the others.

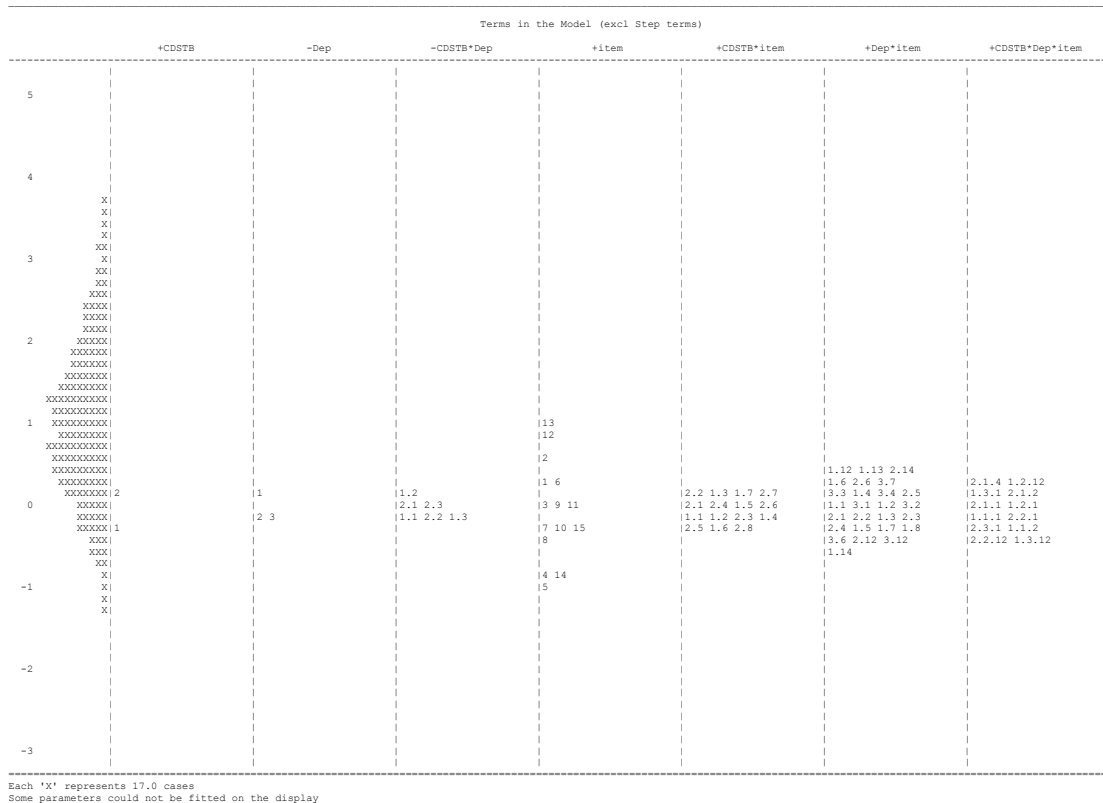


Fig.8. Model 1: item map.

4. Final remarks

Our study evaluated the satisfaction of university students considering the effects of two factors and their interaction. For this purpose, the factorial DIF is extended to PCM. Moreover, factors with more than two levels have been considered. Through this procedure we showed that the three-year degree courses, the departments, and the interaction between these two factors significantly influence the level of student satisfaction as for the infrastructures. This can be due to the fact that the classrooms where the students attend the different courses are in different buildings.

It is possible to conclude that the factorial DIF seems to be a good alternative to one-factor DIF analysis for three reasons: First, interaction effects and more factors can simultaneously be investigated. Second, a deeper decomposition can be obtained. Finally, the formulation of the factorial DIF analysis is intuitive to most data analysts since ANOVA is a well-known analysis. Besides, the factorial DIF can be extended to other multi-

parameter item response models and generalized to more than two factors. In conclusion, several DIF procedures (conventional DIF, factorial DIF Melting-Pot DIF analysis, ANOM, latent DIF detection method) can be compared, through a simulation study, to investigate statistical properties of different methods.

Acknowledgements

Comments and suggestions by the editor and two anonymous reviewers have greatly improved an early version of the manuscript. This work was financially supported by ex-60% 2011 funds of the University of Naples - L'Orientale (I) and by the agreement between University of Nigde and University of Naples Federico II.

References

- Andrich D., 1978. A rating scale formulation for ordered response categories. *Psychometrika*, 43, 561-573.
- ANVUR, 2013. Proposta operativa per l'avvio delle procedure di rilevamento dell'opinione degli studenti per l'A.A. 2013-2014. http://www.anvur.org/attachments/article/26/RilevazioneOpinioneStudDef_06_11_13.pdf
- Baker, F., Kim, S. 2004. Item response theory. Parameter estimation techniques. Dekker, New York.
- Bond, T. G., Fox, C.M., 2001. Applying the Rasch Model. Lawrence Erlbaum, Mahwah, NJ .
- Camminatiello, I., Gallo, M., Menini, T., 2010. The Rasch Model for Evaluating Italian Student Performance. *Journal of Applied Quantitative Methods*, 5 (2), 331-349.
- Draba, R. E., 1977. The identification and interpretation of item bias (Research Memorandum No. 26). Chicago: Statistical Laboratory, Department of Education, University of Chicago.
- Gallo, M., 2011. Measuring passenger satisfaction: a strategy based on Rasch analysis and the ANOM. *Journal of Applied Quantitative Methods*, 5 (2), 331-349.
- Glas, A., Verhelst, N., 1995. Tests of fit for polytomous rasch models. In Fischer, G. H., Molenaar, I. W., Rasch models. Foundations, recent developments, and applications. New York, Springer-Verlag, 325-352.
- Linacre, J.M., 2004. Facets Rasch measurement computer program. Winsteps.com, Chicago.
- Maij-de Meij, A. M., Kelderman, H., van de Flier, H., 2010. Improvement in Detection of Differential Item Functioning Using a Mixture Item Response Theory Model. *Multivariate Behavioral Research*, 45:6, 975-999.
- Rasch, G., 1960. Probabilistic models for some intelligence and attainment tests. Danish Institute for Educational Research, Copenhagen.
- Wang, W-C., 2000. The Simultaneous Factorial Analysis of Differential Item Functioning. *Methods of Psychological Research Online*, 5:1.
- Wu, M., Adams, R., Wilson, M., 1998. Acer Conquest. Generalised item response modelling software, Acer Press.
- Wright, B. D., Masters, G. N., 1982. Rating Scale Analysis. MESA Press, Chicago.
- Wright, B. D., Masters, G. N., 1997. The Partial Credit Model. In *Handbook of Modern Item Response Theory*. New York, Springer-Verlag, 101-121.
- Wright, B. D., Stone, M., 1979. *Best Test Design: Rasch Measurement*. Chicago: MESA Press.