Electronic Journal of Applied Statistical Analysis: Decision Support Systems and Services Evaluation
EJASA:DSS (2012). Vol 3, Issue 1, 49 – 58
e-ISSN 2037-3627, DOI 10.1285/i2037-3627v3n1p49

EVALUATION OF CUSTOMER SATISFACTION OF SERVICES WITH SUBGROUPS DATA

Maria Chiara Zanarotti[*(1)], Laura Pagani[(2)]

[(1)]*Department of Statistical Science, Università Cattolica del S. Cuore, Italy*
[(2)]*Department of Economics and Statistical Science, University of Udine, Italy*

Received 29 October 2011; Accepted 07 May 2012
Available online 28 December 2012

Abstract: *Frequently Customer Satisfaction (CS) surveys in services are carried out with respect to body that supply different services articulated in different contexts. These differences imply that CS surveys cannot be performed as a whole at an agency level, but need to be articulated in specifically devoted sub-surveys for each services. Differences of services and, often, also of customers, imply that CS surveys make use of ad hoc questionnaires that are administered to independent samples from different populations. Nevertheless, the goal of these surveys is to obtain information not only to evaluate each service separately, but also to obtain a valuation regarding the agency/firm as a whole. For this reason it is often desirable to obtain a measure of CS to perform separate as well as global CS evaluation. In this paper two different methods are considered to analyze CS data coming from different populations with the aim to compare and to pool CS measured at service level. By using data from an Italian Public Service, this procedure is applied to analyse the users' satisfaction with services supplied.*

Keywords: *Customer satisfaction, ordinal data, Rasch models, concurrent calibration, heterogeneity and dissimilarity index.*

1. Introduction

Customer Satisfaction (CS) starting from the 80s has become a fundamental and strategic goal in private companies in order to competitiveness and in time has spread to other sectors, particularly to service suppliers in both public and private.

[*] maria.zanarotti@unicatt.it

Contemporaneously, the interest of the academic world has grown remarkably towards CS, an interest certainly due to a great part to "*the difficulty of identifying a unique conceptual model towards which the various researches converge"* [6].
The literature regarding the measuring of CS in various services has become extremely vast: some hints can be taken from the countless works recently published on the topic (see, for example, [11] and [4]).
CS surveys are nowadays a statistical tool designed with the goal of developing continuous improvement of quality in provided services. A CS survey directly involves its catchment and allows to collect information which is fundamental in order to be administered at various levels: starting with the more simple and immediate (for example clearness of forms, effectiveness of employees, timetable and so on) progressing to a more substantial revising regarding the whole procedure of the supply of services. Moreover, CS surveys are also very important for the purpose of communications: it is important for the so called *internal communication* to share with the employees the reasons for any changes; but it is also important for the so called *external communication*, *i.e.* the necessity of the administrators to make the users known the CS results to create a transparency bond and to build consensus.
Frequently CS surveys in services are carried out with respect to bodies that supply differentiated services. For example, in health firms, services are quite different in different sectors of the firm (clinics, wards, first aid, acceptance, etc.), in municipal bodies many offices work towards different targets (registry office, educational services, social services, etc.), in Chambers of Commerce different offices are devoted to very different services (for commerce, for agricultural, for registration, etc.) and so on. All these differences demonstrate that CS surveys cannot be performed as a whole at an agency level, but need to be articulated in specifically directed sub-surveys for each service. Because of the articulated range of services and often also of customers, CS surveys make use of *ad hoc* questionnaires that are administered to independent samples concerning different groups. Nevertheless, the goal of these surveys is to obtain information not only to evaluate each body separately, but also to obtain an evaluation regarding the agency/firm as a whole. For this reason it is often desirable to obtain a measure of CS to perform a separate as well as a global CS evaluation.
In this paper two different partially complementary approaches are considered to analyze CS data gathering with samples coming from different groups with the aim of comparing and pooling all CS measured at the single service level. To perform one of the two above mentioned approaches, one more consideration is necessary; that is, the *ad hoc* questionnaires used for each of the sub-groups must share some items (the so called common items) with each one of the other questionnaire forms. Note that this request is not restrictive in practice, since the designing of questionnaires with some common items and other more specific ones is very widespread.
In section 2, the methodological tools which the authors suggest should be used are summarized. Some empirical results obtained from data from one Italian Public Service sector are sketched in section 3. Concluding remarks are outlined in section 4.

2. Methodological tools

As stressed by Cronin and Taylor [5], the measurement of CS for services is not trivial: *"Service quality is an elusive and abstract construct that is difficult to define and measure"*. No tangible

and objective characteristics (for example, durability, presence/absence of defects, and so on) are available in the case of services, as it happens when CS regards goods. Data about service quality are usually obtained by responses users give to a set of items and these responses are typically on a categorical level. Many statistical tools have been developed to measure CS of services, and the one which is mostly used is the so called SERVQUAL model of Parasuraman *et al*. [20] that is based on the *discrepancy paradigm*, i.e. the difference between *expected quality* and *perceived quality* of the users for each item. But to use SERVQUAL model customer evaluations have to be quantitative, and this is not the case in a lot of CS surveys. For this reason in this paper the tools considered to analyze CS data are specifically devoted to ordinal data.

Since the intended goal is to measure CS for each service as well as for the whole agency, the problem is how to perform disaggregate analysis as well as a global one, this requires that results at the two levels (disaggregate and aggregate/global) can be comparable.

Disaggregate analysis is important for the purpose of improving each service for the less valued aspects of the service, thus achieving a continuous required improvement; it is also important for internal communication. Aggregate analysis, on the other hand, is very important for a lot of reasons: for global evaluation of the agency, for external communications and for monitoring the body during that period.

The methods proposed in this paper make use of dissimilarity indexes between ordinal distribution and the use of the Rasch model.

The first method considered is based on the Dissimilarity Index (DI) among ordinal distribution. The *idea* is to compare, for each item, the observed distribution of users responses with a theoretical one, which has been selected as a comparative model. The comparative model can be, for example, the one that represents the most negative opinion (users choose the lowest category) or the most positive one (users choose the highest category). DI returns a numerical value that reflects the degree of similarity between the observed distribution and the theoretical one. The DI values, one for each item, can then be aggregated with any linear operator to obtain a synthetic measure of CS for each service.

The DI we use in this paper is the one suggested by several authors (see, for example, [3]) and it is the *simple relative dissimilarity index* (see, for example, [15]), denoted with z*:

$$z^* = \frac{1}{K-1}\sum_{k=1}^{K-1}\left|F_k^O - F_k^T\right| \qquad (1)$$

where: F_k^O is the observed cumulative distribution and F_k^T is the theoretical one.

Index z*in (1) takes values from zero to one: it takes value zero if the two distributions are equal and value one when they are very different (*i.e.*: all frequencies are concentrated on the first category for one distribution and all frequencies are concentrated on the last category for the other distribution). The value of z* increases with the growth of the dissimilarity of the two distributions.

Taking as theoretical distributions the most severe one (indicated with F^L) and the most positive (indicated with F^H), it is possible to define, for each item, the following indicators of satisfaction:

$$\mathrm{IS}^{\mathrm{H}} = 1 - \frac{1}{k-1}\sum_{k=1}^{K-1}\left|F_k^O - F_k^H\right| = 1 - \frac{1}{k-1}\sum_{k=1}^{K-1}F_k \tag{2}$$

$$\mathrm{IS}^{L} = \frac{1}{k-1}\sum_{k=1}^{K-1}\left|F_k^O - F_k^L\right| = 1 - \frac{1}{k-1}\sum_{k=1}^{K-1}F_k \tag{3}$$

Note that in (2) is considered the complement of index z* to one so that IS^{H} increases as differences between observed distribution and the highest one (most positive opinion) decrease.
On the other hand, it is not necessary to take the complement to one for index IS^{L} in (3), since it increases as the difference between the observed distribution and the most severe one grows. It is trivial to show that the two indexes lead to the same result, so in the sequel they are simply indicated with IS.
Having the IS values for each item, it is possible to make an aggregation of all these indexes through of a simple or pounded mean so that a synthetic value can be obtained for each service of the body.
Taking the simple arithmetic mean, the synthetic index for the *m-th* services assumes the form:

$$\mathrm{I_m} = \frac{1}{J}\sum_{j=1}^{J} IS_{jm} \tag{4}$$

Index $\mathrm{I_m}$ takes values from zero (greatest dissatisfaction) to one (greatest satisfaction).
Finally, having $\mathrm{I_m}$ indexes available, further aggregations are possible to obtain an overall measurement of satisfaction for the whole agency.
The second method considered makes use of the Rasch Model (RM) and, in particular, of the so called *concurrent calibration* that through RM it is possible to achieve. Concurrent calibration is a linking technique introduced in psychometric analysis. There is a very large literature on test linking and equating (among others: Misley [18], Linn [16], Kolen and Brennan [14] and Dorans and Holland [8]). In psychometrics the term "linking" is referred to all those techniques that permit a transformation of a score on one test to a score on another test. It is a very common situation in many disciplinary contexts (for example: psychology, education, health, rehabilitation, and so on) to handle with some kind of measurements obtained through measurement tools that could be even quite different and that produce results which are not directly comparable. As Dorans and Holland [8] say: "*The comparability of measurements made in differing circumstances by different methods and investigators is a fundamental pre-condition for all science.*". There are different types of links, and Holland and Dorans [12] sketched a possible classification in three basic typologies: *predicting*, *scale aligning* and *equating*. For a comprehensive description of all these categories and sub-categories see Dorans [7]. In particular *scale aligning methods* have the goal to put on a common scale measures expressed originally using different ones. Concurrent calibration is one of these scale aligning methods. To perform all kinds of links between measurements some conditions have to be met: *i.e.* some kind of interconnection between measurement data is necessary. "*The role of data collection is crucial to successful instrument linking*" (Dorans, [7]). If, as in the case considered in this paper, data used

to produce the measurements are the responses of subjects to items collected from questionnaires (also called *tests*), data collection design is crucial to make a link feasible. The linkage between two or more tests can be obtained with two principal data collection designs one that is based on common groups and one that is based on common items. In the first case the same sample of subjects (or equivalent samples from the same population) all take different tests, so that differences in measurements are not imputable to differences in ability/satisfaction but are only due to the differences between tests. In the second case, a set of common items (*anchor items*) is included in each test form so that the common group need is not any more necessary. Each sample takes only one test form and responses on the set of common items can measure the ability/satisfaction differences between such samples that are not necessarily equivalent. In this paper the focus is on this anchor test design and, in particular, on a weaker version of it. In anchor test design there are M populations $P_1,\ldots P_M$, with a sample from each population, each taking a different test form $T_1,\ldots,T_M$. In addition each sample takes also an anchor test T_0, the same for every sample. In the case of only two groups (and two samples), Kolen and Brennan [14] have referred to this design as the *common-item* (or *anchor* test) *non-equivalent group design* (NEAT). As Dorans et al. [9] outline referring to Rosenbaum [21], one way of thinking about the differences between NEAT design and other designs used in linking procedures "*is as the difference between observational studies versus experimental designs(…) the NEAT design is like an observational study where there are two non-randomized study groups that are possibly subject to varying amounts of self-selection"* (Dorans et al., [9]). In this paper we consider a weaker form of NEAT: the design considered is one where each sample from each group P_m takes a test formed by a set of unique items and a linking set of items that are common to one or two or more (or all) the administered tests. These two sets of items can vary from a handful to a large number of questions. More specific details regarding the design used in this paper will be sketched in section 3.

Concurrent calibration (also known as *multiple-group calibration*, see Kim and Kolen [13] and references in it) is applied in this paper under the weaker common-item nonequivalent group design. Through concurrent calibration it is possible to obtain estimated parameters on the same scale and to achieve the goal of comparability when data are referring to several groups.

"*In concurrent calibration, parameters of all items from multiple forms are simultaneously estimated through a single computer run with all response data from multiple forms being combined together, and as a result, all the estimates are placed on a designated common scale."* (Kim and Kolen [13]). In concurrent calibration, the differences in satisfaction between groups is accounted for because the groups are given the linking common-item set so that the satisfaction of all groups is estimated on the same scale for every group. Concurrent calibration is performed in this paper using the one-parameter Item Response Model (IRT), *i.e.* the Rasch Model (RM). Through RM it is possible to measure CS (the so called *latent* treat): this measurement is based on the hypothesis that the probability of the response to each item from each subject is a function of three sets of parameters. The first is the set of *Person Location Parameters* (PLP, denoted θ_i). These parameters express users satisfaction and reflect all individual and context elements that can influence satisfaction. The second set is the one of *Item Location Parameters* (ILP, denoted β_j). These parameters denote the qualitative level embedded in each facet of the service indicated by each item. The last set of parameters is the *Threshold parameters* set (denoted τ_{kj}). These parameters are related to consecutive couples of response categories and are limited by respective intervals. Threshold parameters represent the difficulty of endorsing one response category instead of the previous one. The polytomous Rasch model considered in this paper is

the so called *Partial Credit Model* (PCM) (Master [17]), where the *logit* of the probabilities of two consecutive response categories (k, k-1) is:

$$\ln\frac{P\left(X_{ij}=k\right)}{P\left(X_{ij}=k-1\right)}=\theta_i-\left(\beta_j+\tau_{k_j}\right) \qquad \text{where: } \sum_{k=1}^{K_j}\tau_{kj}=0 \tag{5}$$

and X_{ij} denotes response of person *i (i=1,...,n)* to item *j (j=1,...,J)*.

As it is possible to appreciate by looking at the model in (5), in PCM the *logit* of the probabilities user *i-th* gives the answer *k* instead *k-1* to item *j-th* which is equal to the difference between θ_i (the so called *person location parameter* of user *i*) and the sum of two parameters $\beta_j + \tau_{kj}$ (respectively: *item location parameter* and *threshold parameter)*. It is important to stress that the set of threshold parameters is different (also in its number) for each item.
Some assumptions are fundamental in RM (*unidimensionality* and *local independence*, for example) for parameters identifiability and estimation: a full examination of these assumptions is beyond the scope of this paper. For a detailed discussion see, for example, Fisher and Molenaar [10]. Concurrent calibration via RM (PCM, in particular) as well as the measure of satisfaction obtained through indexes IS are performed and compared in the next section with data detected to measure users satisfaction in an Italian Chamber of Commerce.

3. A case study

In the following section the proposed methods are used to analyze users satisfaction in different offices of an Italian Chamber of Commerce.

Table 1. Structure of questionnaires and number of valid responses.

Request: *Mark each item from 1 (very negative valuation) to 10 (very positive valuation)*		**Office**				
Item		**A**	**B**	**C**	**D**	**E**
I01	Clarity and completeness of the information received	x	x	x	x	x
I02	Competence of operator at the counter	x		x	x	x
I03	Operator courtesy (and availability)	x	x	x	x	x
I04	Length and timetable of the service	x	x			
I05	Availability of the service by e-mail or by internet	x				
I06	Availability of clear and complete information (in office or website)	x		x	x	x
I07	Speed in waiting time at the counter		x			
I08	Operator speed in the delivery of the service at the counter		x			
I10	Availability and functionality of waiting rooms		x			
I11	Clearness and simplicity in forms compilation		x		x	
I12	Easy access to the office by phone			x	x	x
I13	Reliability and accuracy of the delivered service			x	x	
Number of items for each office		**6**	**7**	**6**	**7**	**5**
Number of valid questionnaires for each office		**66**	**81**	**74**	**95**	**30**

The data refer to responses of 346 valid questionnaires administrated during the last two months of 2008 in five offices regarding different types of services and labeled, for reasons of privacy, with a capital letter from A to E. Table 1, summarizing information on the structure of questionnaires and the number of valid responses, shows that some items are common and others are specific to different offices.
For example item I01 (clarity and completeness of information received) is common to all offices but item I08 (operator speed in the delivery of the service at the counter) is specific to office B.
Response categories, for all questionnaires, are in an ordinal (ten-point) scale from 1 (very negative evaluation) to 10 (very positive evaluation).
Following the procedure illustrated in Section 2 the analysis of CS begins with ranking items and offices using the *IS* and I_m indexes calculated with the original ten-point ordinal scale. Results are reported in Table 2.

Table 2. Item - office rankings using IS and I_m indexes (ten-point ordinal scale).

			Item ranking (from best to worst one)						
			Best	←	←		→	→	Worst
Office ranking (from best to worst one)	Best	**E** (0.8482)	**I03** (0.8778)	**I02** (0.8704)	**I01** (0.8556)	**I06** (0.8185)	**I12** (0.8185)		
	↓	**B** (0.8262)	**I08** (0.8793)	**I03** (0.8560)	**I01** (0.8464)	**I07** (0.8395)	**I11** (0.8148)	**I10** (0.7984)	**I04** (0.7490)
		A (0.8145)	**I03** (0.8939)	**I02** (0.8552)	**I01** (0.8030)	**I05** (0.7996)	**I06** (0.7828)	**I04** (0.7525)	
	↓	**C** (0.7988)	**I03** (0.8228)	**I02** (0.8078)	**I01** (0.7943)	**I12** (0.7898)	**I13** (0.7898)	**I06** (0.7883)	
	Worst	**D** (0.7467)	**I03** (0.8105)	**I02** (0.7637)	**I01** (0.7450)	**I12** (0.7427)	**I13** (0.7322)	**I06** (0.7216)	**I11** (0.7111)

Note: in brackets *IS* values for items and I_m values for offices

Item-office ranking shows that:
- the best aspects of the services offered by the different offices are related to staff behavior and the worst aspects regard organization;
- the most appreciated services concern office E while the least appreciated regard office D.

The overall level of satisfaction, taking again values between zero and one, is obtained as an arithmetic mean of I_m indexes in (4). The value is 0.8069, namely about 81% of the maximum.
I_m and GSI indexes are very easy to obtain and simple to interpret and can help the structure to identify critical situations. An alternative method, as already mentioned in Section 2, is to employ concurrent calibration through RM (specifically PCM) with the purpose of analyzing the overall quality of the service. The Rasch (global) analysis considers 13 items and 346 valid questionnaires.

As already mentioned questionnaires of the five offices are partially common, this leads to the design that is a generalization of the so called *Non Equivalent Group with Anchor Test* (NEAT) *design* pointed out in Section 2. Table 3 represents the design considered in this application.

Table 3. Design table.

Item Office	I01	I02	I03	I04	I05	I06	I07	I08	I10	I11	I12	I13	n. of valid responses
A	■	■	■	■	■	■							66
B	■		■	■			■	■	■	■			81
C	■	■	■			■					■	■	74
D	■	■	■			■				■	■	■	95
E	■	■	■			■					■		30

The Rasch analysis on the set of 346 questionnaires is carried out using RUMM 2010 (Andrich, *et al*., [1]), one of the standard software for RM. First results are not suitable because

1. the original ten-point scale is not satisfactory since the set of "unsatisfied" categories (from 1 to 5) has associated zero or very low frequencies creating disordered thresholds;
2. test of fit to the model (related to item fit), based on chi-squared statistics, is significant (X^2=165.035, DF=60, p-value<0.05) indicating that unidimensionality is not reached.

Reliability index (associated to person fit) through *the person separation index* (PSI), on the other hand, is good (PSI= 0.86) meaning that the scale is able to discriminate between people of different satisfaction levels.

To solve the problem of disordered thresholds is necessary to rescore items by collapsing redundant categories into adjacent ones (Bond and Fox [2], Pagani and Zanarotti [19]). The collapsing procedure brings to new appropriate scales:

- a six-point scale: 1 = *negative assessment* (marks from 1 to 5), 2 = *sufficient assessment* (marks 6), 3= *quite good assessment* (mark 7), 4 = *good assessment* (mark 8), 5 = *very good assessment* (mark 9) and 6 = *optimum assessment* (mark 10).
- a five-point scale: 1 = *negative or sufficient assessment* (marks from 1 to 6), 2 = *quite good assessment* (mark 7), 3 = *good assessment* (mark 8), 4 = *very good assessment* (mark 9) and 5 = *optimum assessment* (mark 10).

The five-point scale is appropriate for items I01, I04 and I13, while the six-point scale applies to other items.

The strategy for improving fit and removing lack of unidimensionality is to partition the value of the global chi-square (GCS) into the specific item chi-square (SICS). If SICS is significant then corresponding item has to be deleted. Analysis of SICS (not reported here for brevity) highlights that items I04 and I06 should be deleted. After the process of categories collapsing and item deletion a new concurrent calibration is then carried out applying overall PCM. The new analysis provides satisfactory results: the reliability index is good (PSI=0.889) the GSC is not significant ((X^2=90.53, DF=72, p-value>0.05).

Table 4. Item ranking and ILP using PCM with the new scale and 10 items.

Item	**I08**	**I03**	**I02**	**I07**	**I12**	**I11**	**I10**	**I05**	**I01**	**I13**
Ranking	Best	←	←					→	→	Worst
ILP	-1.959	-0.621	-0.289	0.010	0.222	0.331	0.493	0.494	0.609	0.732

Users satisfaction, measured by person parameters reached with concurrent calibration, is compared with results obtained employing I_m indexes calculated with the new evaluation scales (Table 4). Notice that the I_m values are now lower than the ones in Table 2 and also that the ranking of the offices has changed.
The overall level of satisfaction, is now 0.6803, namely about 68% of its maximum, quite lower than the one calculated with the original ten point scale.

Table 5. I_m values and PLP averages.

Office	B	E	A	C	D
I_m value	**0.7326**	**0.7704**	**0.7070**	**0.5748**	**0.6168**
PLP Average (SE)	**1.7976 (1.9827)**	1.**7923 (2.0360)**	**1.4951 (1.5845)**	**0.8471 (1.3711)**	**0.4497 (1.8105)**

Note: in brackets: PLP standard error.

4. Concluding remarks

This paper suggests (and compares) the use of dissimilarity indexes and concurrent calibration to assess users satisfaction in PA due to the fact that surveys frequently involve organizations that supply different services to different users. In these circumstances the main difficulty regards the data processing when they are collected in a disaggregate way and it is necessary to perform not only a disaggregate analysis, but also an overall evaluation. The procedure reported in this paper can be used to achieve these two goals.

References

[1]. Andrich, D., Sheridan, B., Luo, G. (2004). *RUMM 2020, version 4.0*. RUMM Laboratory Pty Ltd, Perth.
[2]. Bond T.G., Fox C.M. (2001). *Applying the Rasch Model: Fundamental Measurement in the Human Sciences*. London: Lawrence Erlbaum Associates.
[3]. Capursi, V, Porcu, M. (2001). La didattica universitaria valutata dagli studenti: un indicatore basato su misure di distanza fra distribuzioni dei giudizi, in: *SIS 2001"Processi e metodi statistici di valutazione",* Roma, 4-6 June 2001.
[4]. Carpita, M., Manisera, M. (2006). Un'analisi delle relazioni tra equità, motivazione e soddisfazione per il lavoro, in *Valutare la qualità - I servizi di pubblica utilità alla persona*, eds. M. Carpita, L. D'Ambra, M.Vichi and G. Vittadini, Milano: Guerini, 311-360.
[5]. Cronin, J.J. Jr., Taylor, S.A. (1992). Measuring Service Quality: A Reexamination and Extension, *Journal of Marketing*, 56, 55-68.
[6]. D'Ambra, L., Gallo, M. (2006). La valutazione della Customer Satisfaction, in *Valutare la qualità - I servizi di pubblica utilità alla persona*, eds. M. Carpita, L. D'Ambra, M.Vichi and G. Vittadini, Milano: Guerini, 267-289.
[7]. Dorans (2007) Linking scores from multiple health outcomes instruments, *Quality of Life Research*, 16(1), 85-94.

[8]. Dorans, N.J., Holland, PW. (2000). Population invariance and the equatability of tests: basic theory and the linear case. *Journal of Educational Measurement*, 37(4), 281-306.
[9]. Dorans, N.J., Moses, T., Eignor, D. (2010). Principles and practices of test score equating (ETS Research Report No. RR-10-29). Princeton, NJ: ETS.
[10]. Fischer, G.H., Molenaar, I.W. (1995). *Rasch Models. Foundations, Recent Developments, and Applications*. New York: Spriger-Verlag.
[11]. Gori, E., Vittadini, G. (1999). *Qualità e valutazione nei servizi di pubblica utilità*. Torino: Etas.
[12]. Holland, P.W, Dorans, N.J. (2006). Linking and equating, in Educational measurement,4th ed., ed. R.L. Brennan, Westport, CT: Praeger, 187-220.
[13]. Kim, S., Kolen, M. J. (2007). Effects on Scale Linking of Different Definitions of Criterion Functions for the IRT Characteristic Curve Methods. *Journal of Educational and Behavioral Statistics*, 32(4), 371-397.
[14]. Kolen, M.J, Brennan, R.L. (2004). *Test equating, scaling, and linking. Methods and practices,* 2nd ed. New York: Springer-Verlag.
[15]. Leti, G. (1983). *Statistica descrittiva*. Bologna: Il Mulino.
[16]. Linn, R. L. (1993). Linking results of distinct assessments. *Applied Measurement in Education*, *6*(1), 83−102.
[17]. Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47, 149-174.
[18]. Mislevy, R.J. (1992). *Linking Educational Assessments: Concepts, Issues, Methods, and Prospects* . Princeton, NJ: Educational Testing Service.
[19]. Pagani, L., Zanarotti, M.C. (2010). Some Uses of Rasch Models Parameters in Customer Satisfaction Data Analysis. *Quality Technology & Quantitative Management*, 7(1), 83-95.
[20]. Parasuraman, A., Zeithaml, V.A., Berry, L.L. (1988). SERVQUAL: A Multiple-Item Scale for Measuring Consumer Perceptions of Service Quality, *Journal of Retailing*, 64, 12-40.
[21]. Rosenbaum, P. R. (1995). Observational studies. New York, NY: Springer-Verlag.