Electronic Journal of Applied Statistical Analysis
EJASA, Electron. J. App. Stat. Anal.
http://siba-ese.unisalento.it/index.php/ejasa/index
e-ISSN: 2070-5948
DOI: 10.1285/i20705948v9n2p433

Clustering dichotomously scored items through functional data analysis
By Di Battista, Fortuna

Published: 14 October 2016

Electronic Journal of Applied Statistical Analysis
Vol. 09, Issue 02, October 2016, 433-450
DOI: 10.1285/i20705948v9n2p433

Clustering dichotomously scored items through functional data analysis

Di Battista Tonio*and Fortuna Francesca

DISFPEQ, University G. D' Annunzio,
Pescara, Italy

Published: 14 October 2016

In the educational field, it is common to analyze the probability of a correct response to a test item as a continuous function of the item parameters and the subject ability. This relation is given by the item response function. Since test data are expressed as curves, they can be analyzed through the functional data analysis approach. Indeed, several researchers suggest to estimate the shape of the item response function through a non-parametric approach in order to catch unusual or unforeseen features in the curve. On the contrary, item response theory models assume a specific parametric functional form for the item response function. In this paper, we propose an alternative method that combines the parametric specification of the common item response theory with the functional data analysis approach. In particular, we aim to classify the items through some clustering algorithms exploiting the characteristics of convex function spaces. The key idea is to transform the function space of the items in a convex space which guarantees desirable properties. Specifically, we prove that, exploiting the convexity property, the functional mean belongs to the same function space as the item response functions. The applicability of our proposal in the educational field, is demonstrated through a real data set concerning test data of the INVALSI mathematics test administered to upper secondary school students.

keywords: Item response theory, functional k-means algorithm, convex function space, functional hierarchical Ward's method, INVALSI test.

*Corresponding author: dibattis@unich.it

ISSN: 2070-5948
http://siba-ese.unisalento.it/index.php/ejasa/index

1 Introduction

Item Response Theory (IRT) refers to a class of statistical models used to describe the association between the response behaviour of subjects to a set of categorically scored items and the underlying latent trait which is indirectly measured by the items (Lord and Novick, 1968). This relation is given by the item characteristic curve or item response function, which shows how the probability of success on a test item depends on the examinee ability level.

In this context, test data are presented as curves, thus, the functional data analysis (FDA) approach (Ramsay and Silverman, 2005; Ferraty and Vieu, 2006) may be considered. FDA refers to the analysis of curves or functions in a continuous domain. This approach assumes the existence of unknown smooth functions $f(\cdot)$ which generate and underlie the data. However, in real applications, the functions are observed as a sequence of discrete data at a finite number of points of the domain, thus, the FDA approach primarily aims to fit the true form of the underlying function through some suitable techniques, such as basis functions expansion and regularization. This approach has been used in many research fields, such as meteorological (Ramsay and Dalzell, 1991), medical (Pfeiffer et al., 2002), spectometric (Reiss and Ogden, 2007), and ecological (Gattone and Di Battista, 2009; Di Battista et al., 2016) ones. Although FDA is a method to analyze observed curves, Ramsay (1997) highlights that it can also be applied to those implied by and estimated from data that are not at all curves at first sight, such as test data. Indeed, in most IRT models, it is assumed that the item response curve is included in a restricted class of functions defined by a specific mathematical models, such as normal ogive, step, polynomial, logistic functions. However, several researchers suggest to estimate the shape of the item response function in the functional framework, without prior restrictive assumptions about its mathematical form. Ramsay (1991), for example, uses the Kernel regression to non-parametrically estimate the item response function; Rossi et al. (2002) and Matthew (2007) model the log-odds of the item response functions with B-splines, because this transformation of the item response function does not necessarily imply constraints on its value. The authors claim that a non-parametric approach is a more flexible method than the standard IRT, because it allows to model unusual or unforeseen features of the curve. However, parametric models are often motivated by some level of analysis of the psychological processes involved in the choice that an examinee makes when confronted with an item. Moreover, simple parametric models exhibit desirable statistical and mathematical properties, tempting one to wish that they are true (Ramsay, 1997). In this paper, we propose an alternative method for the analysis of test data, which combines the parametric specification of the common IRT models with the FDA approach. Our proposal allows to preserve the usual interpretation of test data in the IRT framework, nevertheless taking advantage of the functional data analysis tools. Indeed, the FDA approach grants a deeper analysis of those phenomena varying in a fixed domain (Ramsay and Silverman, 2005; Ferraty and Vieu, 2006) and, in an educational framework, allows the evaluation of the items behavior throughout the reference domain. In this context, the observed item response curves are expressed by a specific parametric function, thus, the function space is constituted by a set of curves belonging to the same parametric family (De Sanctis and Di Battista, 2012; Di Battista and Fortuna, 2013). The novelty of our method lies in exploiting the known form of the function underlying the data. It allows to work directly on the function space by avoiding the use of smoothing techniques. In particular, we refer to the characteristics of the convex function spaces, in order to achieve some desirable properties. The main advantage in dealing with convex subset of functions is that we can obtain a synthesis measure of the observations that belongs to the same function space as the data. However, not all the functions constitute a convex function space. For example, in this paper, we refer to the parametric logistic IRT models (Lord and Novick, 1968) that specify the probability of a correct response to an item as a logistic function of items and subjects parameters. Thus, the subset of functions is neither a vector subset nor a convex subset. In this setting, we propose a transformation one-to-one that converts non-convex subset of functions to a convex one. In this way, it is possible to obtain functional summary statistics belonging to the same function space as the data.

The advantages of this approach become clear when the functional k-means algorithm is considered. The k-means algorithm is an iterative procedure that assigns curves to the cluster whose centroid is closest according to a specific distance. The functional centroids is the representative function of a cluster; thus, it should belong to the same function space as the observed functions. However, the method may also be applied to other cluster algorithms such as the agglomerative Ward's hierarchical method. Moreover, as the parametric functional form of the observations is known, the functional distance can be directly computed in the explicit form of the functions. Thus, clustering results do not differ depending on how the curves are fitted to the data.

The paper is organized as follows: Section 2 reviews standard parametric logistic IRT models; Section 3 deals with functional clustering algorithms for those functional data that belong to a specific parametric family, focusing on the k-means clustering algorithm and the Ward's hierarchical clustering method. Furthermore, the Section presents a suitable procedure for non-convex subsets, in order to compute functional centroids that belong to the same subset of the observed functions. Section 4 applies our method to a real data set concerning the INVALSI mathematics test administered to upper secondary school students. Section 5 provides some concluding remarks.

2 Parametric IRT models

Item response theory (IRT) refers to a family of latent trait models that is commonly used to study a latent trait (usually the subject ability) associated with a set of categorically scored items in a test (Lord and Novick, 1968). It provides a mathematical model to explain the relationship among item characteristics, subject ability and the probability of a correct response and assume that both persons as well as items have a position on the same underlying dimension the test is measuring (Braeken, 2008). Although IRT models have a predominant role in educational assessment, they become a popular framework in many fields, including customer satisfaction surveys (Bradlow and Zaslavsky, 1999; Valentini et al., 2011) and public health (Aggen et al., 2005; Sharp et al., 2006).

The basic representation of IRT models is given by (Lord and Novick, 1968):

$$P(X = x|\theta) = f(\boldsymbol{\eta}, \theta) \tag{1}$$

where X represents the score on the test item; x is a possible value for the score; $\boldsymbol{\eta}$ is a vector of parameters that denotes the characteristics of the test item, θ represents the single parameter that describes the subject characteristics and f is a function which defines the relationship among the item parameters and the probability of the response. Different IRT models arises from different functional forms assumed for f, and a different number of item parameters. The most common IRT models for dichotomous items assume one (Rasch, 1960), two (Birnbaum, 1968) or three (Birnbaum, 1968) item parameters and define f in Equation (1) as a logistic function. The Rasch model and the two parameters logistic (2PL) model specify the probability of a correct response on an item as a function of items and subject parameters respectively as follows:

$$P_j(\theta_i) = P(X_{ij} = 1|\theta_i, \beta_j) = \frac{\exp(\theta_i - \beta_j)}{1 + \exp(\theta_i - \beta_j)} \tag{2}$$

$$P_j(\theta_i) = P(X_{ij} = 1|\theta_i, \alpha_j, \beta_j) = \frac{\exp[\alpha_j(\theta_i - \beta_j)]}{1 + \exp[\alpha_j(\theta_i - \beta_j)]} \tag{3}$$

where X_{ij} denotes the observed response of the i-th subject ($i = 1, ..., n$) to the j-th item ($j = 1, .., J$), with $X_{ij} = 1$ representing the correct responses and $X_{ij} = 0$ the incorrect one; θ_i indicates the ability of the i-th subject; while α_j and β_j are the discrimination and the difficulty parameter of the j-th item, respectively. The conditional probability functions in Equations (2) and (3) are called item characteristic curves or item response functions; they specify how the probability of an item response changes due to changes in the latent variable and, clearly, they are monotonically increasing functions of the latent trait (Braeken, 2008). The parameters α and β describe the shape of the item response function. In particular, β is a location parameter and represents the point on the latent continuum where the probability of a positive response for the i-th subject is equal to 50%. The larger the difficulty parameter, the more the ability a respondent must have to endorse that item. The discrimination parameter, α, is the slope of the item characteristic curve at the value of the location parameter. It reflects how well the item is able to differentiate between subjects having different ability levels and indicates how strongly related the item is to the latent trait. Items with high discriminations are better at differentiating respondents around the location point, in that small changes in the latent trait lead to large changes in probability. The Rasch model can be viewed as a special case of the 2PL model where the discrimination parameters are constant over items, that is $\alpha_j = 1\ \forall j$, $j = 1, 2, ..., J$.

The three parameters (3PL) logistic model introduces a nonzero lower asymptote for the probability of a correct response. It can be expressed as follows:

$$P_j(\theta_i) = P(X_{ij} = 1|\theta_i, \gamma_j, \alpha_j, \beta_j) = \gamma_j + (1 - \gamma_j)\frac{\exp[\alpha_j(\theta_i - \beta_j)]}{1 + \exp[\alpha_j(\theta_i - \beta_j)]} \tag{4}$$

where γ_j is known as the guessing parameter of the j-th item and represents the probability of getting the item correct just trying a guess. When $\gamma_j = 0$ the 3PL model resembles the 2PL model.
The basic assumptions behind the above models are:

- Unidimensionality: the set of items are indicators of a single continuous latent variable θ;
- Local independence: the subject responses to a set of items are uncorrelated for a given value of θ;
- Monotonicity: the probability of correct response to the item increases as the ability of the examinees increases.

Under these conditions, in case of dichotomous items, the expected score on the j-th item for the i-th subject, is equal to the probability to obtain a correct response at a given ability level:

$$E(X_{ij}|\theta_i) = Pr(X_{ij} = 1|\theta_i) = P_j(\theta_i) \quad i = 1, 2, ..., n; \;\; j = 1, 2, ..., J \tag{5}$$

and the expected test score of the i-th subject is given by (Grayson, 1988; Huynh, 1994; Matthew, 2007):

$$E(X_i|\theta_i) = E\left(\sum_{j=1}^{J} X_{ij}|\theta_i\right) = \sum_{j=1}^{J} E(X_{ij}|\theta_i) = \sum_{j=1}^{J} P_j(\theta_i) \quad i = 1, 2, ..., n \tag{6}$$

In the IRT framework, the plot of the expected test score in Equation (6) against θ is typically called test characteristic curve, $TCC(\theta_i)$, that shows the expected test score as a function of the underlying latent variable, θ.
For dichotomous items, it is more useful to consider the proportion of correct answers in the tests, thus, the expected test score is divided by the maximum possible score (Weiss, 1995) and the TCC turns out to be equal to the average of all the item response functions:

$$TCC(\theta_i) = \sum_{j=1}^{J} \frac{P(X_j|\theta_i, \boldsymbol{\eta}_j)}{J} \tag{7}$$

In this formulation, it is possible to compare TCC from tests with different numbers of test items.

3 Functional clustering on convex function spaces

Test data may do not appear explicitly as functional data, because they are usually zeros and ones indicating unsuccessful and correct answers to test items. Nevertheless, in the educational field, it is common to analyze the probability of a correct response as a continuous function of the item parameters and the subject ability. In this context, the observed data consist of a set of item response functions, $P_j(\theta_i)$, $j = 1, 2, ...J$, one per test item that, usually, are expressed by a specific parametric model. Thus, the functional data observed for each unit, belong to a function space, say S, with m real parameters, that is:

$$S = \{f(\boldsymbol{\eta}; \theta)\} \tag{8}$$

where f is a parametric function, $\boldsymbol{\eta} = (\eta_1, \eta_2, ..., \eta_m)^T$ represents a set of unknown item parameters taking values in a parameter space Ω; θ is the domain of the functions ($\theta \in \mathbb{R}$), and S is a subset of some $L^p(X)$ space. In particular, according to the parametric IRT models introduced in Section 2, $\boldsymbol{\eta}$ represents the items parameters vector and f is a logistic function. Under parametric IRT models, $P_j(\theta_i)$ can be viewed as the regression of item score on the underlying variable θ (Lord and Novick, 1968) and the logistic function specifies a monotonically increasing function such that higher ability results in a higher probability of success.

In this setting, test data can be analyzed in the FDA framework. Moreover, since the functional form of the observations is known in advance, the approximation of the function underlying the data through smoothing techniques is not required (Di Battista et al., 2016). The main advantage to be gained by the FDA approach is the analysis of curve characteristics with functional tools. In particular, each item can be studied through the shape of the item characteristic curve. Indeed, the curves in a functional data set may present a variety of distinctive patterns corresponding to different shapes and variation that can be identified by clustering the functions (Tarpey, 2007; Sangalli et al., 2010b). Starting from J parametric item response functions, $P_j(\theta_i) \in S$, we aim to identify a set of homogeneous clusters in L^p by determining a partition of the space according to the minimal distance. In particular, an L^2 metric in function space is applied combined with both a k-means algorithm and hierarchical methods for finite dimensional data. Specifically, the k-means algorithm (Forgy, 1965; MacQueen, 1967) is an iterative procedure initialized by fixing K clusters, C_k, $k = 1, 2, ..., K$, and by randomly selecting in S a set of arbitrary initial centroids, $\{\phi_1^{(q)}(x), ..., \phi_K^{(q)}(x)\}$, one for each cluster. At each q-th iteration, $q = 0, 1, ..., Q$, the curves are assigned to the cluster whose centroid is closest according to a specific distance. Then, the cluster means are updated based on the assignment of curves to clusters and the algorithm continues to iterate until no more curves are reassigned to clusters (Sangalli et al., 2010a). Specifically, let $P_j(\theta_i)$, $j = 1, 2, ..., J$ be J item response functions in a subset S of L^p and let $\phi_k(\theta_i)$ be the centroid of the k-th cluster, then, the k-means algorithm finds a partition of the subset S into K clusters by minimizing the sum of squared error criterion (Jain and Dubes, 1988; Tan and Witten, 2015) between the cluster center and the functions belonging to the cluster as follows:

$$J(C) = \min \sum_{k=1}^{K} \sum_{P_j(\theta_i) \in C_k} d^2\Big(P_j(\theta_i), \phi_k(\theta_i)\Big) = \min \sum_{k=1}^{K} \sum_{P_j(\theta_i) \in C_k} ||P_j(\theta_i) - \phi_k(\theta_i)||^2 \tag{9}$$

In standard applications of k-means clustering, data points in $\mathbb{R}^n$ are assigned to clusters using the minimal Euclidean distance to the cluster centers. If the data are functions, the following L^2 metric in function space may be more appropriate for clustering (Tarpey, 2007):

$$d(P_j(\theta_i), \phi_k(\theta_i)) = ||P_j(\theta_i) - \phi_k(\theta_i)||^2 = \left(\int |P_j(\theta_i) - \phi_k(\theta_i)|^2 d\,\theta \right)^{\frac{1}{2}} \tag{10}$$

Since the centroid is the representative function of a cluster, it should belong to the same function space as the item response functions (Di Battista et al., 2016). According to the standard FDA approach, functional summary statistic are computed by averaging the functions across the replications (Ramsay and Silverman, 2005). This procedure leads to a function belonging to the same subset of the functional data only in a convex space. However, under the IRT models, the subset of functions, S, is neither a vector subset nor a convex subset. In order to obtain functional centroids belonging to the same function space as the data, we refer to a one-to-one transformation, T, which converts a non-convex subset of functions, S, in a convex subset, C. Through T, it is possible to consider the functional mean in the convex subset, then, we can return it to S with the inverse transformation T^{-1}. In particular, for test data, T is expressed as the logit of the item response functions:

$$T\Big(P_j(\theta_i)\Big) = logit\Big(P_j(\theta_i)\Big) = \psi_j(\theta_i) \tag{11}$$

where $\psi_j(\theta_i)$ are the corresponding functions in the convex subset C. It is easy to verify that the centroid of the k-th cluster belong to the same function space as the item response functions. Indeed, the functional centroids with respect to T are given by:

$$\phi_{T,k}(\theta_i) = T^{-1}\left(\sum_{P_j(\theta_i)\in C_k} \frac{T\Big(P_j(\theta_i)\Big)}{J_k} \right) \tag{12}$$

where J_k is the number of item response functions in the k-th cluster, with $\sum_{k=1}^{K} J_k = J$. Taking advantage of the convexity properties, the transformation T allows us to define the functional mean in the usual way obtaining an element of S. Thus, the centroids in Equation (12) belong to the same function space as the functions. Then, the squared error criterion with respect to T can be defined as follows:

$$J_T(C) = \min \sum_{k=1}^{K} \sum_{P_j(\theta_i)\in C_k} d^2\Big(T(P_j(\theta_i)), T(\phi_k)\Big) = \min \sum_{k=1}^{K} \sum_{P_j(\theta_i)\in C_k} ||T(P_j(\theta_i)) - T(\phi_k)||^2 \tag{13}$$

and the cluster membership of the observations is determined by minimizing (13) over all the K clusters.

Moreover, since S is a parametric set of functions, the functional distance in (10) can be computed directly on the explicit form of the functions. For this reason, clustering results do not depend on how the curves are smoothed to the data, contrary to the standard FDA approach. Indeed, it is well known that, functional k-means clustering

results vary according to the method used for fitting the curves (Tarpey, 2007). Thus, in the standard FDA framework, the primary question of interest is how best to linearly transform the data prior to clustering.
Hierarchical clustering techniques may be also adopted for functional data (Ferreira and Hitchcock, 2009). In this setting, the classification strategy consists of a series of partitions, which may run from a single cluster containing all the functions (divisive methods), to J clusters, each containing a single function (agglomerative methods). In order to determine which groups should be merged (for agglomerative approach) or divided (for divisive approach), different metrics and linkage methods can be used. For functional data, the L^2 metric in (10) represents a suitable choice to define dissimilarities among functions. Concerning the linkage methods, the most common are: the single linkage (Sneath, 1957), the complete linkage (McQuitty, 1966), the average linkage (Sokal and Michener, 1958), or the Ward's minimum variance method (Ward, 1963). In particular, if we refer to the Ward's minimum variance method for agglomerative hierarchical clustering, the problem of defining a mean function that belongs to the same function space as the item response functions arises again. Indeed, this method minimizes the total within-cluster error sum of squares. Thus, at each step, the algorithm finds the pair of clusters that leads to minimum increase in total within-cluster variance after merging. This increase is a weighted squared distance among the cluster mean functions. The latter can be suitably computed through Equation (12). Ward's method presents some similarities with the k-means algorithm as it is the only one, among the agglomerative hierarchical clustering methods, that is based on a classical sum-of-squares criterion, producing groups that minimize within-group dispersion at each binary fusion. However, Ward's method uses merging of sub-clusters to achieve this goal as opposed to k-means algorithm which employs an iterative reassignment of points.

4 Application: INVALSI mathematics test for the upper secondary school students

The framework previously described is applied to a real data set drawn from annual surveys conducted by the Italian national institute for the evaluation of the school system (INVALSI). The INVALSI regularly develops standardized tests to assess Italian language and mathematics skills of students at different school grades: primary (second and fifth grade), lower secondary (sixth and eight grade), and upper secondary (tenth grade) school. In this paper, we consider the INVALSI mathematics test administered to upper secondary school pupils at the end of the scholastic year 2011-2012. These data are based on a nationally representative sample of about 42000 students. The INVALSI mathematics test consists of 54 items covering four main content domains: numbers (17 items), shapes and figures (12 items), functions and relationships (11 items) and data and previsions (14 items) (INVALSI, 2012). Several types of items are designed: multiple-choice with one correct answer and three distractors (21 items), complex multiple-choice with more true-false items (13 items), open-ended items with a unequivocal answer (13 items), and open-ended items that require students to give both a numeric answer and the adopted procedure (7 items). All type of items are dichotomously scored, by assigning 1 point to correct answer and 0 otherwise. The INVALSI mathematics test may be found on the 'test area' of the web-site www.invalsi.it.

The responses of upper secondary school students on 54 dichotomous items are analyzed applying the parametric IRT models introduced in Section 2. In this context, the observed data consist of a set of J parametric item response functions, $P_j(\theta_i)$, displaying the smooth relationship between the probability of success on an item, the item characteristics and the latent ability continuum. A preliminary analysis is conducted to test the equal discrimination across items by comparing the Rasch model and the 2PL model in Equations (2) and (3), respectively. According to the likelihood ratio (LR) test (LR=37786, with 54 degree of freedom, $p < 0.001$), the 2PL model fits the data significantly better than the Rasch model. Information criteria supported this result, as the Akaike's Information criterion (AIC) (Akaike, 1974) and the Bayesian Information criterion (BIC) (Schwarz, 1978) for the 2PL (AIC=2378763, BIC=2379696) are less than those of the Rasch model (AIC=2416441, BIC=2416908). Thus, we conclude for a significant difference in discriminating power among the items. The 3PL model should not be considered because in the INVALSI mathematics test for the upper secondary school, the presence of significant distortions due to cheating or guessing phenomena has not been detected (INVALSI, 2012). This is mainly due to the presence of an outside observer during the test administration in the sample classes. Thus, the INVALSI mathematics test data are analyzed according to a 2PL parametrization. Item parameters estimates of the 2PL model are obtained through the R package "ltm" (Rizopoulos, 2006), using marginal maximum likelihood estimation. Table 1 shows the discrimination and the difficulty parameters for each item, together with their standard errors and z-values. The items are reported in ascending order, according to the difficulty parameter.

In the IRT literature, the parameter estimates are useful descriptors of the data, as they provide a numerical summary of item characteristics. However, we point out that the analysis of the shape of each item response function through functional tools may provide further information. In particular, we aim to identify specific common patterns among the INVALSI mathematics test items applying the functional k-means algorithm introduced in Section 3. We initialized the k-means procedure by fixing three clusters ($K = 3$), as suggested by the analysis of the average silhouette values (Rousseeuw, 1987) computed for different numbers of clusters (Table 2). Then, each item of the INVALSI mathematics test is assigned to its nearest cluster center, by minimizing the sum of squared error criterion in (9). In particular, we use the one-to-one transformation T between the non-convex subset of the item response functions, S, and a convex subset C in $L^p(X)$; that is, the logit transformation in Equation (11) which yields to the corresponding item response functions, $\psi_j(\theta_i)$, in the convex subset C. Then, the functional centroid is computed in the subset C as follows:

$$T(\phi_k) = \frac{1}{J_k} \sum_{\psi_j(\theta_i) \in C_k} \psi_j(\theta_i) = \frac{1}{J_k} \sum_{\psi_j(\theta_i) \in C_k} (\alpha_{jk}\theta_i - \alpha_{jk}\beta_{jk}) = \overline{\alpha}_k \theta_i - \overline{\alpha_k \beta_k} \quad \in C \quad k = 1, 2, ..., K \tag{14}$$

Item	difficulty (β)			discrimination (α)			Item	difficulty (β)			discrimination (α)		
	Value	Std.err	z.value	Value	Std.err	z.value		Value	Std.err	z.value	Value	Std.err	z.value
M1_a	-3.77	0.10	-36.19	0.78	0.02	31.62	M10_b	0.45	0.02	28.93	0.78	0.01	60.82
M4_b	-2.83	0.08	-37.34	0.51	0.01	35.64	M15	0.63	0.01	46.42	1.04	0.01	70.64
M2_a	-2.32	0.05	-51.04	0.77	0.02	45.62	M21	0.66	0.01	53.54	1.23	0.02	75.60
M4_a	-2.07	0.04	-49.66	0.70	0.02	46.14	M13	0.73	0.01	50.04	1.00	0.01	69.11
M2_b	-1.80	0.02	-75.35	1.30	0.02	57.47	M24	0.78	0.01	77.62	1.91	0.02	80.58
M7_b	-1.70	0.03	-51.46	0.71	0.01	49.53	M22	0.86	0.03	29.93	0.48	0.01	42.21
M2_c	-1.70	0.02	-74.20	1.23	0.02	58.87	M1_c	0.86	0.02	37.64	0.62	0.01	51.28
M6_a	-1.66	0.02	-71.57	1.15	0.02	58.99	M20	0.87	0.01	70.12	1.41	0.02	77.01
M11_a	-1.63	0.02	-83.71	1.50	0.02	60.70	M9_d	0.88	0.03	33.02	0.53	0.01	45.34
M9_c	-1.11	0.02	-51.30	0.78	0.01	56.00	M5	0.92	0.02	48.46	0.81	0.01	60.82
M14_b	-1.08	0.02	-69.29	1.17	0.02	67.17	M1_b	0.95	0.03	37.80	0.59	0.01	48.98
M4_c	-0.98	0.02	-56.67	0.94	0.02	62.47	M16	1.18	0.02	63.91	1.00	0.02	66.34
M29_a	-0.93	0.02	-45.37	0.73	0.01	55.01	M3	1.28	0.02	56.61	0.82	0.01	59.87
M23_a	-0.78	0.01	-63.28	1.28	0.02	72.32	M25	1.34	0.02	67.61	1.02	0.02	65.85
M9_a	-0.64	0.04	-16.41	0.31	0.01	28.35	M11_c	1.36	0.02	80.43	1.31	0.02	70.58
M26_a	-0.50	0.01	-41.03	1.10	0.02	70.58	M12	1.44	0.02	95.09	1.70	0.02	70.90
M18	-0.46	0.01	-47.43	1.54	0.02	78.57	M14_c	1.51	0.02	75.80	1.17	0.02	66.75
M6_b	-0.46	0.01	-32.63	0.89	0.01	63.96	M6_c	1.52	0.02	92.86	1.62	0.02	69.61
M26_b	-0.36	0.01	-29.77	1.03	0.01	69.40	M2_e	1.59	0.02	98.54	1.79	0.03	68.28
M10_a	-0.30	0.01	-22.89	0.89	0.01	64.85	M23_b	1.65	0.02	76.80	1.18	0.02	65.40
M29_c	-0.24	0.02	-15.77	0.72	0.01	57.00	M9_b	1.69	0.05	35.64	0.45	0.01	38.30
M7_a	-0.16	0.01	-12.22	0.90	0.01	65.67	M11_b	1.94	0.03	60.71	0.84	0.02	55.61
M27	-0.13	0.01	-11.90	1.16	0.02	74.29	M17	1.94	0.03	74.11	1.13	0.02	61.07
M28	-0.05	0.01	-4.27	1.14	0.02	73.88	M19	2.99	0.10	30.27	0.36	0.01	30.19
M2_d	0.02	0.01	2.17	1.86	0.02	83.97	M30	3.28	0.10	33.06	0.42	0.01	32.27
M8	0.07	0.01	7.76	1.46	0.02	80.83	M14_a	4.16	0.16	26.39	0.33	0.01	25.92
M26_c	0.14	0.01	11.68	1.02	0.01	70.48	M29_b	4.69	0.62	7.59	0.08	0.01	7.72

Table 1: Item Parameters estimates of the 2PL model for the INVALSI mathematics test

Number of clusters (K)	Average silhouette values
2	0.49
3	0.54
4	0.51
5	0.47

Table 2: Average silhouette values for different numbers of K for functional k-means algorithm

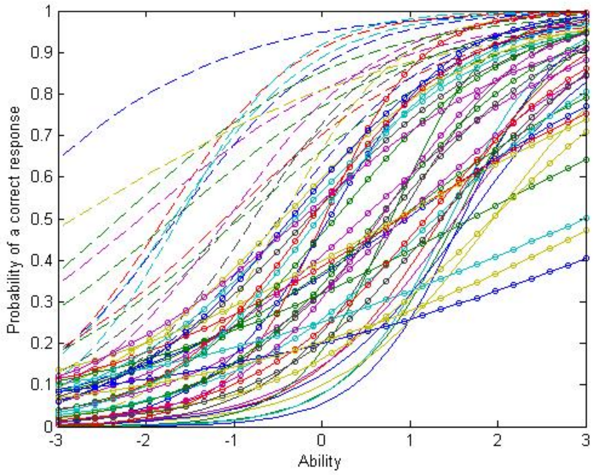


Figure 1: Classification of the $P_j(\theta_i)$ into three clusters (the first marked by dashed lines, the second by solid lines and the third one by circle-solid lines) through the functional k-means algorithm

where $\overline{\alpha}_k = \frac{1}{J_k}\sum_{j=1}^{J_k}\alpha_{jk}$; $\overline{\alpha_k\beta_k} = \frac{1}{J_k}\sum_{j=1}^{J_k}\alpha_{jk}\beta_{jk}$; and J_k is the number of functions belonging to the k-th cluster. Applying the inverse transformation of the logit in (11), the functional centroid of the $P_j(\theta_i)$, with respect to T, is obtained as follows:

$$\phi_{T,k}(\theta_i) = T^{-1}(\phi_k) = \frac{\exp\Big(T(\phi_k)\Big)}{1+\exp\Big(T(\phi_k)\Big)} = \frac{\exp(\overline{\alpha}_k\theta_i - \overline{\alpha_k\beta_k})}{1+\exp(\overline{\alpha}_k\theta_i - \overline{\alpha_k\beta_k})} \tag{15}$$

where $\phi_{T,k}(\theta_i)$ belongs to S and, thus, presents the same functional form as the observed data. Then, the cluster membership is determined by minimizing Equation (13) overall the K clusters, on the basis on the L^2 distance in (10). Figures 1 shows the partition of the subset S of item response functions into three clusters. The first group (dashed lines in Figure 1) represents the easier and less discriminating items. It is composed by 16 items (M1_a, M2_a, M2_b, M2_c, M4_a, M4_b, M4_c, M6_a, M7_b, M9_c, M11_a, M14_b, M18, M23_a, M26_a, M29_a) which are mainly open-ended type related to data and previsions content domain. The second cluster (solid lines in Figure 1) represents the most difficult and more discriminating items. It is composed by 9 items (M2_e, M6_c, M11_c, M12, M14_c, M17, M20, M23_b, M24) which are all open-ended type and are mostly related to numbers content domain. Finally, the third cluster (circle-solid lines in Figure 1) contains quite easy and quite discriminating items. It consists of 29 items (M1_b, M1_c, M2_d, M3, M5, M6_b, M7_a, M8, M9_a, M9_b, M9_d, M10_a, M10_b, M11_b, M13, M14_a, M15, M16, M19, M21, M22, M25, M26_b, M26_c, M27, M28, M29_b, M29_c, M30) which are chiefly of multiple-choice type and mainly refer to numbers and to shapes and figures content domain. The characteristics of each cluster may be captured in an immediate way by analyzing the shape of the functional centroids

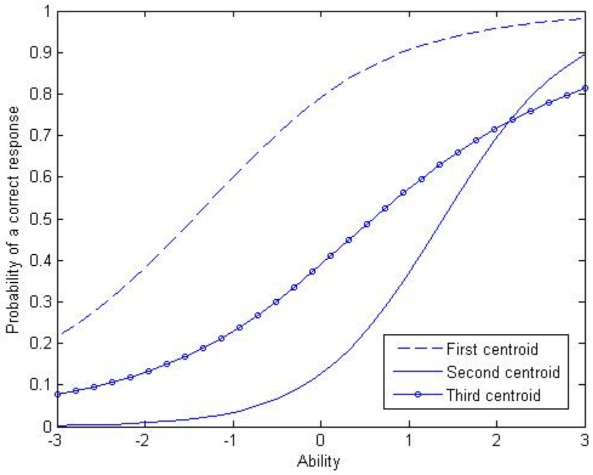


Figure 2: Functional centroids of the three clusters

in Figure 2.

The same data set is analyzed through a functional agglomerative hierarchical Ward's method. Also in this case, we consider the one-to-one transformation T in order to obtain a convex subset C of the item response functions and the cluster mean of each group is computed as in (15). As shown by the dendrogram in Figure 3, it is possible to clearly identify three clusters among the INVALSI mathematics test items. Figures 4 shows the classification of the subset S of item response functions into three clusters, according to the hierarchical method. The first group (dashed lines in Figure 4) represents the easiest and less discriminating items. It is composed by 12 items (M1_a, M2_a, M2_b, M2_c, M4_a, M4_b, M4_c, M5, M6_a, M7_a, M9_b, M10_b) which are mainly of complex multiple choice type, related to both data and previsions and function and relationship content domains. The second cluster (solid lines in Figure 1) represents difficult and very discriminating items. It is composed of 16 items (M3, M6_c, M8, M11_b, M11_c, M12, M13, M14_c, M15, M16, M17, M19, M20, M23_a, M23_b, M24) which are chiefly of open-ended type and are related especially to numbers content domain. Finally, the third cluster (circle-solid lines in the Figure 1) contains fairly easy and quite discriminating items. It consists of 26 items (M1_b, M1_c, M2_d, M2_e, M6_b, M7_b, M9_a, M9_c, M9_d, M10_a, M11_a, M14_a, M14_b, M18, M21, M22, M25, M26_a, M26_b, M26_c, M27, M28, M29_a, M29_b, M29_c, M30) which are mainly of multiple-choice type and mostly refer to number and to shapes and figures content domains. Figure 5 shows the INVALSI mathematics test items per clusters, according to the two functional clustering methods. These algorithms provide a quite similar classification. However, we highlight that the functional k-means algorithm works well because, as shown in Figure 5, it better merges in the same cluster items with similar characteristics and shapes. On the contrary, with the functional hierarchical Ward's method, there are some item response functions whose shape is clearly in contrast with the trend of the cluster.

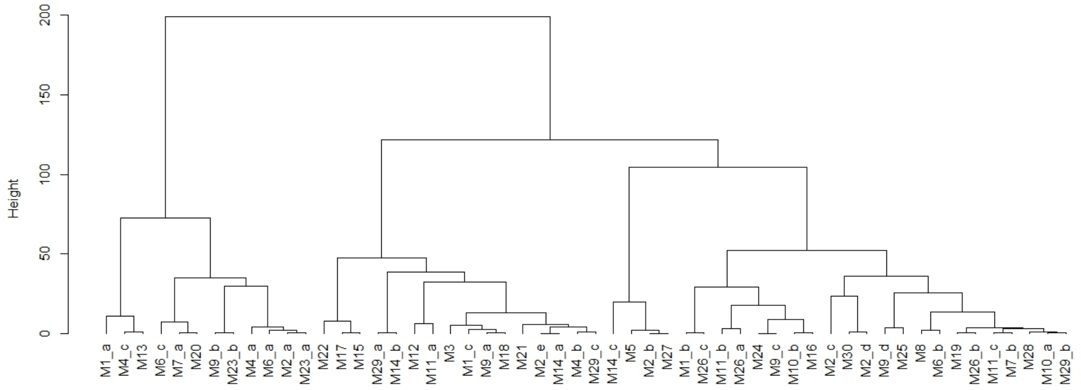


Figure 3: Cluster dendrogram for functional hierarchical Ward's method applied to the INVALSI mathematics test

Number of clusters (K)	Average silhouette values
2	0.51
3	0.40
4	0.48
5	0.41

Table 3: Average silhouette values for different numbers of K for non-functional k-means method

In order to increase the soundness of the theoretical framework and to assess the results of our method, the same data set is analyzed through a traditional (non-functional) approach for clustering test items. In particular, a non-functional k-means clustering is implemented for the INVALSI mathematics test data. A key difference lies in the choice of the number of clusters. Indeed, through a traditional k-means clustering approach, the analysis of the average silhouette values suggests to only consider two clusters (Table 3). Figures 6 shows the classification of the INVALSI mathematics test items into two clusters, according to non-functional k-means algorithm. The two clusters are both composed by 27 items. The first cluster (M1_a, M2_a, M2_b, M2_c, M2_d, M4_a, M4_b, M4_c, M6_a, M6_b, M7_a, M7_b, M8, M9_a, M9_c, M10_a, M11_a, M11_b, M18, M23_a, M26_a, M26_b, M26_c, M27, M28, M29_a, M29_c) represents items that are enough easy and less discriminating. The items are principally of multiple-choice type and refer to both data and previsions and to functions and relationships content domain. The second cluster (M1_b, M1_c, M2_c, M3, M5, M6_c, M9_b, M9_d, M10_b, M11_b, M11_c, M12, M13, M14_a, M15, M16, M17, M18, M19, M20, M21, M22, M23_b, M24, M25, M29_b,

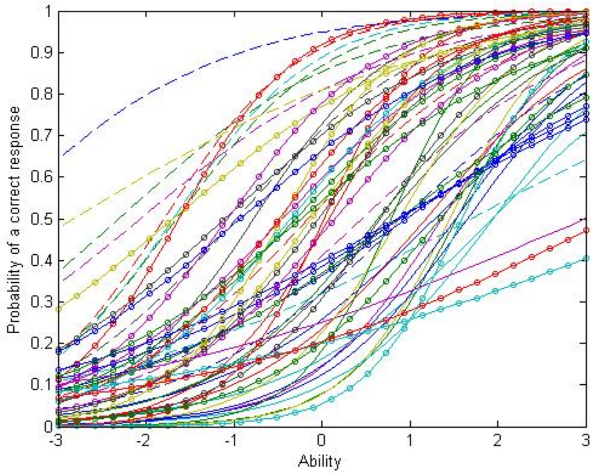
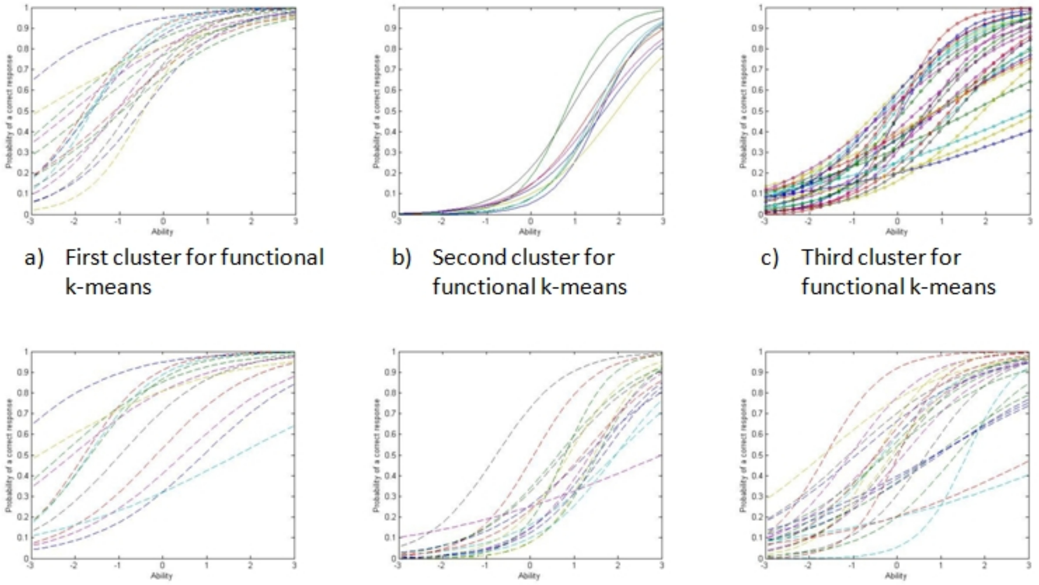


Figure 4: Classification of the $P_j(\theta_i)$ into three clusters (the first marked by dashed lines, the second by solid lines and the third one by circle-solid lines) through functional hierarchical Ward's method

M30) contains difficult and more discriminating items. The items in the cluster are especially of multiple-choice and open-ended type and refer to the numbers content domain. The traditional k-means algorithm hides a very important cluster (the third clusters in both the functional clustering algorithms) which includes quite easy and enough discriminating items. As shown in Figure 6, the latter are mixed in the first and in the second clusters; thus, the traditional k-means approach does not reveal an important feature of the INVALSI mathematics test. The Rand index (RI) (Rand, 1971) and its adjusted version (ARI) proposed by Hubert and Arabie (1985) are computed in order to provide a measure of agreement between the two partitions obtained from functional and non-functional k-means algorithms. As we expected, the values of both indices (RI=0.63; ARI=0.25) indicate a low level of agreement between the two partitions. Compared to the traditional k-means approach, our method is more suitable to highlight different kinds of items, by returning a greater degree of detail for the INVALSI mathematics test items. The reason is that the functional k-means approach takes into account the behavior of the functions for each point of the domain; whereas, a traditional k-means approach is based on just some item response function feature.

5 Concluding remarks

The aim of this paper is to develop a functional approach within the IRT context, in order to classify the items in homogeneous groups, by considering the difficulty and the discrimination of the items and the subject ability. The study of test data through the combined use of the IRT parametric specification and the functional data analysis



a) First cluster for functional k-means

b) Second cluster for functional k-means

c) Third cluster for functional k-means

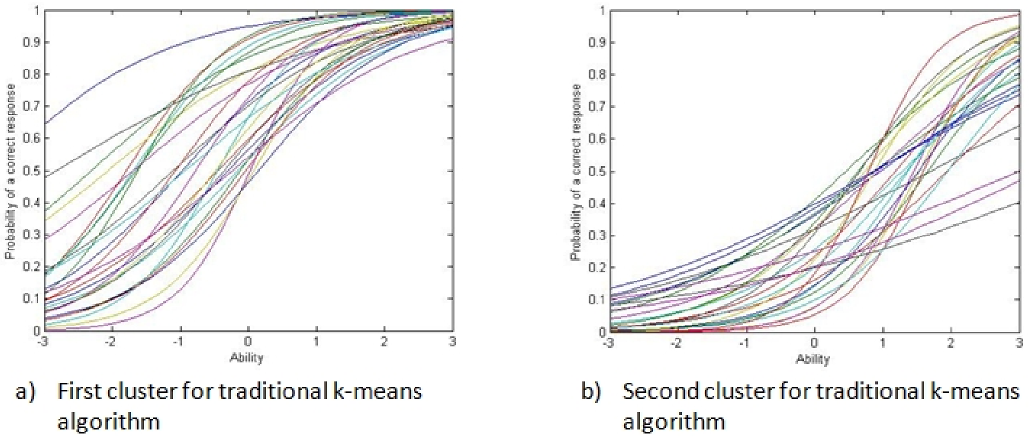
d) First cluster for functional hierarchical clustering

e) Second cluster for functional hierarchical clustering

f) Third cluster for functional hierarchical clustering

Figure 5: Classification of the $P_j(\theta_i)$ into three clusters according to functional k-means and functional hierarchical Ward's method

approach yields several advantages. Firstly, the proposed method allows to analyze the item as a function, by preserving the usual interpretation of test data in the IRT framework. Indeed we exploit the known form of the function underlying the data by avoiding the use of smoothing techniques. Secondly, since the items are analyzed as curves, the functional data analysis tools are able to evaluate the functions at each point of the domain. Thirdly, with regard to clustering algorithms, the conversion of the IRT function space to a convex one through the one-to-one transformation T, allows to compute a functional centroid for each cluster that belongs to the same function space as the item response functions. In this way, the typical properties of the mean function are ensured. We also proved that the functional cluster algorithms are more sensitive than the traditional ones to identify distinct items features. Finally, our methodological results are particularly useful in order to identify the characteristics of each cluster, evaluating the behaviour of the functional centroid.



a) First cluster for traditional k-means algorithm

b) Second cluster for traditional k-means algorithm

Figure 6: Classification of the $P_j(\theta_i)$ into two clusters according to non-functional k-means algorithm

References

Aggen, S., Neale, M., and Kendler, K. (2005). DSM criteria for major depression: evaluating symptom patterns using latent-trait item response models. *psychological medicine*, 35:475–487.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19 (6):716–723.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In Lord, F. and Novick, M., editors, *Statistical Theories of Mental Test Scores*, pages 397–479. MA: Addison-Wesle.

Bradlow, E. and Zaslavsky, A. (1999). A hierarchical latent variable model for ordinal data from a customer satisfaction survey with 'no answer' responses. *Jornal of the American Statistical Association*, 94:43–52.

Braeken, J. (2008). *Modeling Residual Dependencies in Latent Variable Models with Copulas*. Katholieke Universiteit, Leuven.

De Sanctis, A. and Di Battista, T. (2012). Functional analysis for parametric families of functional data. *International Journal of Bifurcation and Chaos*, 22 (9):1250226–1–1250226–6.

Di Battista, T., De Sanctis, A., and Fortuna, F. (2016). Clustering functional data on convex function spaces. In Di Battista, T., Moreno, E., and Racugno, W., editors, *Selected Papers of the 47th Scientific meeting of the Italian Statistical Society*, pages 101–109. Springer, In press.

Di Battista, T. and Fortuna, F. (2013). Assessing biodiversity profile through FDA. *Statistica*, 1:69–85.

Di Battista, T., Fortuna, F., and Maturo, F. (2016). Environmental monitoring through functional biodiversity tools. *Ecological Indicators*, 60:237–247.

Ferraty, F. and Vieu, P. (2006). *Nonparametric functional data analysis.* Springer, New York.

Ferreira, L. and Hitchcock, D. (2009). A comparison of hierarchical methods for clustering functional data. *Communications in Statistics - Simulation and Computation*, 38(9):1925–1949.

Forgy, E. (1965). Cluster analysis of multivariate data: efficiency vs interpretability of classifications. *Biometrics*, 21:768–769.

Gattone, S. and Di Battista, T. (2009). A functional approach to diversity profiles. *Journal of the Royal Statistical Society*, 58:267–284.

Grayson, D. (1988). Two group classification in latent trait theory: scores with monotone likelihood ratio. *Psychometrika*, 53:383–392.

Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of classification*, 2(1):193–218.

Huynh, H. (1994). A new proof for monotone likelihood ratio for the sum of independent bernoulli random variables. *Psychometrika*, 59:77–79.

INVALSI (2012). Rilevazioni nazionali sugli apprendimenti 2011-2012. Technical report, INVALSI.

Jain, A. and Dubes, R. (1988). *Algorithms for Clustering Data.* Englewood Cliffs, New York: Prentice Hall.

Lord, F. and Novick, M. (1968). *Statistical Theories of Mental Test Scores.* Addison-Wesley, Reading, MA.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability.* University of California Press.

Matthew, S. (2007). Modeling dichotomous item responses with free-knot splines. *Computational Statistics & Data Analysis*, 51:4178–4192.

McQuitty, L. (1966). Similarity analysis by reciprocal pairs for discrete and continuous data. *Educational and Psychological Measurement*, 27:21–46.

Pfeiffer, R., Bura, E., Smith, A., and Rutter, J. (2002). Two approaches to mutation detection based on functional data. *Statistics in Medicine*, 21 (22):3447–3464.

Ramsay, J. (1991). Kernel smoothing approaches to nonparametric item characteristic curve estimation. *Psychometrika*, 56:611–630.

Ramsay, J. (1997). A functional approach to modeling test data. In van der Linden, W. and Hambleton, R., editors, *Handbook of modern Item Response Theory*, pages 381–394. Springer, New York.

Ramsay, J. and Dalzell, C. (1991). Some tools for functional data analysis. *Journal of the Royal Statistical Society: Series B*, 53(3):539–572.

Ramsay, J. O. and Silverman, B. W. (2005). *Functional Data Analysis, 2nd edn.* Springer, New York.

Rand, W. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66:846–850.

Rasch, G. (1960). *Probabilistic models for some intelligence and achievement tests*. Danish Institute for educational research, Copenhagen.

Reiss, P. and Ogden, R. (2007). Functional principal component regression and functional partial least squares. *Journal of the American Statistical Association*, 102:984–996.

Rizopoulos, D. (2006). ltm: An r package for latent variable modeling and item response theory analysis. *Journal of Statistical Software*, 17 (5):1–25.

Rossi, N., Wang, X., and Ramsay, J. (2002). Nonparametric item response function estimates with the em algorithm. *Journal of Educational and Behavioral Statistics*, 27:291–317.

Rousseeuw, P. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65.

Sangalli, L., Secchi, P., Vantini, S., and Vitelli, V. (2010a). Functional clustering and alignment methods with applications. *Communications in Applied and Industrial Mathematics*, 1:205–224.

Sangalli, L., Secchi, P., Vantini, S., and Vitelli, V. (2010b). k-mean alignment for curve clustering. *Computational Statistics & Data Analysis*, 54:1219–1233.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6:461–4664.

Sharp, C., Goodyer, I., and Croudace, T. (2006). The short mood and feelings questionnaire SMFQ: a unidimensional item response theory and categorical data factor analysis of self-report ratings from a community sample of 7- through 11- year-old children. *Journal of Abnormal Child Psychology*, 34:379–391.

Sneath, P. (1957). The application of computers to taxonomy. *Journal of General Microbiology*, 17:201–226.

Sokal, R. and Michener, C. (1958). A statistical method for evaluating systematic relationships. *University of Kansas Science Bulletin*, 38:1409–1438.

Tan, K. and Witten, D. (2015). Statistical properties of convex clustering. *Electronic Journal of Statistics*, 9:2324–2347.

Tarpey, T. (2007). Linear transformations and the k-means clustering algorithm: applications to clustering curves. *The American Statistician*, 61(1):34–40.

Valentini, P., Di Battista, T., and Gattone, S. (2011). Heterogeneneity measures in customer satisfaction analysis. *Jornal of classifications*, 28:38–52.

Ward, J. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58:236–244.

Weiss, D. (1995). Improving individual differences measurement with item response theory and computerized adaptive testing. In Lubinski, D. and Dawis, R., editors, *Assessing individual differences in human behavior: New concepts, methods, and findings*, pages 19–79. Davies-Black Publishing, Palo Alto, CA.