Soc Indic Res (2019) 146:19–39
https://doi.org/10.1007/s11205-018-1893-4

Testing Equality of Functions Across Multiple Experimental Conditions for Different Ability Levels in the IRT Context: The Case of the IPRASE TLT 2016 Survey

Fabrizio Maturo[1] · Francesca Fortuna[2] · Tonio Di Battista[2]

Accepted: 28 March 2018 / Published online: 2 April 2018

Abstract In the educational field, it is common to analyze test data through item response theory models. In this context, a key role is played by item characteristic curves (*ICC*s) and item information curves (*IIC*s). In many real cases, practitioners are interested in understanding if some factors have a significant influence on the probability of correctly answering items. In the literature, this problem has been addressed by applying the standard analysis of variance model, which is based on the total scores or the proportion of correct responses. However, this method needs to meet some strong assumptions and may present some limitations because it does not consider useful information typical of the IRT, such as the shapes of the *ICC*s and *IIC*s, which provide interesting insights for different ability levels. To overcome these issues, this research suggests the use of the functional analysis of variance approach and a novel functional tool in the IRT context. The main advantages of this approach are that it is distribution-free and allows us to check the degree of consistency with the hypothesis of equality among mean curves for different ability levels. Specifically, the proposed method is applied on *ICC*s and *IIC*s for improving the existing techniques in the educational studies. A real dataset drawn from the IPRASE Trentino Language Testing Survey 2016 is considered. The final purpose of this study is to provide additional tools for scholars and practitioners in defining specific educational plans.

Keywords IRT · *ICC* · *IIC* · FANOVA · *P*-Statistic

✉ Fabrizio Maturo
f.maturo@unich.it

Francesca Fortuna
francesca.fortuna@unich.it

Tonio Di Battista
dibattis@unich.it

[1] Department of Management and Business Administration, "G. d' Annunzio" University, Pescara, Italy

[2] DISFPEQ, "G. d' Annunzio" University, Pescara, Italy

1 Introduction

In the global economy, foreign languages proficiency has become an increasingly important issue. The ability to understand and communicate in other languages is recognized as a driver of economic growth, wealth, and competition because it contributes to the cultural richness of our society, increases personal fulfilment, and encourages global citizenship. As a consequence, the promotion of foreign language instruction represents an integral part of learning process and a considerable effort is made in order to redefine educational systems in the modern global world. In this context, the Council of Europe has developed the Common European Framework of Reference (CEFR) for the language learning, teaching, and assessment (Council of Europe 2011). This document promotes methodological innovations for designing teaching programmes, encourages the development of a communicative approach, and enables comparison across educational systems. Obviously, the revision of foreign languages teaching policies raises the need of assessing the level of students competencies through periodic standardized tests. However, in Italy, such information is not yet available. This aspect highlights the gap between foreign language and other disciplines, such as maths and Italian language, which are subject to periodic national evaluations. Thus, it seems necessary to foster initiatives that provide valid, reliable, and accurate information on students skills, as is the case for other disciplines. For this reason, in 2016, the Autonomous Province of Trento (in Northern Italy) has approved the “Trentino Trilingual Plan 2016–2020” to improve the knowledge of English and German languages in the provincial schools (Covi and Dutto 2017). This plan defines learning objectives, modalities, and application tools for the provincial school system. The research centre IPRASE (Provincial Institute for Experimental Research and Educational Research) has been designated for monitoring the implementation of the strategic plan. At this purpose, the IPRASE has developed the IPRASE Trentino language testing (TLT) 2016 for assessing students language skills in German and English. The TLT 2016 represents the first systematic initiative at the provincial level for evaluating foreign language skills and places the Trentino educational system within a common international framework. Indeed, it is a commonplace to claim the importance of assessment in language teaching and learning. Teachers need to know learner acquired skills in order to individuate their strength and weakness, and plan their instruction appropriately.

In the educational field, tests for students assessment are analyzed through item response theory (IRT) models. They refer to a family of latent trait 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 1968a). This relation is given by the item characteristic curve (*ICC*), which shows how the probability of success on a test item changes as the level of the latent trait increases or decreases. Thus, the probability of a correct response to an item is modeled as a continuous function of the item parameters and the subject ability. In most parametric families of IRT models, it is assumed that the *ICC* is included in a restricted class of functions defined by specific mathematical models, such as the logistic function. IRT models allow to examine latent constructs in a rigorous way, providing more accurate and consistent results compared to traditional measurement instruments, such as the Classical Test Theory (Ayala 2009). Indeed, IRT offers a powerful framework for understanding and optimizing tests in that they reveal how items and tests perform at all levels of the latent trait (O’Connor et al. 2015; Chen et al. 2013; Ceccatelli et al. 2013).

In the IRT context, another source of information is given by the item information curve (*IIC*), which reveals the amount of information that an item provides at a particular value of the latent trait. *IIC*s play an important role in test development and item evaluation because they may highlight ability levels that are inadequately covered by the items.

Generally, tests assessment through IRT models focuses on the analysis of item parameters (i.e. difficulty, discriminatively, and so on). However, because *ICC*s and *IIC*s are functions in a continuous domain (the subject ability), it is possible to analyze them through the functional data analysis (FDA) approach (Ramsay and Silverman 2005; Ferraty and Vieu 2006). In this context, most of the proposed approaches recommend to estimate the shape of the item characteristic curve with non-parametric techniques, such as Kernel regression (Ramsay 1991) or B-splines expansion (Rossi et al. 2002; Matthew 2007), without prior assumptions about the *ICC* functional form. On the contrary, Battista and Fortuna (2016) have suggested an alternative method that combines the parametric specification of the common IRT with the FDA framework. This approach allows to preserve the usual interpretation of item parameters but takes advantage of functional data analysis tools. Indeed, the FDA approach grants a deeper analysis of those phenomena varying in a fixed domain, and, in an educational framework, it enables the evaluation of items behavior for all levels of the latent trait. For example, the graphical inspection of *ICC*s and *IIC*s plays a key role in the IRT framework because it might show problematic items, which should be revised or excluded from the test. However, this procedure is not always so straightforward. In this context, the FDA approach may be helpful to improve the graphical interpretation by analyzing the behaviour of *ICC*s and *IIC*s with additional functional tools (such as integrals, derivatives, and so on; Fortuna and Maturo 2018). Moreover, functional versions of classical statistical methods can be extended to IRT without losing information about different ability levels. For example, a functional k-means algorithm could allow to identify a set of homogeneous clusters among the items by jointly consider items and students properties over the whole domain (Battista and Fortuna 2016). Moreover, functional principal component analysis can be used to explore the variations across items in the shapes of the *ICC*s (Rossi et al. 2002), and functional linear discriminant analysis may be useful to study differential item functioning (Liu 2016) for each ability levels. Thus, the FDA approach can be viewed as a powerful tool to improve the existing techniques in educational studies. This aspect is particularly evident when one wishes to test for a treatment effect on the probability of correctly answering an item. This problem is usually dealt with a standard one-way analysis of variance (ANOVA). However, the applicability of such method is not always straightforward in the educational context because we consider a latent trait, whereas ANOVA uses an observable outcome variable. Indeed, the ANOVA model has been applied to the total scores or proportion of correct responses, which are derived from classical test theory (Carpita 2017). The FDA approach provides a useful solution to this problem because it grants to extend ANOVA to the IRT context. Indeed, functional analysis of variance (FANOVA) allows to test the possible differences in responses according to the treatments used, considering that the data are curves. Moreover, this method permits to analyze possible treatment effects for different levels of students ability by providing the instructor with useful directions for defining specific educational plans. In this study, we extend the FANOVA model (Ramsay and Silverman 2005) to the IRT context; specifically, we test the equality of *ICC*s and *IIC*s across multiple experimental conditions by combining the classical parametric specification of IRT models and the FDA approach (Battista and Fortuna 2016). Hence, we propose the FANOVA approach for a parametric model with a functional form that is known in advance, preserving the usual interpretation of item parameters.

In this context, a new functional measure is proposed for evaluating the degree of consistency with the null hypothesis for different ability levels: the functional *P*-statistic (functional ratio).

The proposed functional approach is applied to a real dataset drawn from the IPRASE Trentino Language Testing (TLT) Survey 2016. The test has been administered with four different versions (QSET-Question set) whose difficulty has been recognized by a team of experts as equivalent from a linguistic point of view. With reference to these data, our purpose is to check if the factor QSET (with treatments given by the different versions of the test) has a statistical significant effect on the functional response variables *ICC*s and *IIC*s. The remainder of this paper is organized as follows. Section 2 illustrates the materials and methods. In particular, it begins with the analysis of standard parametric IRT models in a functional framework and continues with a brief overview of the FANOVA model. In addition, it provides the proposed functional tool for assessing the degree of consistency with the null hypothesis of equality of mean functions. In Sect. 3, we illustrate the main results obtained by applying the proposed approach to test data drawn from the TLT 2016. Finally, Sect. 4 presents the discussion and conclusions of this study.

2 Materials and Methods

2.1 Functional Parametric IRT Models

IRT encompasses a variety of statistical methods to study a latent trait (usually the subject ability) associated with a set of categorically scored items in a test (Lord and Novick 1968b). The core of IRT is the *ICC*, which provides a mathematical model to explain the relationship among item characteristics, subject ability and the probability of a correct response. The basic representation of IRT models for dichotomous items (Lord and Novick 1968b) is given by:

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

where X represents the observed response on the test item, with $X = 1$ representing the correct responses and $X = 0$ the incorrect one; η is a vector of parameters, which denotes the characteristics of the item; θ represents a proficiency parameter for the subject; and f is a function which defines the relationship among the item parameters, the subject ability and the probability of a correct response. Different IRT models arise from different functional forms assumed for f, and different number of item parameters. The most common parametric IRT models define f in Eq. 1 as a logistic function and specify one, two, or three item parameters, giving rise to the so-called one (1PL), two (2PL), and three (3PL) parametric logistic models, respectively. The 3PL model (Birnbaum 1968) represents the general form and specifies the *ICC* of the j-th item ($j = 1, \dots, J$) as follows:

$$ICC_j(\theta) = P(X_{ij} = 1|\theta_i, \gamma_j, \alpha_j, \beta_j) = \gamma_j + (1 - \gamma_j)\frac{\exp[\alpha_j(\theta_i - \beta_j)]}{1 + \exp[\alpha_j(\theta_i - \beta_j)]} \tag{2}$$

where X_{ij} denotes the observed response of the i-th subject ($i = 1, \dots, n$) to the j-th item ($j = 1, \dots, J$), θ_i indicates the ability of the i-th subject, while γ_j, α_j and β_j are the pseudo-guessing, the discrimination, and the difficulty parameter of the j-th item, respectively. For a detailed explanation of the item parameters see Lord and Novick (1968b) and Birnbaum (1968).

The 1PL and 2PL models can be considered as special cases of Eq. 2 (Lord 1980). Indeed, the 2PL model (Birnbaum 1968) can be obtained by fixing γ_j to 0; while the 1PL model or Rasch model (Rasch 1960) can be derived by additionally fixing α_j to 1. The basic assumptions behind the above models are that the probability of a correct response increases as the ability of the examinees increases (monotonicity), the items are indicators of a single continuous latent variable (unidimensionality), and the item responses are distributed independently, given the value of the latent trait (local independence).

Because, in the IRT framework, test data are expressed as curves, the FDA approach (Ramsay and Silverman 2005) can be adopted. In particular, this study refers to the approach proposed by Battista and Fortuna (2016) that extends parametric IRT models by using some techniques of the FDA framework. In this context, the observed data consist of a set of item response functions, $ICC_j(\theta_i)$, $j = 1, 2, \dots, J$, one per test item, that are expressed by a specific parametric model. Thus, *ICC*s belong to a function space, say S, with m real parameters, that is:

$$S = \{f(\boldsymbol{\eta};\theta)\} \tag{3}$$

where f is a parametric function; $\boldsymbol{\eta} = (\eta_1, \eta_2, \dots, \eta_m)'$ represents the item parameter vector; θ is the domain of the function; and S is a subset of some L^p space. The peculiarity of test data is that, under the IRT models, the functional form of the *ICC* is known in advance. For this reason, the approximation of the function underlying the data through smoothing techniques can be avoided (Battista and Fortuna 2016, 2017; Battista et al. 2017; Maturo et al. 2016, 2018; Maturo and Di Battista 2018; Maturo 2018). The study of test data through the combined use of the IRT parametric specification and the FDA approach yields several advantages. Firstly, it allows to preserve the usual interpretation of test data in the IRT framework by exhibiting desirable statistical and mathematical properties of their parametric specification. Secondly, it allows to analyze the item as a function, providing more accurate information about the data. Indeed, the main advantage of the FDA approach is the analysis of curve characteristics with functional tools. In this way, each item can be studied through the shape of the item characteristic curve, which can be evaluated at each point of the latent trait. Finally, contrary to the standard FDA approach, the use of the *ICC* parametric form avoids that the results vary according to the method used for fitting the curves.

This approach can be easily extended to *IIC*. The latter reveals the amount of information that an item provides at a particular value of the latent trait and can be viewed as a statistical indicator of the quality of ability estimation. Specifically, the *IIC* for the j-th item can be defined as follows (Hambleton and Linden 1997):

$$IIC_j(\theta) = \frac{\left[ICC_j'(\theta)\right]^2}{ICC_j(\theta)\left(1 - ICC_j(\theta)\right)} \tag{4}$$

where $ICC_j'(\theta)$ represents the derivative of $ICC_j(\theta)$ with respect to θ. Because *IIC* provides information for all values of θ, it is usually represented as a function of the latent trait rather than as a single value. *IIC*s play an important role in test development and item evaluation. Indeed, item selection for test construction is based on the analysis of *IIC*s which might reveal poorly informative items, and thus suggest to substituting other questions in their place. On the other hand, *IIC*s can also show areas in the θ-scale that are inadequately covered, and which would benefit by the addition of extra test items.

Although the IRT parametric specification yields several advantages, it is important to highlight that it entails the risk of not accounting for unusual features of the items, such as the non-monotonicity and other systematic shape departures that can not be accommodated in existing models (Ramsay 1991). However, if the parametric IRT models are able to capture the characteristics of test data, then they represent a suitable choice due to their statistical and mathematical properties (Ramsay 1997) and their easy interpretation for practitioners.

2.2 The FANOVA Model

The functional version of the ANOVA model, called FANOVA, focuses on cases where the independent variable is a scalar and the response variable is a curve, which is decomposed into contributions of the overall mean and the main effects, which are both functional (Ramsay and Silverman 2005; Battista et al. 2016; Maturo et al. 2017; Di Battista et al. 2014). Our research extends this method to the IRT context in order to understand if some factors have significant influence on *ICC*s and *IIC*s. As for the FDA approach, the first step is to reconstruct the functional form of the data. To this purpose, the most common methods are the parametric approach of Ramsay and Silverman (2005) and the non-parametric one of Ferraty and Vieu (2006). However, in the IRT context, *ICC*s and *IIC*s can be expressed by parametric models known in advance, such as the 3PL model in Eq. 2. The estimation of this class of models is usually performed using the marginal maximum likelihood estimation (Bock and Aitkin 1981), which allows us evaluating the goodness of fit of the model. Once the raw-data are expressed as functions, we can adopt the FANOVA model following the parametric point-wise approach of Ramsay and Silverman (2005).

Assuming that there is a single factor with K different levels ($k = 1, \dots, K$), the FANOVA model for the j-th item ($j = 1, \dots, J$) in the k-th treatment can be expressed as follows:

$$\boldsymbol{\Phi}_{jk}(\theta) = \mu(\theta) + a_k(\theta) + \epsilon_{jk}(\theta) \quad j = 1, \dots, J;\ k = 1, \dots, K \tag{5}$$

where $\boldsymbol{\Phi}_{jk}(\theta)$ can be the *ICC* or the *IIC* of the j-th item in the k-th treatment; $\mu(\theta)$ is the average function across all treatments; $a_k(\theta)$ represents the specific functional effect of being in the k-th treatment, with:

$$\sum_{k=1}^{K} a_k = 0 \ \ \forall \theta \in [\theta_{min}, \theta_{max}] \tag{6}$$

for their unique identification; and $\epsilon_{ij}(\theta)$ is the residual error function, which expresses the unexplained variation of the j-th item within the k-th treatment. The FANOVA model in Eq. 5 can be expressed in matrix notation as follows:

$$\boldsymbol{\Phi}(\theta) = \mathbf{Z}\boldsymbol{\gamma}(\theta) + \boldsymbol{\epsilon}(\theta) \tag{7}$$

where $\boldsymbol{\Phi}(\theta)$ is a vector of J functional observations; $\boldsymbol{\gamma}(\theta) = [\gamma_1(\theta) = \mu(\theta), \gamma_2(\theta) = a_1(\theta), \dots, \gamma_{K+1}(\theta) = a_K(\theta)]'$ is a vector of $K + 1$ parameter functions; $\mathbf{Z}$ is a $(J, K + 1)$ design matrix, similar to that used in traditional ANOVA for coding group membership; and $\boldsymbol{\epsilon}(\theta)$ is a vector of J residual functions. The parameter vector $\boldsymbol{\gamma}(\theta)$ can be estimated using the standard least squares criterion, that is by minimizing the residual sum of squares:

$$LMSSE(\boldsymbol{\gamma}) = \int \left[\boldsymbol{\Phi}(\theta) - \mathbf{Z}\boldsymbol{\gamma}(\theta)\right]' \left[\boldsymbol{\Phi}(\theta) - \mathbf{Z}\boldsymbol{\gamma}(\theta)\right] d\theta \tag{8}$$

subject to the constraint (6). It is possible to minimize the discrete version of Eq. 8, that is: $||\boldsymbol{\Phi}(\theta) - \boldsymbol{Z\gamma}(\theta)||^2$, individually for each point of the domain, leading to the pointwise least squares estimates of the functional parameters (Ramsay and Silverman 2005):

$$\hat{\boldsymbol{\gamma}}(\theta) = (\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T\boldsymbol{\Phi}(\theta) \tag{9}$$

Thus, we calculate $\hat{\boldsymbol{\gamma}}(\theta)$ for a suitable grid of values of θ using ordinary regression analysis, and then interpolate between these fitted values generating continuous estimates.

Within the FANOVA framework, one wants to test for a difference in mean curves from K groups anywhere in θ. The FANOVA testing problem can be expressed as follows:

$$\begin{aligned} H_0 &: a_1(\theta) = a_2(\theta) = \cdots = a_K(\theta) \\ H_1 &: a_i(\theta) \neq a_{i'}(\theta) \quad \textit{for at least one } \theta \quad \textit{and } i \neq i' \end{aligned} \tag{10}$$

To solve this problem, a pointwise F statistic can be used (Ramsay and Silverman 2005):

$$F_{obs}(\theta) = \frac{\left[SSY(\theta) - SSE(\theta)\right]/(K-1)}{SSE(\theta)/(J-K)} \tag{11}$$

where $SSY(\theta)$ represents the total sums of squares function:

$$SSY(\theta) = \sum_{j=1}^{J}\sum_{k=1}^{K}\left[\boldsymbol{\Phi}_{jk}(\theta) - \hat{\mu}(\theta)\right]^2 \tag{12}$$

where $\hat{\mu}(\theta)$ is an estimate of the overall mean function. $SSE(\theta)$ is the error sums of squares function:

$$SSE(\theta) = \sum_{j=1}^{J}\sum_{k=1}^{K}\left[\boldsymbol{\Phi}_{jk}(\theta) - \hat{\boldsymbol{\Phi}}_{jk}(\theta)\right]^2 \tag{13}$$

where $\hat{\boldsymbol{\Phi}}_{jk}(\theta)$ is the predicted *ICC* or *IIC* value for the *j*-th item in the *k*-th treatment from a fitted FANOVA model such as that in Eq. 7.

Equation 11 provides the observed F-statistic function; that is a function built by calculating the Fisher test statistic at each point of the domain. Thus, in the IRT framework, $F_{obs}(\theta)$ is a function of θ and varies over the whole domain, unlike classical statistical tests.

In many circumstances (e.g. in a functional context) the null distribution of the test statistic is unknown, and thus a permutation test can be adopted. The latter is an alternative way to test for differences in population means in a non-parametric fashion; hence, we do not need to make specific assumptions about the sampling distribution of the test statistic. The basic idea is to consider the sampling distribution of some test statistic under all possible permutations of the labels of the observed data. Specifically, in a functional context, this method is based on randomly shuffling curve labels and calculating the test statistic each time, obtaining $F_{perm}(\theta)$. This process is repeated several times, say *nperm*, in order to obtain the sampling distribution of the test statistic when the null hypothesis is true. To perform inference across each point of the domain, we can compare $F_{obs}(\theta)$ in Eq. 11 with the permutation α-level critical value (Ramsay and Silverman 2005). To find the pointwise 0.05 critical value of the null distribution, at each point of the domain, we calculate the 95th percentile of the permuted F-statistic. Calculating the 95th percentile of the maxima of the *nperm* permutations, the maximum 0.05 critical value of the null distribution can be used as a reference to test the null hypothesis overall the domain (Ramsay and Silverman 2005).

2.3 The Functional *P*-Statistic (Functional Ratio) for Evaluating the Degree of Consistency with the Null Hypothesis for Different Ability Levels

Following the classical FANOVA approach (Ramsay and Silverman 2005), we obtain the plot of $F_{obs}(\theta)$, which allows us to test if there are any significant differences among groups. However, given the specificity of the IRT context, practitioners and scholars may be interested in understanding how much the test is poorly formulated for different ability intervals. For this reason, we do not focus on an overall test of the null hypothesis, but we analyze different domain intervals for finding a measure of the degree of consistency with H_0 for different ability levels. Specifically, when the hypothesis testing provides conflicting results for different intervals (i.e. H_0 is supported only in some parts of the domain), this approach may help in the questionnaire assessment according to different evaluation purposes. Therefore, we propose a functional tools for facilitating the detection of those ability levels for which the degree of consistency/inconsistency with H_0 is relevant.

Certainly, a first rough measure of the degree of consistency with H_0 is the functional F-statistic. Indeed, if $F_{obs}(\theta)$ is high, then we have evidence that the means are different (against H_0), and viceversa. This should be a good clue but it is not a reliable measure because the value of $F_{obs}(\theta)$ should be compared with the critical value. Thus, to introduce a measure for assessing the degree of consistency with the null hypothesis, we propose the functional *P*-statistic, which is based on the functional *p* value. The significance level of a statistic is the proportion of values that are as extreme as, or more extreme than the test statistic in the reference distribution, which is either obtained by permutations or found in a table of the appropriate statistical distribution. The level of significance should be regarded as the strength of evidence against the null hypothesis (Manly 1997). In the context of permutation-tests, the *p* value is the proportion of permutations that are greater or equal to F_{obs}. Thus in our context, the *P*-statistic is the following:

$$P_\theta = \frac{\#(F_{obs_\theta} \geq F_{perm_\theta})}{nperm} \quad \theta \in [-4, 4] \tag{14}$$

where θ is a generic point of the domain, and *nperm* is the total number of permutations. Computing Eq. 14, we obtain the functional *p* value. The plot of P_θ could be directly compared with the critical value α to understand, if for which ability level, there is (or not) enough evidence against H_0.

For each θ, the *p* value is an exact measure of the degree of consistency with the null hypothesis. Because it may happen that the graph of P_θ intersects the critical value in some θ intervals, we introduce this functional tool for understanding if, on average, the amount of information against H_0 is greater than that in favour of H_0. Hence, the following functional statistic, say $P(\theta_m, \theta_n)$, may be defined as follows:

$$P(\theta_m, \theta_n) = \frac{\int_{\theta_m}^{\theta_n} P(\theta) d\,\theta}{\int_{\theta_m}^{\theta_n} \alpha d\,\theta} \tag{15}$$

where the denominator of the fraction is introduced to emphasize the p value magnitude relative to the significance level α. Following, we consider $\alpha = 0.05$ and the ability domain $\theta \in [-4, 4]$ divided into four equally-spaced intervals: $P(-4, -2)$, $P(-2, 0)$, $P(0, 2)$, and finally $P(2, 4)$, for low, middle–low, middle–high, and high ability levels, respectively.

Computing the P-statistic, two circumstances may occur:

- $P(\theta_m, \theta_n) \geqslant 1$ if $P(\theta)$ is, on average higher or equal to α in the interval $[\theta_m, \theta_n]$; thus, for this ability interval there is not enough evidence against H_0;
- $P(\theta_m, \theta_n) < 1$ if $P(\theta)$ is, on average lower than α in the interval $[\theta_m, \theta_n]$; hence, for this ability level, there is enough evidence against H_0.

Consequently, the higher is $P(\theta_m, \theta_n)$, the more there is not enough evidence against H_0, and equality among group means can be declared. The P-statistic, being a ratio, emphasizes this inconsistency (or consistency) with the null hypothesis because it is characterized by a multiplicative effect. Hence, we can use it to compare the degree of coherence with the null hypothesis for different intervals of ability.

3 Application: Trentino Language Testing 2016

The proposed functional approach is applied to a real dataset, which has been drawn from the IPRASE Trentino Language Testing (TLT) Survey 2016. The latter is part of the “Trentino Trilungual Plan 2016–2020”, which is promoted by the Autonomous Province of Trento (Northern Italy) for improving the knowledge of English and German languages in the provincial schools (Covi and Dutto 2017). The questionnaire was developed by the research centre IPRASE (Provincial Institute for Experimental Research and Educational Research) by referring to the levels of the Common European Framework of Reference (CEFR) for language skills assessment (Council of Europe 2011). The tests has been administered in the spring of 2016 and involved pupils in the last year of primary and lower secondary school, and in the second year of the upper secondary school, for a total of about 3000 students. In this paper, we focus on the available data, i.e. a sample of 300 students of the last year of first grade secondary school that has responded to the English test. The latter consists of 66 items, which cover three main language skills: listening (A); reading (L); and formal skills (F). Each of them is composed of subgroups of questions: two sets for listening and reading (A1, A2, L1 and L2, respectively) and one set for formal skills (F1). In order to reduce the possibility of cheating, each set of items has been administered in four different versions (QSET-Question set), whose difficulty has been evaluated by a group of experts as perfectly similar from a linguistic point of view. With reference to these data, our interest is not to evaluate students performance or to check the difficulty or the discriminating power of the items, but rather to verify if the use of different QSETs has significantly influenced the average probability of correctly response to items (*ICC*s) and also the average amount of information that items provide (*IIC*s). Specifically, the main purpose of this study is to check the effect of different QSETs for various ability levels. Therefore, the FANOVA model (see Sect. 2.2) has been applied to the English test. For the sake of brevity, in this paper, we focus only on the two subgroups of items of the listening section (A1 and A2 tests). In particular, we investigate the possible effect of the four different QSET versions of the questionnaire, i.e. QSETs 46, 47, 48, and 49 for the A1 test, and QSETs 92, 93, 94, and 95 for the A2 test. *ICC*s and *IIC*s for both tests have been obtained by considering a 2PL IRT model (Birnbaum 1968) carried out through the

R package “*ltm*” (Rizopoulos 2006), using marginal maximum likelihood estimation (Bock and Aitkin 1981).

A preliminary analysis has been conducted to compare the performance of the Rasch model and the 2PL model, for each of the eight IRT models (4 subgroups for test A1 and 4 subgroups for test A2). According to the likelihood ratio (LR) tests, the 2PL models fit the data significantly better than the Rasch models. Information criteria supported these results, as the Akaike’s Information criterion (AIC) (Akaike 1974) and the Bayesian Information criterion (BIC) (Schwarz 1978) for the 2PL are lower than those of the Rasch models. Thus, we conclude for a significant difference in discriminating power among the items. Hence, test data are analyzed according to the 2PL parametrization. However, we underline that our approach can be extended to other IRT models.

Moreover, to check the fit of the models to the data, the “*margins()*” function of the “*ltm*” R package has been used, adopting as a rule of thumb the value 3.5 (Rizopoulos 2006); particularly, it consists in constructing the 2×2 contingency tables obtained by taking the variables two at a time, and then comparing the observed and expected two-way margins using the so called Chi-squared residuals. As stressed by Rizopoulos (2006), this method is analogous to comparing the observed and expected correlations when judging the fit of a factor analysis model, and moreover it can be extended to the three-way margins. According to our results, the 2PL models present a good fit to the data (values lower than 3.5).

3.1 Application to A1 Test

Figure 1 shows the *ICC*s for each QSET of the A1 test, and also the four average *ICC*s. Figure 2a illustrates the functional effects of each QSET. The plot highlights that the QSET 49 has a positive effect on the probability of correctly answer for each level of ability. Conversely, the QSET 46 appears to be the more difficult over the whole domain. The other two QSETs have different functional effects according to the levels of ability. Hence, we perform the FANOVA model for understanding if this first evidence of difference among groups is statistically significant.

Figure 2b displays the permutation functional F-test of the FANOVA model based on the *ICC*s of the A1 test. The observed functional F-statistic lies below both the 0.05 maximum and point-wise critical values over the whole domain. In this case, there is not enough evidence against the null hypothesis of equality of group means; thus the use of different QSETs has not a significant effect on *ICC*s. Nevertheless, we apply the functional P-statistic proposed in Sect. 2.3 to check its coherence with the evidence of Fig. 2b. Figure 2c shows the plot of the functional p value whereas Fig. 2d displays the functional P-statistic, which is computed using Eq. 15 and whose values are shown in Table 1. The functional p value is always higher than the 0.05 line, and the P-statistic is greater than 1 over the whole domain. In particular, for middle–low abilities, we obtain a high degree of consistency with H_0. Thus, we can state that the four QSETs are equally difficult according to the probability of correctly answer the items; however, the values of the functional instrument confirms that this statement is particularly true for middle–low abilities.

The second FANOVA is performed on the *IIC*s of the A1 test for checking the possible effect of QSET on the level of information that items provide. Figure 3 shows the *IIC*s and their means for different QSETs whereas Fig. 4 displays the functional effect of being in each of the four QSETs, functional F-test, functional p value, and functional P-statistic. We observe that the QSET 49 has a strong information power for

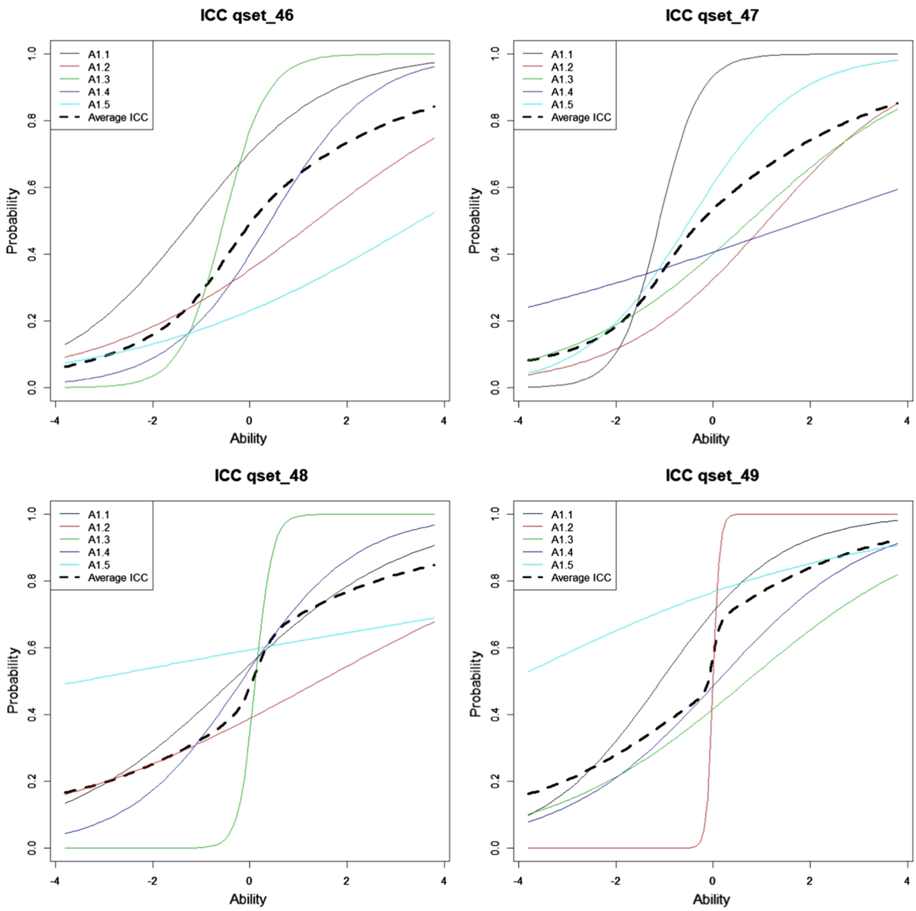

Fig. 1 ICCs and their average for the four QSETs of the A1 test

middle abilities due to the high peak of the item A1.2. In summary, the QSETs 46 and 47 have a negative impact for middle abilities whereas the QSET 48 varies between positive and negative effects for different ability levels.

The analysis of Fig. 4 and Table 2 confirms that, also for the *IIC*s, there is not enough evidence against H_0 for each ability level. Thus, there are not significant differences among group means according to their informative power. Indeed, the P-statistic is always greater than 1, and $F_{obs}(\theta)$ is always below the critical level α. Therefore, we can state that the four QSETs are similar according to their information power; and this statement is particularly true for low abilities (P-statistic = 14.12).

In summary, the experts have done a good job in preparing the four QSETs with similarly difficult and informative items for the A1 test.

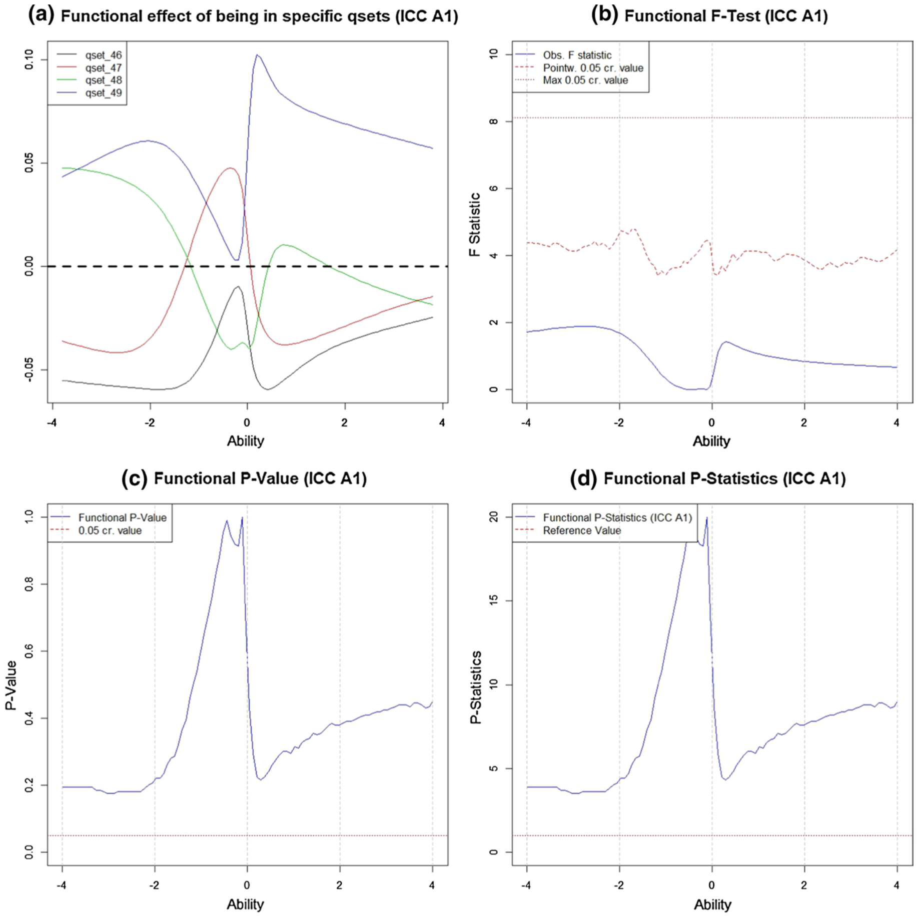

Fig. 2 F-test and *P*-statistic for *ICC*s of the A1 test

Table 1 Functional *P*-statistic on FANOVA model of *ICC*s of the A1 test

Abilities	*P*-Statistic	Evidence
Low	$5.24 > 1$	Not enough evidence against H_0
Middle–low	$12.59 > 1$	Not enough evidence against H_0
Middle–high	$7.55 > 1$	Not enough evidence against H_0
High	$9.52 > 1$	Not enough evidence against H_0

3.2 Application to A2 Test

The same analysis is performed on the A2 test. Figure 5 shows the *ICC*s and their means for the QSETs 92, 93, 94, and 95 of the A2 test.

Figure 6a displays the functional effects of each QSET. The most difficult QSET is the number 93 whereas the most easy is the number 95, which is positive over the whole

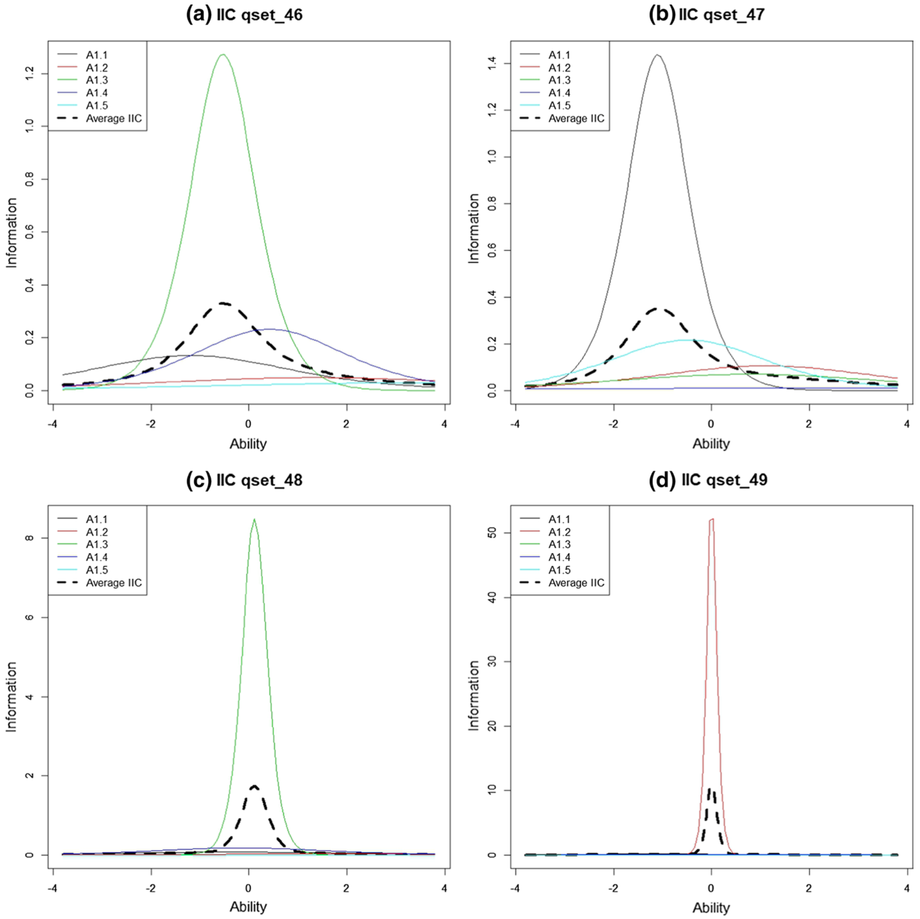

Fig. 3 *IIC*s and their average for the four QSETs of the A1 test

domain. Figure 6b illustrates the permutation functional F-test for the FANOVA model. The $F_{obs}(\theta)$ provides conflicting results because the curve intersect many times $F_{perm}(\theta)$. An overall judgement based on the whole domain would lead us to state that there is enough evidence against H_0. However, due to the specificity of the educational context, we focus on different intervals of abilities for assessing the equality of the four QSETs. Figure 6c displays the functional p value whereas Fig. 6d and Table 3 show the graph of the functional P-statistic and its values, respectively. We observe that, for low abilities, there is not enough evidence against the hypothesis of equality among group means, and we can establish that the use of different QSETs has not a significant effect on the *ICC*s. Indeed, in the interval $[-4, -2]$, the P-statistic is equal to 2.11. On the contrary, for middle–low, middle–high, and high abilities, there is enough evidence against H_0 because the P-statistic is lower than 1.

This result provides an additional information with respect to the classical FANOVA model. In effect, in this context, experts should revise items according to these evidences (i.e. according to the probability of correctly answer items, the four QSETs are similar

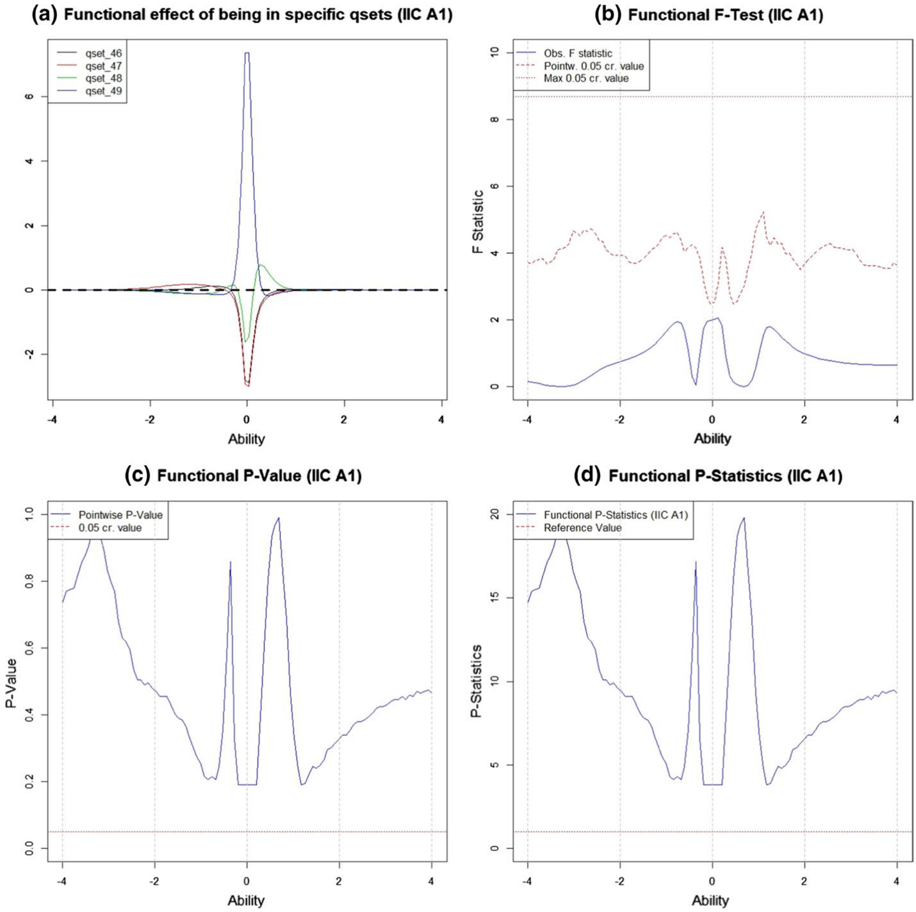

Fig. 4 F-test and functional statistics for *IIC*s of the A1 test

Table 2 Functional *P*-statistic on FANOVA model of *IIC*s of the A1 test

Abilities	*P*-Statistic	Evidence
Low	$14.12 > 1$	Not enough evidence against H_0
Middle–low	$7.04 > 1$	Not enough evidence against H_0
Middle–high	$9.17 > 1$	Not enough evidence against H_0
High	$8.52 > 1$	Not enough evidence against H_0

only for low-ability levels). Because the functional *P*-statistic is a measure of the degree of consistency (or "inconsistency") with H_0, we have to consider that the lower the *P*-statistic, the more different the QSETs. This circumstance highlights that the test should be revised with particular attention to those levels of ability for which the incongruence is more evident. In this application, Table 3 underlines that the difference among groups is particularly strong for high abilities (*P*-statistic = 0.30). We point out that the *P*-statistic, being a ratio,

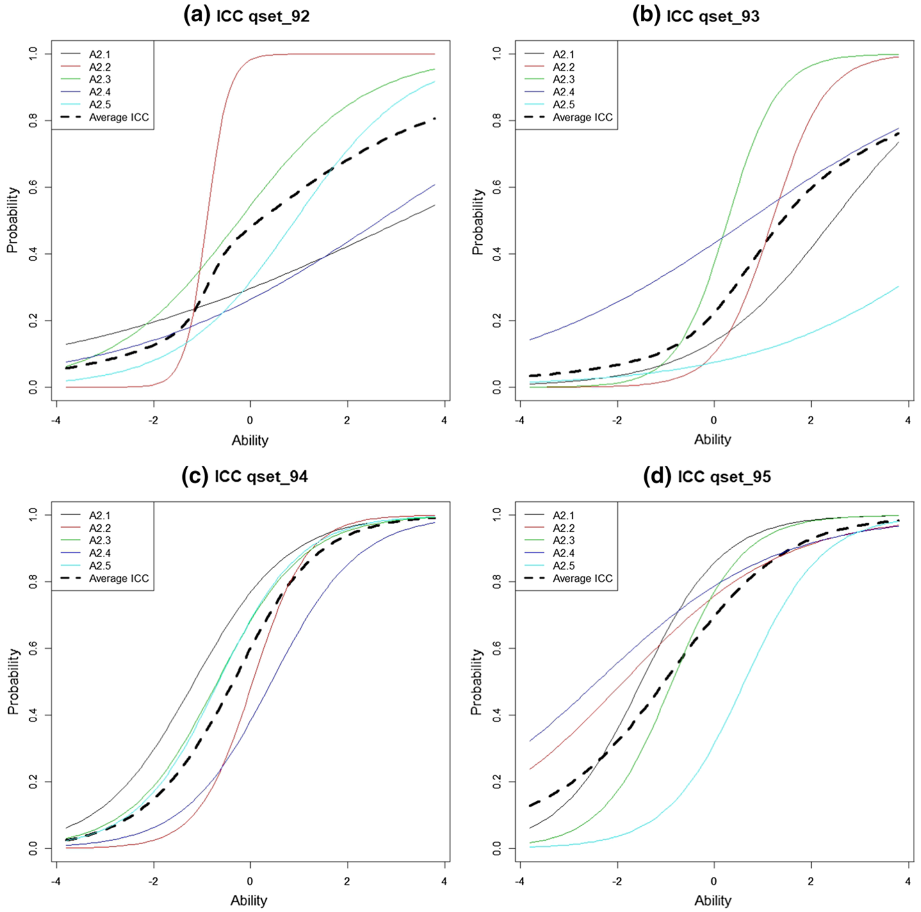

Fig. 5 *ICC*s and their average for the four QSETs of the A2 test

emphasizes the gap among group means because it is characterized by a multiplicative effect.

Figure 7 shows the *IIC*s of the A2 test for different QSETs, and suggests that the four QSETs are informative for different ability levels (the average peaks are in different points of the θ domain). Indeed, the QSETs 92 and 95 are useful for testing middle–low abilities whereas the QSET 93 plays a key role for assessing middle–high skills. Also Fig. 8a highlights this circumstance because the QSET 92 has a positive functional effect for middle–low abilities.

To statistically test this first evidence, we refer to the FANOVA model by plotting the functional F-statistic over the whole domain (Fig. 8b). Also in this case, an overall test may be inconsistent for our purposes because the curves intersect. Thus, we consider the *P*-statistics for different ability intervals to understand if there is enough evidence against H_0 (equal average information power among QSETs). Contrary to the FANOVA model on *ICC*s, Table 4 shows that group means are not equal for low ability levels whereas there is not enough evidence against H_0 for middle–low, middle–high, and high abilities. This

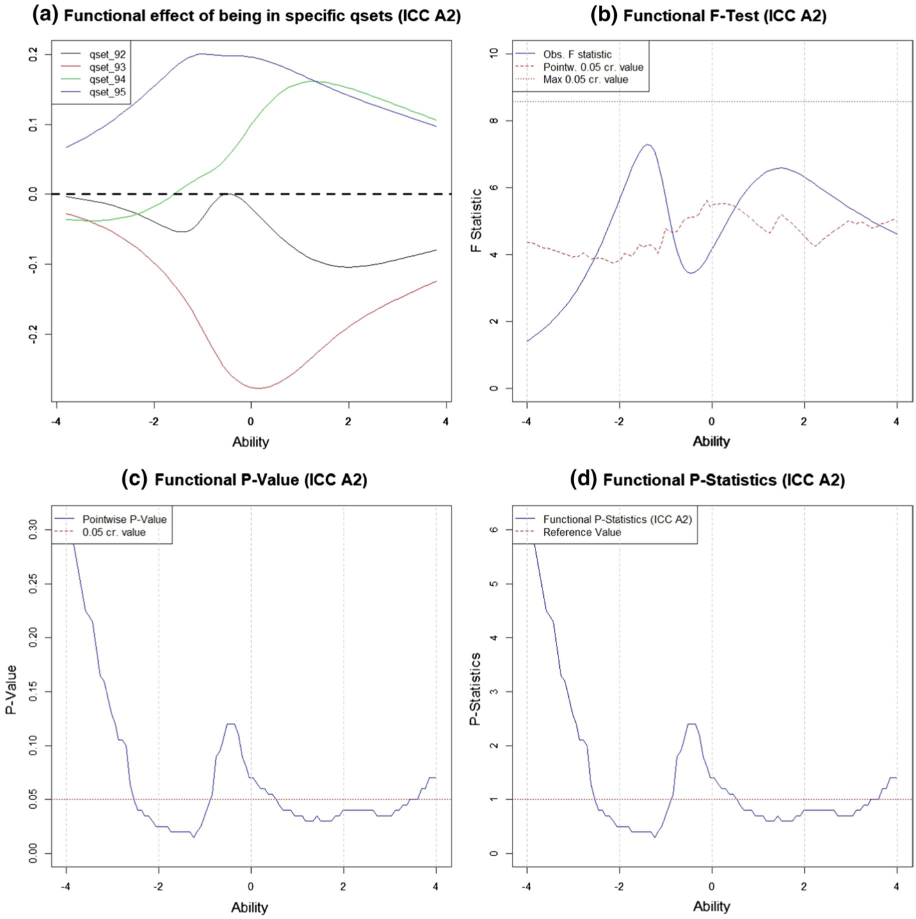

Fig. 6 F-test and functional statistics for *ICC*s of the A2 test

Table 3 Functional *P*-statistic on FANOVA model of *ICC*s of the A2 test

Abilities	*P*-Statistic	Evidence
Low	$2.11 > 1$	Not enough evidence against H_0
Middle–low	$0.73 < 1$	Enough evidence against H_0
Middle–high	$0.49 < 1$	Enough evidence against H_0
High	$0.30 < 1$	Enough evidence against H_0

result is quite interesting in the context of IRT models. Indeed, the FANOVA model on *ICC*s suggests that the group means are equal only for low ability levels whereas the FANOVA model on *IIC*s asserts the contrary (i.e. the group means are equal for ability levels greater than $\theta = -2$). This is a useful additional information provided by the functional approach. In effect, if we consider the probability of correctly answering the items (*ICC*s), the QSETs result to be well formulated for $\theta < -2$. Nevertheless, regarding the

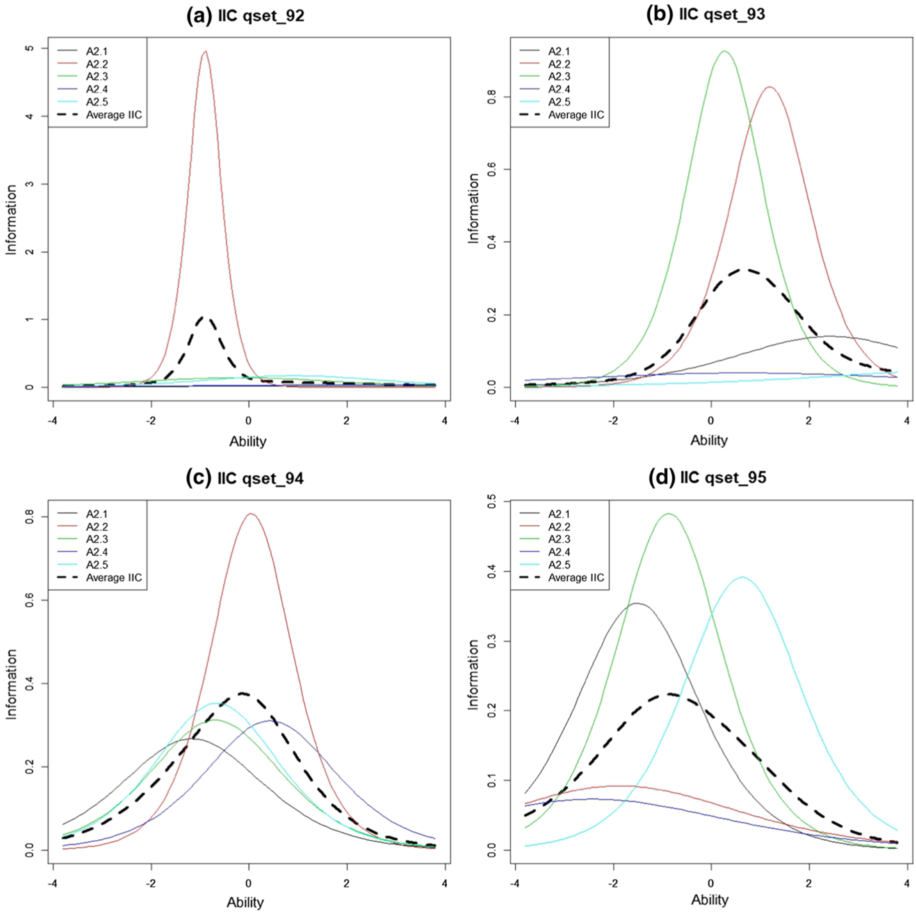

Fig. 7 *IIC*s and their average for the four QSETs of the A2 test

information power (*IIC*s), we observe that the four QSETs differently represent various ability levels for $\theta < -2$. This information naturally has an interesting insight into the educational field and should be used by practitioners to better calibrate the questionnaire and the different QSETs.

4 Discussion and Conclusions

IRT models refer to a family of latent trait 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 item. In the IRT context, mainly two functions are considered of great importance for the evaluation of items and tests: the item characteristic curve and item information curve. The item characteristic curve provides the probability of correctly response to items whereas the item information function shows the amount of information that each item provides. The

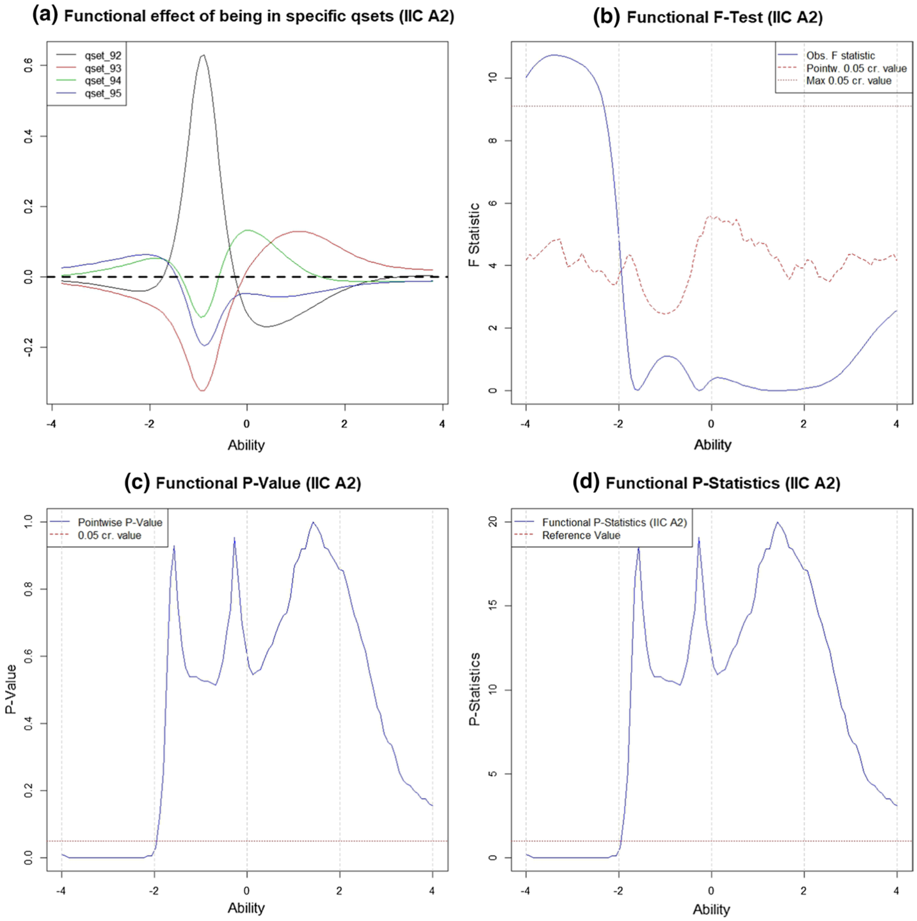

Fig. 8 F-test and functional statistics for *IIC*s of the A2 test

Table 4 Functional *P*-statistic on FANOVA model of *IIC*s of the A2 test

Abilities	*P*-Statistic	Evidence
Low	$0.10 < 1$	Enough evidence against H_0
Middle–low	$10.63 > 1$	Not enough evidence against H_0
Middle–high	$16.26 > 1$	Not enough evidence against H_0
High	$7.57 > 1$	Not enough evidence against H_0

latter is calculated by multiplying the probability of endorsing a correct response by the probability of answering incorrectly (in the case of the 2PLM model, this product is further multiplied by the square of the discrimination parameter). Thus, the larger the discrimination parameter, the greater the information provided by the item. Hence, the information contained in these two functions is linked but different. In this work, using the functional analysis of variance applied to IRT, we focused on these two functions by considering them as dependent variables of our model. Our purpose was to check if different versions of the questionnaire (QSETs), which were judged as equivalent from experts, have had a significant effect on *ICC*s and *IIC*s. At the same time, we have proposed a functional tools to assess the degree of consistency with the null hypothesis in a functional framework: the *P*-statistic. Due to the specificity of the IRT context, the latter has been separately considered for different skill ranges. This choice is due to the need to find a tool for evaluating the degree of consistency with the null hypothesis in various parts of the domain, and thus provides useful indications for modifying the questionnaire (in case of need).

In recent decades, many scholars have focused on the problem of differential item functioning (DIF) analysis. DIF is present when the relationship between the item score and the latent trait is not identical across subpopulations (Drasgow 1984). Despite the purposes of DIF methods have a common point with our proposal, that is to understand if there are significant differences between groups, our approach checks differences between two or more tests whereas DIF techniques focus on discovering differences between pairs of items. Moreover, the DIF analysis primarily considers differences among item parameters or total scores whereas our method allows us to treat data directly as curves. In addition, IRT DIF analysis is typically conducted between two groups where the reference group is often the majority socio-demographic group that has a larger sample and the focal one is the minority socio-demographic group with a smaller sample.

Recently, some studies have dealt with the issue of differential test functioning (DTF). Assessing DTF is considered an important extension of DIF for understanding tests and the effect of scoring bias, but unfortunately it has received relatively little attention in the applied measurement literature. DTF indicates that the test as a whole is not equivalent across groups, and naturally it can depend on cumulative DIF across all items. The seminal work on this topic is by Roju et al. (1995). Similarly to our method, Roju et al. aim also to discover differences considering the questionnaire as a whole and adopt the area between curves for testing statistical differences among items (or tests). Despite there is this common idea of considering areas to find the differences between tests, our method is quite different from these based on DTF. First, the latter adopts the expected score, i.e. they consider the expected test score for a trait level given all the estimated item parameters; in this context, the expected scores is also used as a measure of effect size because it is a direct measure of the difference in expected scores across groups. Contrary, we do not take into account the total score but focus on the behavior of *ICC*s and *IIC*s over the whole domain. This because our purpose is to use a functional approach for not losing important information about different domain ranges in considering both the probability of correctly answer to items and their informative power. Second, we do not consider the difference between a reference group and a focal group; indeed, our approach is generalizable to multiple groups without selecting anyone as a focal one. Finally, the DFT hypothesis test applies to the whole domain and is based on the assumption that the differences between the focal and control groups are normally distributed; on the contrary, our approach is based on decomposition of the domain in different skill ranges and the normality assumption is not required, i.e. our method is distribution-free.

The FDA approach can be viewed as a powerful tool to improve the existing techniques in the educational studies. Indeed, it allows us to preserve the usual interpretation of item parameters but takes advantage of functional data analysis tools. Specifically, the FDA approach grants a deeper analysis of those phenomena varying in a fixed domain, and, in an educational framework, it enables the evaluation of items behavior for all levels of the latent trait. In addition, the graphical inspection of the *P*-statistic may improve information for practitioners showing possible limitations of tests in specific ability intervals. Hence, this method may provide useful directions for defining specific educational plans.

Following this perspective, future developments of this study could focus on defining procedures to compare treatment groups when there is evidence against H_0 by introducing functional contrasts in the context of FANOVA.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723. https://doi.org/10.1109/TAC.1974.1100705.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. Lord & M. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–479). Boston: Addison-Wesle.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*(4), 443–459. https://doi.org/10.1007/bf02293801.

Carpita, M. (2017). L'analisi psicometrica dei test. In L. Covi & M. Dutto (Eds.), *Rapporto TLT 2016 Trentino Language Testing Esito delle rilevazioni delle competenze linguistiche degli studenti trentini* (pp. 71–86). Provincia Autonoma di Trento: IPRASE. (**ISBN 978-88-7702-426-8**).

Ceccatelli, C., Di Battista, T., Fortuna, F., & Maturo, F. (2013). Best practice to improve the learning of statistics: The case of the national olympics of statistics in italy. *Procedia: Social & Behavioral Sciences*, *XCIII*, 2194–2199. https://doi.org/10.1016/j.sbspro.2013.10.186.

Chen, S., Hwang, F., & Lin, S. (2013). Satisfaction rating of QOLPAV: Psychometric properties based on the graded response model. *Social Indicator Research*, *110*, 367–383.

Council of Europe. (2011). *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge: Cambridge University Press.

Covi, L., & Dutto, M. (2017). *Rapporto TLT 2016 Trentino Language Testing. Esito delle rilevazioni delle competenze linguistiche degli studenti trentini*. Provincia Autonoma di Trento: IPRASE. (**ISBN 978-88-7702-426-8**).

de Ayala, R. (2009). *The theory and practice of item response theory*. New York: The Guilford Press.

Di Battista, T., & Fortuna, F. (2016). Clustering dichotomously scored items through functional data analysis. *Electronic Journal of Applied Statistical Analysis*, *9*, 433–450.

Di Battista, T., & Fortuna, F. (2017). Functional confidence bands for lichen biodiversity profiles: A case study in Tuscany region (central Italy). *Statistical Analysis and Data Mining: The ASA Data Science Journal*, *10*, 21–28.

Di Battista, T., Fortuna, F., & Maturo, F. (2014). Parametric functional analysis of variance for fish biodiversity. In *International conference on marine and freshwater environments, iMFE 2014*. www.scopus.com.

Di Battista, T., Fortuna, F., & Maturo, F. (2016). Parametric functional analysis of variance for fish biodiversity assessment. *Journal of Environmental Informatics*, *28*(2), 101–109. https://doi.org/10.3808/jei.201600348.

Di Battista, T., Fortuna, F., & Maturo, F. (2017). BioFTF: An R package for biodiversity assessment with the functional data analysis approach. *Ecological Indicators*, *73*, 726–732. https://doi.org/10.1016/j.ecolind.2016.10.032.

Drasgow, F. (1984). Scrutinizing psychological tests: Measurement equivalence and equivalent relations with external variables are the central issues. *Psychological Bulletin*, *95*(1), 134–135. https://doi.org/10.1037/0033-2909.95.1.134.

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

Fortuna, F., & Maturo, F. (2018). K-means clustering of item characteristic curves and item information curves via functional principal component analysis. *Quality & Quantity*. https://doi.org/10.1007/s11135-018-0724-7.

Hambleton, R., & van der Linden, W. (1997). *Handbook of modern item response theory*. New York: Springer.

Liu, Y. (2016). Modelling and testing differential item functioning in unidimensional binary item response models with a single continuous covariate: A functional data analysis approach. *Psychometrika*, *81*, 371–398.

Lord, F. (1980). *Applications of item response theory to practical testing problems*. Hillsdale: Lawrence Erlbaum.

Lord, F., & Novick, M. (1968a). *Statistical theories of mental test scores (with contributions by A. Birnbaum)*. Reading, MA: Addison-Wesley.

Lord, F., & Novick, M. (1968b). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Manly, B. F. J. (1997). *Randomization, Bootstrap and Monte Carlo Methods in Biology*. Chapman and Hall, London. (**ISBN 0412721309**).

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

Maturo, F. (2018). Unsupervised classification of ecological communities ranked according to their biodiversity patterns via a functional principal component decomposition of hill's numbers integral functions. *Ecological Indicators*, *90*, 305–315. https://doi.org/10.1016/j.ecolind.2018.03.013.

Maturo, F., & Di Battista, T. (2018). A functional approach to Hill's numbers for assessing changes in species variety of ecological communities over time. *Ecological Indicators*, *84*(C), 70–81. https://doi.org/10.1016/j.ecolind.2017.08.016..

Maturo, F., Di Battista, T., & Fortuna, F. (2016). BioFTF: Biodiversity assessment using functional tools. https://cran.r-project.org/web/packages/BioFTF/index.html.

Maturo, F., Migliori, S., & Paolone, F. (2017). *Do institutional or foreign shareholders influence national board diversity? Assessing Board diversity through functional data analysis* (pp. 199–217). Cham: Springer. https://doi.org/10.1007/978-3-319-54819-7_14.

Maturo, F., Migliori, S., & Paolone, F. (2018). Measuring and monitoring diversity in organizations through functional instruments with an application to ethnic workforce diversity of the U.S. federal agencies. *Computational and Mathematical Organization*. https://doi.org/10.1007/s10588-018-9267-7.

O'Connor, B., Crawford, M., & Holder, M. (2015). An item response theory analysis of the subjective happiness scale. *Social Indicator Research*, *124*, 249–258.

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 W. van der Linden & R. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 381–394). New York: Springer.

Ramsay, J. O., & Silverman, B. W. (2005). *Functional data analysis* (2nd ed.). New York: Springer.

Rasch, G. (1960). *Probabilistic models for some intelligence and achievement tests*. Copenhagen: Danish Institute for Educational Research.

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

Roju, N. S., van der Linden, W. J., & Fleer, P. F. (1995). IRT-Based Internal Measures of Differential Functioning of Items and Tests. *Applied Psychological Measurement*, *19*(4), 353–368. https://doi.org/10.1177/014662169501900405.

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

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464. https://doi.org/10.1214/aos/1176344136.