Identifying and Classifying Aberrant Response Patterns through Functional Data Analysis

Abstract

We propose new methods for identifying and classifying Aberrant Response Patterns (ARP) by means of Functional Data Analysis (FDA). These methods take the Person Response Function (PRF) of an individual and compare it with the pattern that would correspond to a generic individual of the same ability according to the Item-Person Response Surface (IPRS). ARP correspond to atypical difference functions. The ARP classification is done with functional data clustering applied to the PRFs identified as ARP. We apply these methods to two sets of simulated data (the first is used to illustrate the ARP identification methods, and the second demonstrates classification of the response patterns flagged as ARP) and a real data set (a Grade 12 science assessment test, SAT, with 32 items answered by 600 examinees). For comparative purposes, ARP are also identified with three non-parametric person-fit indices (Ht, Modified Caution Index, and ZU3). Our results indicate that the ARP detection ability of one of our proposed methods is comparable to that of person-fit indices. Moreover, the proposed classification methods enable ARP associated with either spuriously low or spuriously high scores to be distinguished.

Keywords: Person-fit, Person Response Function, Item-Person Response Surface, Functional Data Analysis, Outlier detection, Functional Clustering.

1 Introduction

Knowledge and skills reflect individual characteristics that are evaluated indirectly through a respondent's performance on certain tasks that are grouped into a test. The respondent's answers to the test items are summarized into an individual score that is interpreted as an indicator of ability. Inferences from scores will be valid only if the individual level of achievement can be correctly inferred from them (AERA/APA/NCME, 2014). Various reasons may be adduced for inferring that an individual score is invalid. For example, respondents may copy the answers they do not know for the most difficult items from a more competent examinee. In this case, their test score will overestimate their ability. In other cases a test score may underestimate the abilities of examinies, e.g. when capable examinees fail to pay sufficient attention to the easiest items and thus provide incorrect answers. Such situations give rise to Aberrant Response Patterns (ARP), as a result of which inferences made about examinees' abilities based on their test score will not be valid.

Numerous person-fit indices have been proposed for identifying ARP (e.g., Meijer & Sijtsma, 2001), which to a greater or lesser extent efficiently identify response patterns that underestimate or overestimate the latent ability of the respondent being evaluated. However, none of these indices have been designed to identify directly the type of bias involved in estimating this latent ability. Several authors (Emons, Sijtsma, & Meijer, 2004, 2005; Nering & Meijer, 1998; Sijtsma & Meijer, 2001; Walker, Engelhard, Hedgpeth, & Royal, 2016) argued to identify the type of ARP from the Person Response Function (PRF), a synonym for *Person Response Curve*. A PRF provides the probability of a certain respondent giving the correct answer to each test item (Trabin & Weiss, 1983).

The objectives of this study are the identification and classification of ARP based on Functional Data Analysis (FDA). Analyzing PRF as functional data was first mentioned by Emons et al. (2004) as a possibility for further research in person-fit analysis. Our proposal essentially consists in comparing the PRF of an individual with the PRF that would correspond to a generic individual of the same ability according to the Item-Person Response Surface (IPRS), which also takes into account the estimated item parameters and which is estimated on the basis of the responses of all examinees. We compute the difference between both functions for each individual. The further the difference function is from zero, the more aberrant is the response pattern of the individual it represents. Functional outlier detection techniques are used to identify ARP. Finally, functional cluster analysis is applied to the PRFs that are flagged as aberrant in order to classify them into different types of ARP.

The paper is structured as follows. Section 2 introduces the Functional Data Analysis concepts employed throughout this paper. Section 3 presents the IPRS, which are defined in this section from a functional perspective. Section 4 presents our proposed methods for estimating the global Item-Person Response Surface and the PRF, as well as for detecting and classifying ARP. In order to illustrate the performance of these methods, two simulation studies are discussed in Section 5. Our proposed methods for identifying and classifying ARP are applied to a real data example in Section 6. Final conclusions are discussed in Section 7.

2 Functional Data Analysis

2.1 Basic concepts

FDA deals with the statistical description and modeling of samples in which a whole function is observed for each individual. For instance, in the Berkeley Growth Study, the heights of 39 boys and 54 girls were measured and registered at 31 specific time points from one to 18 years old. Each individual in the study contributed a complete function to the sample, namely his or her own growth curve. This is one of the examples included in Ramsay and Silverman (2005, first edition in 1997), which constitutes the first FDA monograph and includes functional versions for a wide range of statistical tools. A recent general introduction to FDA can be found in Kokoszka and Reimherr (2017). Two packages in R deserve special mention because of their broad coverage of functional tools: `fda` (Ramsay & Silverman, 2005) and `fda.usc` (Febrero-Bande & Oviedo de la Fuente, 2012). We use the latter one throughout this paper.

A *functional random variable* is a random variable $\boldsymbol{f}$ that takes values in an infinite functional space, usually the set of all the square-integrable functions defined in an interval $[a, b] \subseteq \mathbb{R}$: $L_2([a, b]) = \Big\{f : [a, b] \to \mathbf{R},\ \text{with}\ \int_a^b f(t)^2 dt < \infty\Big\}$. An observation f of $\boldsymbol{f}$ is called a *functional datum.* A *functional dataset* $f_1, \dots, f_n$ is the observation of n independent functional random variables $\boldsymbol{f}_1, \dots, \boldsymbol{f}_n$ that are identically distributed as $\boldsymbol{f}$. For convenience, we assume that the functional space is always $L_2([a, b])$ because it has a Hilbert space structure: it is a vector space with an inner product, $\langle f, g\rangle = \int_a^b f(t)g(t)dt$, inducing a norm, $\|f\| = \langle f, f\rangle^{1/2}$, known as the L_2-norm on $[a, b]$.

The parameters describing the probability distribution of a functional random variable $\boldsymbol{f}$ are defined in the usual way. The expected value function and the variance function $\boldsymbol{f}$ are $\mathbb{E}(\boldsymbol{f}(t))$ and $\mathrm{Var}(\boldsymbol{f}(t))$, respectively, for all $t \in [a, b]$. The covariance function of $\boldsymbol{f}$ is given by $c(s, t) = \mathrm{Cov}(\boldsymbol{f}(s), \boldsymbol{f}(t))$, for all s, and $t \in [a, b]$. These theoretical functions can be estimated by the corresponding descriptive statistics computed on a functional dataset coming from $\boldsymbol{f}$: $\bar{f}(t)$, $\widehat{\mathrm{Var}}(f(t))$ (respectively, the sample mean and variance of the values that functions $f_1, \ldots, f_n$ take at a particular value $t \in [a, b]$), and $\hat{c}(s, t)$ (the sample covariance between the values that functions $f_1, \ldots, f_n$ take at two values s and t in $[a, b]$). They can be computed with `func.mean` and `func.var` in library `fda.usc`.

2.2 Functional depth

The measure of depth is intended to quantify the centrality of an observation within a given sample; see Cuevas, Febrero-Bande, and Fraiman (2007) for a comprehensive treatment of the concept of statistical depth for functional data. Depth measures are useful for detecting functional outliers, identified as functional data with the least depth (Febrero-Bande, Galeano, & González-Manteiga, 2008).

In this paper, we use *modal depth* (through the function `depth.mode`, in the library `fda.usc`), which for each functional data f_i quantifies how densely it is surrounded by other data in the functional dataset. Given a metric or a semi-metric $d(\cdot, \cdot)$ between functions, for fixed $h > 0$ the h-modal-depth is

$$\mathrm{MD}_h(f_i) = \sum_{j \neq i} \frac{1}{h} K\left(\frac{d(f_i, f_j)}{h}\right)$$

where $K(z)$ is a kernel function (a unimodal symmetric univariate density function, which is typically the standard normal density). In the library `fda.usc`, the *tuning parameter* h is selected by default as the 15% quantile of $d(f_j, f_k)$, $j, k = 1, \ldots, n$, $j \neq k$, because empirical experience shows that the relative order of depths does not vary too much for values of h between 10% and 25% quantiles.

2.3 Clustering of functional data

The distance matrix between observations is the only information required for hierarchical clustering. Clustering functional data can therefore be performed as soon as a distance between functional data is defined. For instance, the L_2-norm can be used. Another classical clustering method, namely k-means, also requires computation of the averages of the observations allocated to the same cluster. We use the function `metric.lp` (in the package `fda.usc`) to compute distances and the function `hclust` for hierarchical clustering. We perform functional k-means using `kmeans.fd` (also from the package `fda.usc`).

3 Item-Person Response Surface

3.1 Definition and estimation

For the sake of simplicity we consider a unidimensional one-parameter IRT model. Let us assume that an exam has m items that differ only in their (latent) difficulties $b_1, \ldots, b_m$ ($b_j \in \Omega \subseteq \mathbb{R}$) and that n examinees with (latent) abilities $\theta_1, \ldots, \theta_n$ ($\theta_i \in \Theta \subseteq \mathbb{R}$) take the exam. The examination produces a $n \times m$ matrix $\boldsymbol{X}$ with entries $x_{ij} \in \{0, 1\}$, such that $x_{ij} = 1$ if and only if examinee i has answered item j correctly.

The IPRS is a function p defined from $\Theta \times \Omega$ to $[0, 1]$, such that $p(\theta, b)$ is the probability that a generic individual with ability θ gives the right answer to a generic item of difficulty b. It is expected that $p(\theta, b)$ increases in θ and decreases in b. An identifiability problem exists, consisting in the fact that abilities, difficulties and the IPRS p are not unambiguously determined. More specifically, let $\tau : \Theta \longrightarrow \Theta^*$ and $\nu : \Omega \longrightarrow \Omega^*$ be two increasing monotonic functions. Then the abilities $\theta^* = \tau(\theta)$, the difficulties $b^* = \nu(b)$ and the IPRS p^* defined from $\Theta^* \times \Omega^*$ to $[0, 1]$ as $p^*(\theta^*, b^*) = p(\tau^{-1}(\theta^*), \nu^{-1}(b^*))$ are equivalent to θ, b and $p(\theta, b)$. It follows that only the order of abilities and difficulties are relevant.

A simple statistical model for the data coming from an IPRS is as follows. Let $b_1, \ldots, b_m$ be m independent identically distributed (i.i.d.) observations from a random variable with known distribution function F_b. Let $\theta_1, \ldots, \theta_n$ be n i.i.d. observations from a random variable with known distribution function F_θ, which are independent from difficulties b_j. Given difficulties b_j, abilities θ_i, and the IPRS $p(\theta, b)$, let $X_{ij} \sim \text{Bern}(p(\theta_i, b_j))$ be $n \times m$ independent binary random variables. Entries x_{ij} in the matrix $\boldsymbol{X}$ are realizations of random variables X_{ij}.

The inference goal is to estimate abilities θ_i, difficulties b_j and the IPRS $p(\theta, b)$. Sometimes a parametric expression is assumed for $p(\theta, b)$. This is the case of the one-parameter logistic model (1PLM): $p(\theta, b) = 1/(1 + e^{-D(\theta - b)})$. Observe that this is a particular case of logistic regression $p(\theta, b) = 1/(1+e^{-(\beta_0+\beta_1\theta+\beta_2 b)})$, when $\beta_0 = 0$, $\beta_1 = -\beta_2 = D$. The logistic regression model can also be stated as

$$\text{logit}(p(\theta, b)) = \log\left(\frac{p(\theta, b)}{1 - p(\theta, b)}\right) = \beta_0 + \beta_1\theta + \beta_2 b. \qquad (1)$$

We use the 1PLM with $D = 1.7$ to generate a data set consisting of $n = 100$ individuals responding to a multiple-choice test with $m = 50$ items, each of them having four response options with only one being correct. The abilities θ_i and the difficulties b_j are independent observations of a standard normal distribution. We refer to this data set as the *Illustrative Example*.

When working with a non-parametric model we assume that

$$\operatorname{logit}(p(\theta, b)) = s(\theta, b), \tag{2}$$

where $s(\theta, b)$ is a smooth function (for instance, it has continuous second partial derivatives) that can take any value in $\mathbb{R}$. Additionally, it can be assumed that $s(\theta, b)$ increases in θ and decreases in b.

We estimate the models in Equations 1 and 2 in the following way. First, the total score is computed for examinee i on the exam as $t_i = \sum_{j=1}^{m} x_{ij}$. Then the ranks of $t_1, \ldots, t_n$, say $r_1, \ldots, r_n$, are calculated and transformed in estimated abilities by $\hat{\theta}_i = F_\theta^{-1}((r_i - 1/2)/n)$, and the same is done for the items. The ranks $s_i, \ldots, s_m$ for the number of wrong responses to items $1, \ldots, m$ are computed and the estimated difficulties are $\hat{b}_j = F_b^{-1}((s_j - 1/2)/m)$. The monotonicity of $p(\theta, b)$ is a sufficient condition for the consistency of $\hat{\theta}_i$ and $\hat{b}_j$ (e.g., Ramsay, 1991). Finally, a binary response regression model (either parametric or non-parametric) is fitted to the data $(\hat{\theta}_i, \hat{b}_j; x_{ij})$ to estimate $p(\theta, b)$. The maximum likelihood estimation is used for parametric regression models, and the penalized maximum likelihood estimation for non-parametric models (see, for instance, Ramsay, 2000). Figure 1 shows the estimation results for the *Illustrative Example* data.

3.2 Aberrant Response Patterns

The PRF for individual i is a function PRF_i from Ω to $[0, 1]$, such that $\text{PRF}_i(b)$ computes its probability of giving a correct answer to an item with difficulty $b \in \Omega$. Under the previous IPRS model, if θ_i is the ability of individual i, then $\text{PRF}_i(b) = p(\theta_i, b)$ for all $b \in \Omega$. That is, $\text{PRF}_i(b)$ is the slice of the IPRS $p(\theta, b)$.

Nevertheless, it is possible that the PRF for certain individuals do not coincide

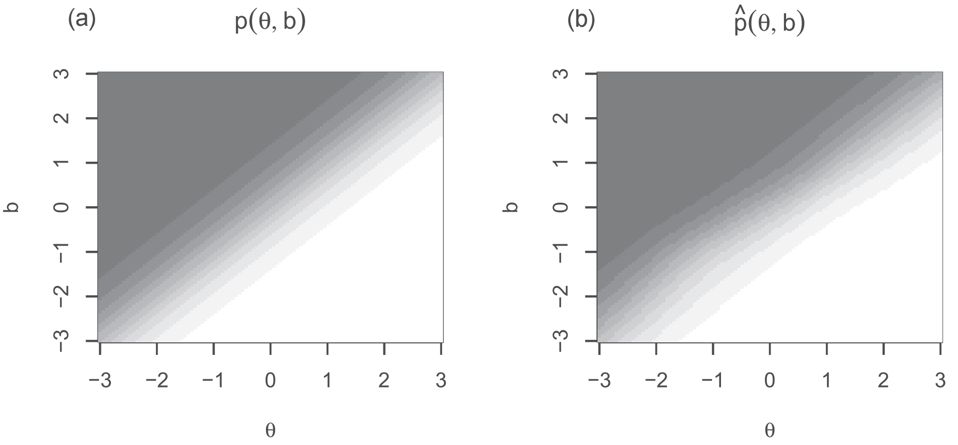

Figure 1: *Illustrative Example.* (a) Two-dimensional representation of the IPRS $p(\theta, b)$. (b) Estimation $\hat{p}(\theta, b)$. The lighter color in both panels (resp. darker) corresponds to the higher (resp. lower) probability of a correct answer.

with their corresponding IPRS profiles. In such cases we say that these individuals present an ARP. More formally, we say that individual i follows an ARP when the functions $\text{PRF}_i(\cdot)$ and $p(\theta_i, \cdot)$ are different. In accordance with Karabatsos (2003), the types of ARP used in this study are listed in Table 1. A comprehensive list of ARP types discussed in the literature and their relationships to real-life testing situations can be found in the review by Rupp (2013).

An IPRS model including the ARP is as follows. Consider the difficulties b_j and the abilities θ_i generated as explained before. Let $p(\theta, b)$ be the IPRS defined as before. Let $\pi_A \in [0, 1]$ be the probability that an examinee has an ARP. For individuals $i = 1, \ldots, n$, let their PRF be $\text{PRF}_i(b) = p(\theta_i, b)$ with probability $1 - \pi_A$, and $\text{PRF}_i(b) = g_i(b)$ with probability π_A where $g_i(b)$ is an ARP (corresponding, for instance, to cheaters, lucky guessers, random respondents, careless respondents or creative respondents). The random selection of individuals with ARP is mutually independent. Finally, X_{ij} follows a Bernoulli distribution with probability of success $\text{PRF}_i(b_j)$. Figure 2 shows the PRF corresponding to the *Illustrative Example* data and ordered by the ranking of abilities θ_i. For 10 ex-

Table 1: Types of ARP considered in this study, their characteristics and operationalization for simulating them.

Type of ARP	Definition	Operationalization
Cheaters	These are low ability examinees who unfairly obtain the correct answers on test items that they are unable to answer correctly. For instance, they copy from another examinee with a higher ability.	An individual with low ability ($\theta_i < z_{.375}$) is randomly selected. Then $p(\theta_i, b)$ is replaced by $g_i(b)$ defined as $$g_i(b) = 1 \text{ if } b > z_{.85},$$ otherwise $g_i(b) = p(\theta_i, b)$.
Lucky-guessers	These examinees guess the correct answers to some test items, for which they do not know the correct answer.	An individual with low ability ($\theta_i < z_{.375}$) is randomly selected. Then $p(\theta_i, b)$ is replaced by $g_i(b)$ defined as $$g_i(b) = .25 \text{ if } b > z_{.85},$$ otherwise $g_i(b) = p(\theta_i, b)$.
Random respondents	These examinees randomly choose the answer for each item on the multiple-choice test.	An individual is randomly selected. Then $p(\theta_i, b)$ is replaced by the constant $g_i(b) = .25$.
Careless respondents	These examinees answer certain items incorrectly, even though they are able to answer them correctly.	An individual with high ability ($\theta_i > z_{.625}$) is randomly selected. Then $p(\theta_i, b)$ is replaced by $g_i(b)$ defined as $$g_i(b) = .5 \text{ if } b < z_{.15},$$ otherwise $g_i(b) = p(\theta_i, b)$.
Creative respondents	These are examinees with a high ability who give incorrect responses to the easiest items, because they interpret them creatively.	An individual with high ability ($\theta_i > z_{.625}$) is randomly selected. Then $p(\theta_i, b)$ is replaced by $g_i(b)$ defined as $$g_i(b) = 0 \text{ if } b < z_{.15},$$ otherwise $g_i(b) = p(\theta_i, b)$.
Mixed	This category is a mixture of the previous types.	One fifth of the simulated ARP comes from each different type.

aminees (randomly chosen), their profiles $p(\theta_i, b)$ are replaced by $g_i(b)$.

4 Implementation

4.1 Outline of the procedure

We propose the following procedure for identifying examinees with ARP. First, the IPRS is estimated using data from all the individuals in such a way that the estimation is not significantly affected when ARP are present. Secondly, the PRF of individual i is estimated using data from this individual only; this estimation will be affected by the presence of ARP. Thirdly, for each individual the corresponding profile derived from the estimated IPRS is compared with its individually estimated PRF. Individuals for which both estimations are extremely different are identified as ARP. The estimation procedures are detailed in Section 4.2, and the ARP identification process is described in Section 4.3.

4.2 IPRS and PRF estimation

Following the estimation strategy for the IPRS model introduced in Section 3, we propose using a specific type of non-parametric regression model: the Generalized Additive Model (GAM; see, for instance, Wood, 2017), which assumes that

$$\operatorname{logit}(p(\theta, b)) = \log\left(\frac{p(\theta, b)}{1 - p(\theta, b)}\right) = \beta_0 + s_1(\theta) + s_2(b),$$

where s_1 and s_2 are smooth functions that can take any value in $\mathbb{R}$. Additionally it can be assumed that s_1 is an increasing function and that s_2 is decreasing. One may observe that the GAM model is more flexible than the logistic model in Equation 1, which is a particular case when $s_1(\theta) = \beta_1\theta$ and $s_2(b) = \beta_2 b$, but not

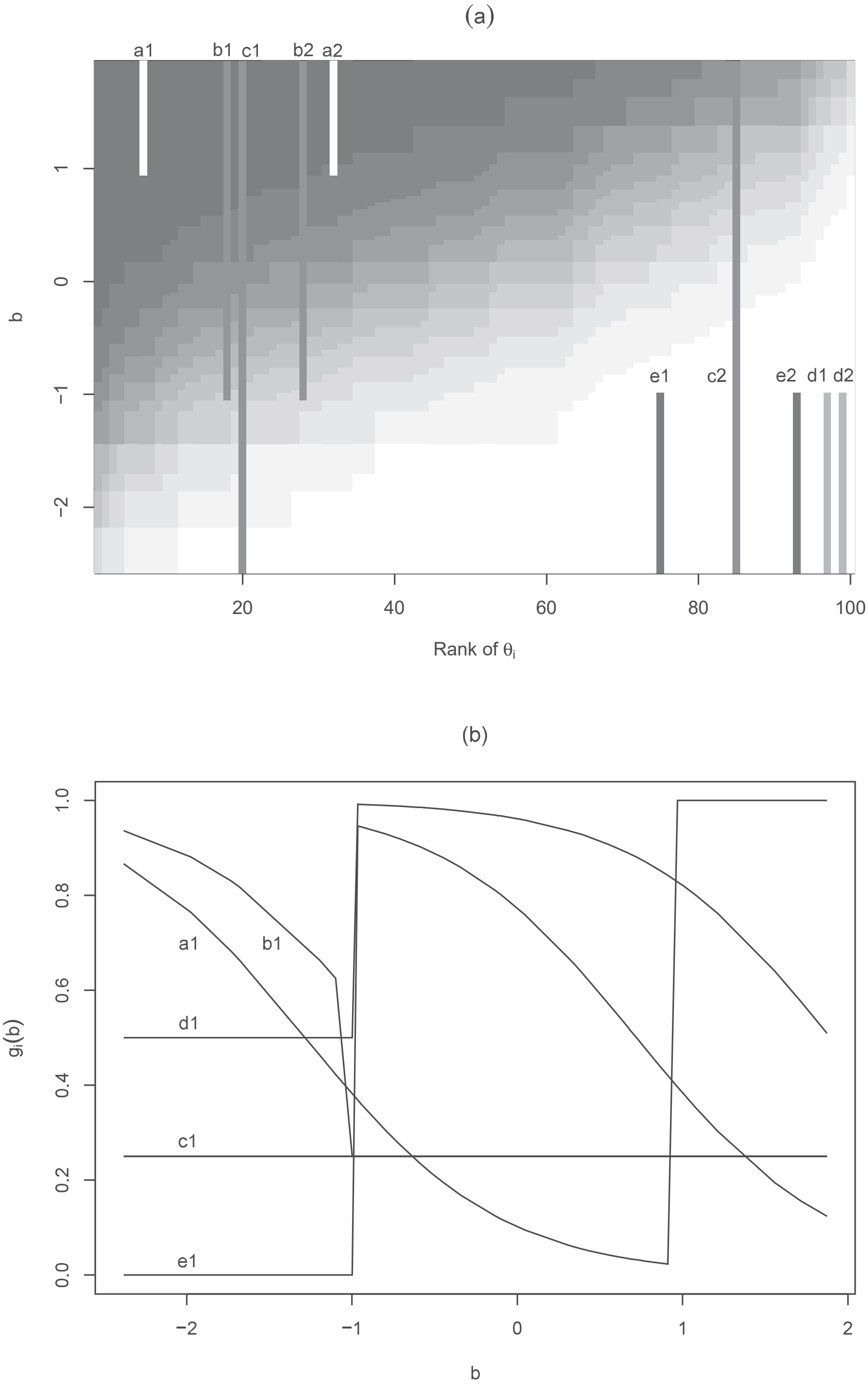

Figure 2: *Illustrative Example* including 10 ARP of the following types: cheaters (a1 and a2), lucky guessers (b1 and b2), random respondents (c1 and c2), careless respondents (d1 and d2), and creative respondents (e1 and e2). Panel (a) gives their PRF and Panel (b) gives the ARP profiles for one examinee of each type.

as flexible as the non-parametric model in Equation 2 (both coincide only when $s(\theta, b) = s_1(\theta) + s_2(b)$). Note that GAM achieves a balance between the flexibility of a full non-parametric regression model and the low sensitivity against ARP of parametric models (e.g., logistic regression). The estimations are carried out using the function `gam` in the library `mgcv`.

In order to estimate an individual's PRF_i, we fit a non-parametric estimator with a logistic link (e.g., by using penalized spline regression as in Wood, 2017) for every individual in the dataset to the pairs $(\hat{b}_j;\, x_{ij})$, $j = 1, \ldots, m$. Let $\widehat{\mathrm{PRF}}_i(b)$ be the estimated function. We fit penalized spline regression using the function `gam` in the library `mgcv`. In order to avoid degeneracies, it is necessary to remove the following response patterns from the data set before estimation: (a) individuals with constant answers (e.g., all answers are correct or all are incorrect); (b) those with a perfect Guttman pattern (e.g., correct answers to a certain number of the easiest items and incorrect answers to the remaining items that are more difficult); and (c) those with an anti-Guttman pattern (e.g., incorrect answers to a certain number of the easiest items and correct answers for the remaining items that are more difficult).

For the *Illustrative Example*, Figure 3a shows the estimated IPRS and Figure Figure 3b the PRF for the 100 individuals whose PRF was plotted in Figure 2. The ranking and relative order of the 10 individuals presenting ARP have changed with respect to Figure 2 because these individuals were chosen at random before their IPRS profiles were replaced by the ARP functions that correspond to abilities different from those they originally had. Figure 3b shows that there are more than 10 estimated PRF that could be declared ARP.

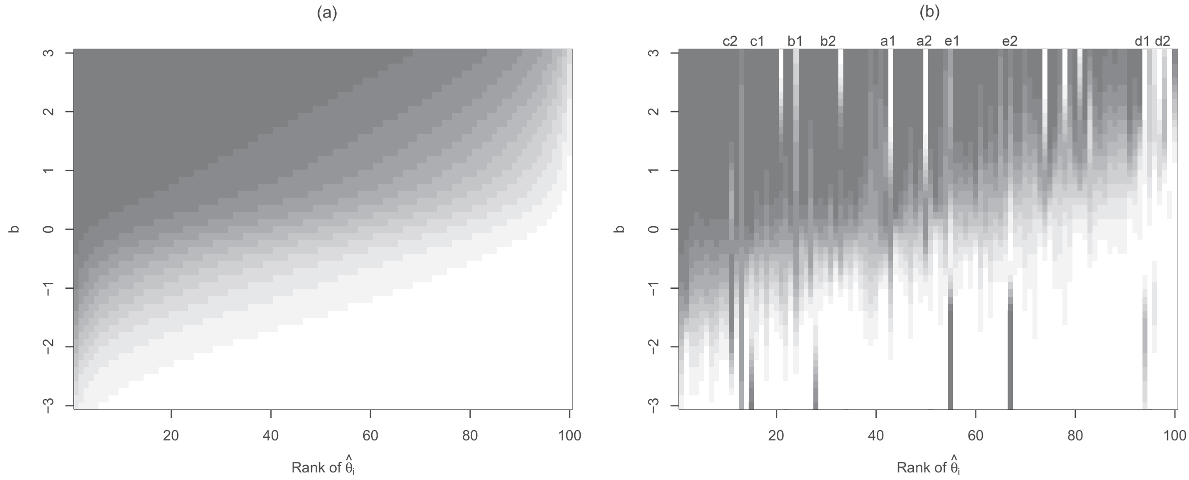

Figure 3: Continuation of the *Illustrative Example*. (a) IPRS and (b) PRF estimates of the 100 individuals whose PRF was plotted in Figure 2.

4.3 Detection of ARP

Rationale. We propose to compute the differences between $\widehat{\mathrm{PRF}}_i(b)$ and the profile function $\hat{p}(\hat{\theta}_i, b)$ in the logit scale, $D_i(b) = \mathrm{logit}\left(\widehat{\mathrm{PRF}}_i(b)\right) - \mathrm{logit}\left(\hat{p}(\hat{\theta}_i, b)\right)$, $b \in \Omega$, which leads to the functional data set $\{D_i(b) : i = 1, \ldots, n\}$. If individual i does not have an ARP, $D_i(b)$ should be close to 0 for all $b \in \Omega$ because $\widehat{\mathrm{PRF}}_i(b)$ and $\hat{p}(\hat{\theta}_i, b)$ both estimate the same function of b: $p(\hat{\theta}_i, b)$.

When individual i has an ARP, $D_i(b)$ is expected to be far from 0 because $\hat{p}(\hat{\theta}_i, b)$ estimates $p(\hat{\theta}_i, b)$ (especially when the non-parametric bivariate estimator is not so flexible, as is the case with GAMs) but $\widehat{\mathrm{PRF}}_i(b)$ estimates $g_i(p)$. So $D_i(b)$ will appear as functional outliers for individuals with an ARP.

ARP detection. We propose three methods for ARP detection, the first two of which are based on FDA. In particular, the first method consists in computing the modal depth of the functional data $\{D_i(b) : i = 1, \ldots, n\}$. The least deep functions are candidates for being functional outliers (a bootstrap test exists for finding atypicality; see Febrero-Bande et al., 2008) and consequently correspond to ARP. We label this method `Foutl`. The second proposal consists in applying the first ARP detection method to the first derivatives $D_i'(b)$, because the functional outliers are sometimes revealed much more clearly when looking at derivatives than when looking at the original functions (see, for instance, Febrero-Bande et al., 2008). We label this method `Foutl.d1`.

Finally, our third proposal is based on a log-likelihood ratio test approach. Consider individual i. There are two ways to compute the log-likelihood of the observed data $x_{ij}, j = 1, \ldots, m$, the first of which uses the estimated PRF_i:

$$\mathrm{loglik.PRF}_i = \sum_{j=1}^{m} \left\{ x_{ij} \log \left(\widehat{\mathrm{PRF}}_i(\hat{b}_j) \right) + (1 - x_{ij}) \log \left(1 - \widehat{\mathrm{PRF}}_i(\hat{b}_j) \right) \right\}.$$

The second one uses the estimated IPRS:

$$\mathrm{loglik.IPRS}_i = \sum_{j=1}^{m} \left\{ x_{ij} \log \left(\hat{p}(\hat{\theta}_i, \hat{b}_j) \right) + (1 - x_{ij}) \log \left(1 - \hat{p}(\hat{\theta}_i, \hat{b}_j) \right) \right\}.$$

$\mathrm{Loglik.PRF}_i - \mathrm{loglik.IPRS}_i$ is expected to take abnormally high values for individuals with ARP. We then identify ARP by looking for outliers in the one-dimensional dataset $\{\mathrm{loglik.PRF}_i - \mathrm{loglik.IPRS}_i : i = 1, \ldots, n\}$. We label this method `LikRat`. Observe that $\mathrm{loglik.PRF}_i - \mathrm{loglik.IPRS}_i$ is the log-likelihood ratio test statistic for testing the null hypothesis $H_0 : \mathrm{PRF}_i(b) = p(\hat{\theta}_i, b)$ for all b, against the alternative that these two functions are different.

Illustration. Figure 4a shows for the *Illustrative Example* the differences $\{D_i(b), i = 1, \ldots, n\}$, between individual PRF and the IPRS profile estimates. Their first derivatives, $D_i'(b)$, are shown in Figure 4b. Black curves correspond to the 10 ARP introduced in Figure 2. The ARP of the cheater and creative respondent types were the most clearly identifiable, followed by careless respondents, lucky guessers and random respondents (more clearly distinguished by their derivatives). This result was expected given the way in which the ARP were simulated: cheaters and creative respondents are intensified versions of careless respondents and lucky guessers, respectively. Moreover, there were other curves, $D_i(b)$ and/or $D'_i(b)$, far from zero even if they did not correspond to ARP, indicating that ARP identification may not be an easy task.

Given that the method based on the log-likelihood ratio test statistic is a non-standard way of detecting outliers, its performance is shown for the illustrative data. Figure 5 displays the scatter plot of $(\text{loglik.IPRS}_i, \text{loglik.PRF}_i)$, $i = 1, \ldots, n$. The dashed line corresponds to differences between loglik.PRF_i and loglik.IPRS_i equal to the upper whisker value of the box-plot for the one-dimensional data $\{\text{loglik.PRF}_i - \text{loglik.IPRS}_i,\ i = 1, \ldots, n\}$. So only the individuals with points in the upper left part of the plot would be regarded as ARP. In this case 9 out of the 10 ARP were detected as ARP (the unidentified tenth ARP is a lucky guesser) and no other points were wrongly declared as such.

4.4 Classification of identified ARP

Rationale. In order to characterize the PRF of individuals identified as ARP, we propose using clustering techniques for functional data (k-means and hierarchical clustering, as in Section 2.3). The working functional dataset is formed by

$$G_o(b) = \text{logit}(\text{PRF}_o),\ o = 1, \ldots, \mathcal{O},$$

where $\mathcal{O}$ is the number of functional outliers identified as ARP. The distances between functions are calculated as the weighted L_2-norm of the logit differences:

$$d_{h,k} = d(G_h, G_k) = \left(\int_{-2}^{2} \left(G_h(b) - G_k(b)\right)^2 \phi(b) db \right)^{1/2}, \qquad (3)$$

where ϕ is the density function of a standard normal (as difficulties b are $N(0, 1)$).

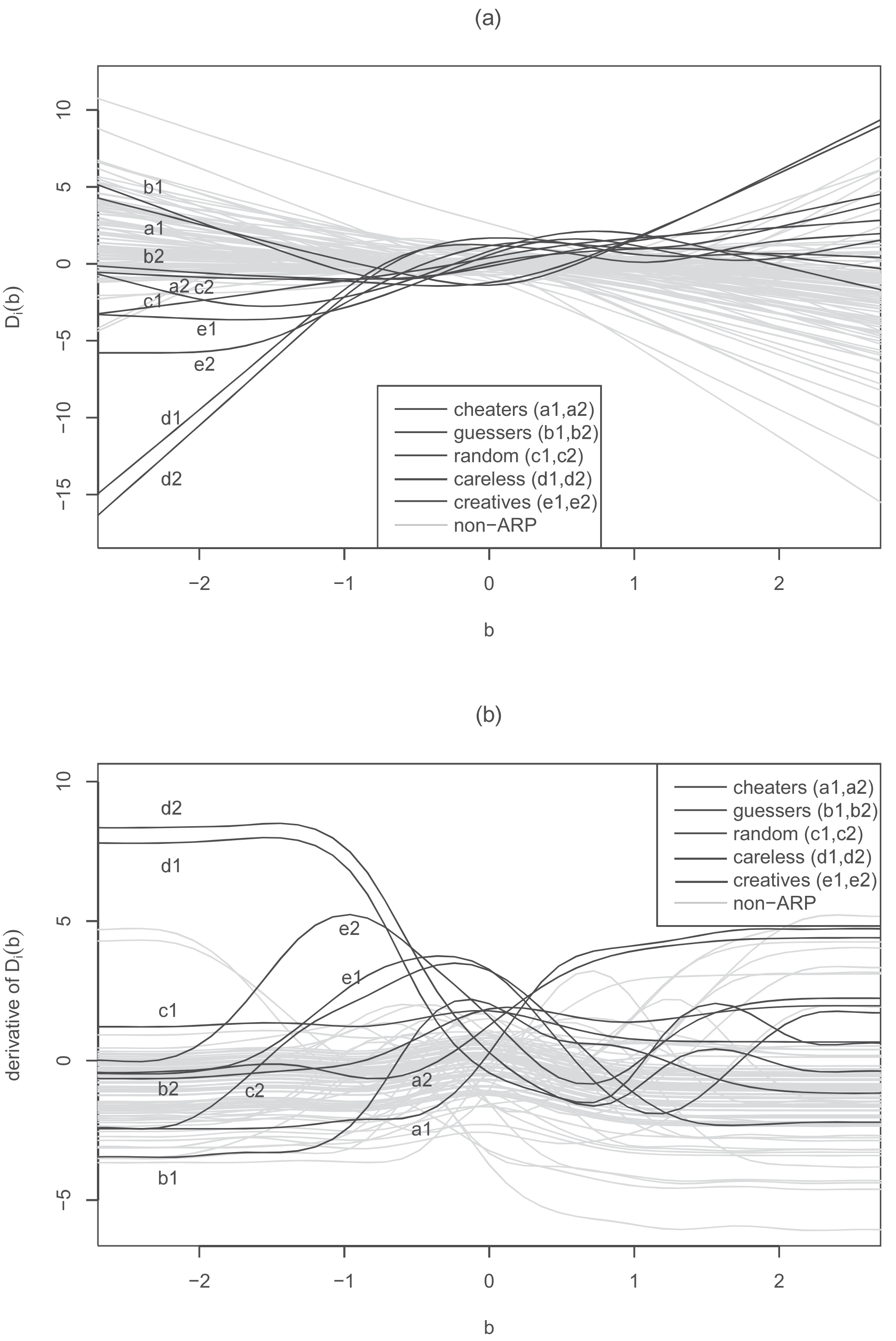

Figure 4: (a) $\{D_i(b) : i = 1, \ldots, n\}$, differences and (b) first derivatives $\{D'_i(b) : i = 1, \ldots, n\}$ in the logit scale of individual PRF and the IPRS profile estimates shown in Figure 3.

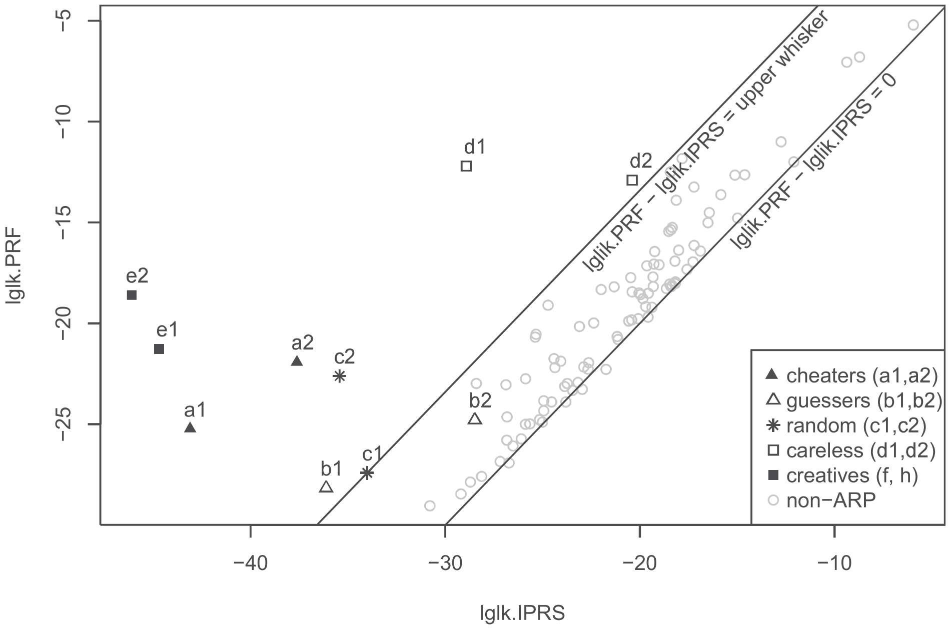

Figure 5: Scatter-plot of (loglik.IPRS$_i$, loglik.PRF$_i$), $i = 1, \ldots, n$, for the *Illustrative Example*. The dashed line corresponds to the upper whisker value of the box-plot for $\{\text{loglik.PRF}_i - \text{loglik.IPRS}_i,\ i = 1, \ldots, n\}$.

We choose $[-2, 2]$ as the integration interval, because on the one hand it has a high probability under the standard normal distribution, and on the other hand by excluding more extreme values of b we avoid numerical problems when taking the logit transformation of $\text{PRF}_h(b)$ values that are too close to zero or one.

Procedure. For functional k-means clustering we used the function `kmeans.fd` in the package `fda.usc`. The hierarchical clustering was performed by using the standard clustering R function `hclust` while working on the functional distances d_{hk} computed in Equation 3 and using Ward's linkage method. The number of clusters (k in the k-means method, or the number of groups after cutting the dendogram into hierarchical clustering) was automatically selected by minimizing the ratio of the average distance within clusters divided by the average distance between clusters (we have used the function `cluster.stats` of the R package `fpc`, Hennig, 2018).

Illustration. As an example, we worked with a simulated dataset of 500 individuals and 50 items with a proportion of 10% of ARP. Specifically, we had 10 of each type: cheaters, lucky guessers, random respondents, careless respondents, and creative respondents (details on the simulation are given below in Section 5).

When applying the ARP detection method based on log-likelihood differences (using an ARP threshold based on a resampling procedure that mimics that of the `cutoff` function in the R library `PerFit` from Tendeiro, Meijer, & Niessen, 2016), a total of $\mathcal{O} = 49$ cases were identified as such, which were distributed as follows: 10 cheaters, seven guessers, eight random, four careless respondents, 10 creative respondents and 10 non-ARP.

For the functional k-means clustering, the optimal number of clusters was $k = 6$ when using the ratio within/between distances as a criterion. Randomly selected initial centers were used. The results are summarized in Table 2 and in Figure 6a. Looking row-wise at Table 2, it follows that the different types of ARP could be retrieved except for the careless respondents, that were in different clusters. Column-wise this table shows that the clusters were heterogeneous with respect to the different ARP types, except for the smaller cluster with cheaters and for two non-ARP clusters.

Figure 6a gives a graphical representation of the discovered clusters. The thin gray dashed curves are the PRF of the 49 individuals identified as ARP. Each thick curve represents a summary of one cluster: they are the logit inverse transformation of the average of the logit transformation of the PRF for the individuals in each cluster. These summary curves had the expected shape: the two solid

Table 2: *Cross table of the true ARP type by the assigned cluster by k-means clustering for the 49 individuals identified as ARP. The cluster numbers correspond to those in Figure 6a.*

	Clusters by k-means					
Truth	1	2	3	4	5	6
non-ARP	2	3	0	0	1	4
cheaters	0	0	3	7	0	0
guessing	0	0	0	3	4	0
random	0	0	0	2	5	1
careless	0	0	0	0	0	4
creative	0	0	0	3	0	7

gray curves correspond to Clusters 1 and 2, including non-ARP respondents who differed in ability but shared the characteristic of moving too suddenly from easier items with correct responses to more difficult items with incorrect responses; the dot-dashed curve (corresponding to Cluster 3 comprising only three cheaters) increases abnormally for the difficult items ($b > 0$); the 2-dotted curve is similar but somewhat flatter (this corresponds to Cluster 4 containing cheaters, guessers and random respondents); the solid black curve (Cluster 5, mostly random and guesser respondents) is almost flat; the long-dashed curve (Cluster 6 with creative and careless respondents) is abnormally low for easy items ($b < 0$). The dashed gray curve corresponds to the average of all truly non-ARP cases.

The dendogram resulting from hierarchical clustering is shown in Figure 7. It suggests that a cut defining three clusters is appropriate, leading to results summarized in Table 3 and in Figure 6b. One may observe that one cluster with only non-ARP individuals was identified (Cluster 1, which average is the gray solid line in Figure 6b). Two more clusters also appeared: Cluster 2 (2-dotted curve in Figure 6b) included all cheaters, guessers and random respondents plus one non-ARP case, while Cluster 3 (long-dashed curve in Figure 6b) consisted of

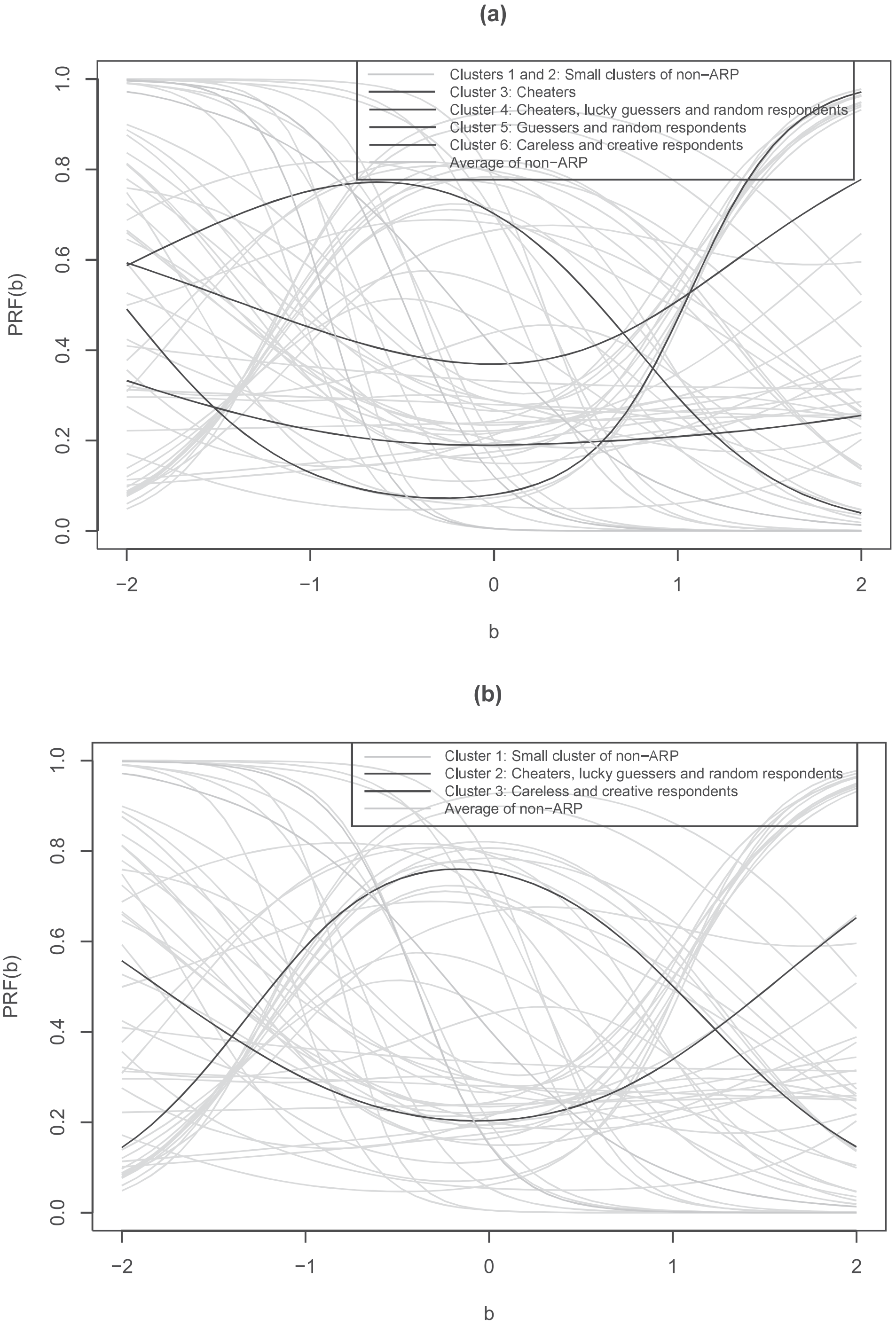

Figure 6: Clustering with (a) k-means and (b) hierarchical clustering of the 49 $\mathrm{PRF}_h(b)$ curves identified as ARP (gray lines).

creative and careless respondents plus one non-ARP.

Table 3: *Cross table of the true ARP type by the assigned cluster by hierarchical clustering for the 49 individuals identified as ARP. The cluster numbers correspond to those in Figure 6b.*

	Hierarchical clustering		
Truth	1	2	3
non-ARP	8	1	1
cheaters	0	10	0
guessing	0	7	0
random	0	8	0
careless	0	0	4
creative	0	0	10

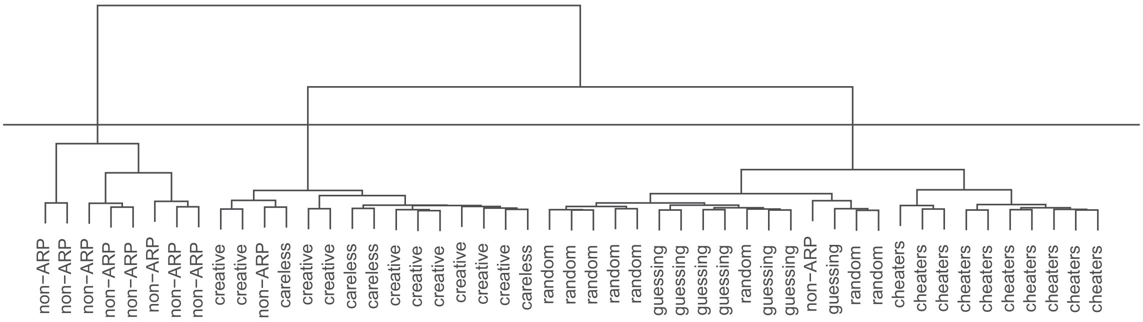

Figure 7: Dendogram for the hierarchical clustering applied to ARP from simulated data.

There was a high degree of concordance between the clusters obtained by k-means and those by hierarchical clustering, as shown in Table 4. The main difference was that Clusters 3, 4 and 5 obtained by k-means were grouped when using hierarchical clustering as Cluster 2.

Table 4: *Cross table of the cluster composition through k-means and hierarchical clustering for the 49 individuals identified as ARP. The cluster numbers correspond to those used in Figure 6.*

	Hierarchical clustering		
Clusters by k-means	1	2	3
1	2	0	0
2	3	0	0
3	0	3	0
4	0	12	3
5	0	10	0
6	3	1	12

5 Simulation studies

5.1 Simulation of the ARP detection

Methods. We conducted a simulation study to evaluate the ARP identification power of the three different methods proposed in Section 4.3. Furthermore, we compared them with three well-known non-parametric person-fit statistics, which have been reported to be among the best at identifying aberrant-responding examinees by Karabatsos (2003): Sijtsma's Ht, Harnisch & Linn's Modified Caution Index (labeled as `Cstar` in the figures), and van der Flier's ZU3 (the standardized form of van der Flier's U3 index). We used their implementation in the library `PerFit` (Tendeiro et al., 2016) by means of the functions `Ht`, `Cstar` and `ZU3`, respectively. In order to determine the reference values according to which these person-fit statistics indicate that a respondent can be regarded as an ARP, the `PerFit cutoff` function was used. This function employs a resampling procedure to determine the required reference values.

Instead of the default value of 1000, 100 resamples were used in our simulations in order to reduce computing time. The other function parameters in the library `PerFit` were left at their default values. In order to compare the six ARP detection methods (the three proposed methods plus the three non-parametric person-fit statistics) on an equal footing, for our proposal based on the log likelihood-ratio test statistic we mimicked the resample scheme included in the `cutoff` function. For the methods based on functional depth, we used the bootstrap procedure already implemented in the function `outliers.depth.trim` from the library `fda.usc`.

The design of the simulation study was in accordance with what is usual in this field (Rupp, 2013). A total of $S = 100$ exams were independently simulated, each with $m = 50$ items. The number of examinees was $n = 200$ for the first 50 exams and $n = 500$ for the other 50. The individual abilities θ_i, $i = 1, \ldots, n$ were independent values of a standard normal. The same applied for the item difficulties b_j, $j = 1, \ldots, m$. The true IPRS was the 1PLM with $D = 1.7$.

Three different proportions $(0.05, 0.1, 0.25)$ of ARP were used with each of the five different types of ARP (cheater, lucky guesser, random respondent, careless respondent, and creative respondent) separately, and a separate "mixed" group in which 20% of the ARP were each of the five types. This yielded a total of 18 combinations. The way these ARP were generated was based on the Karabatsos (2003) simulations. These and other ways to operationalize different types of ARP were reported by Rupp (2013).

Results. Figures 8 and 9 summarize the results of the simulation study for $n = 500$. While not included here, results for $n = 200$ leaded to similar conclusions. These figures show the sensitivity (Figure 8) and the specificity (Figure 9) of each ARP detection method for all of the 18 simulated scenarios.

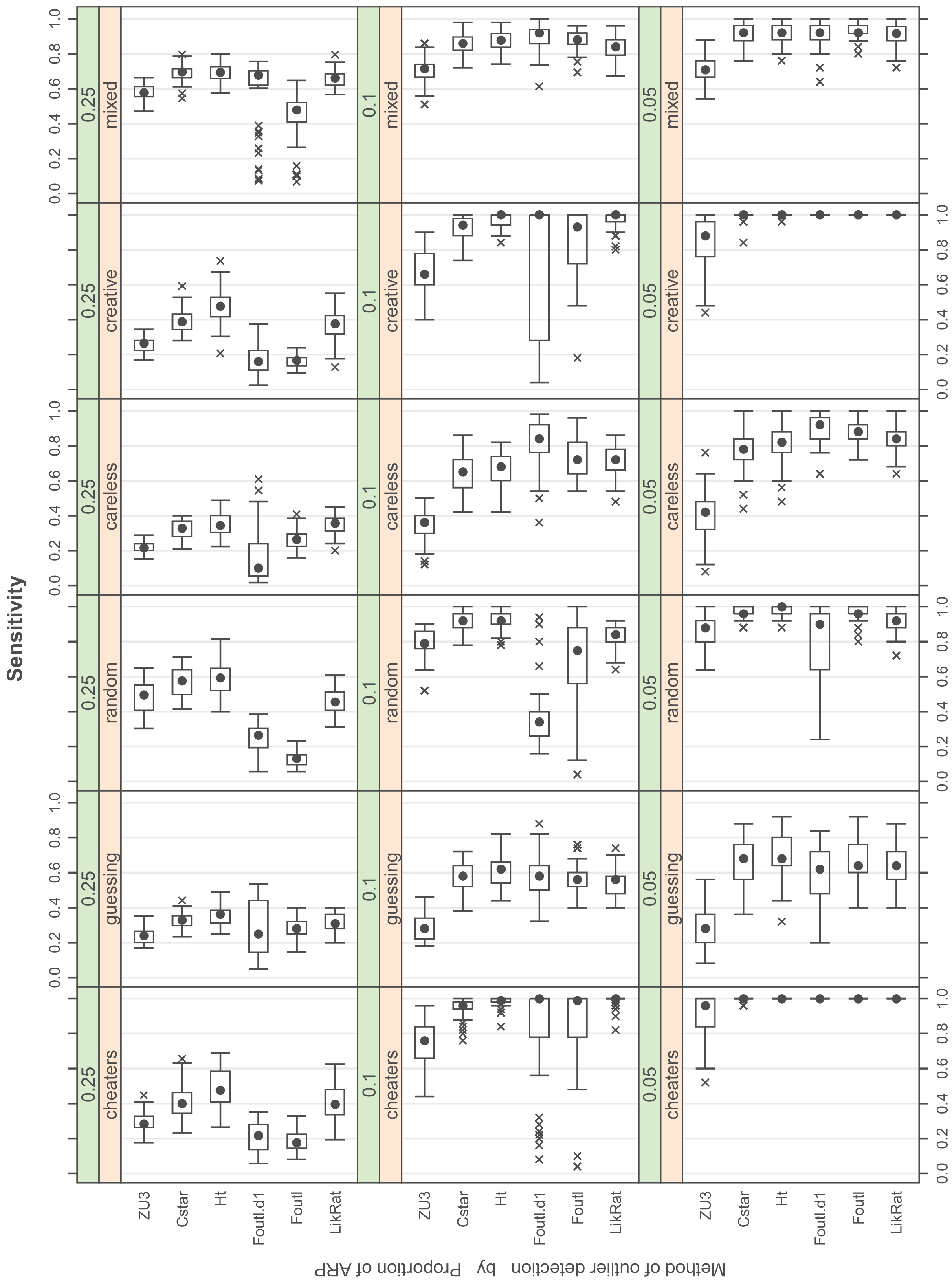

Figure 8: Simulation results for $n = 500$: Sensitivity (probability of correctly detecting an ARP).

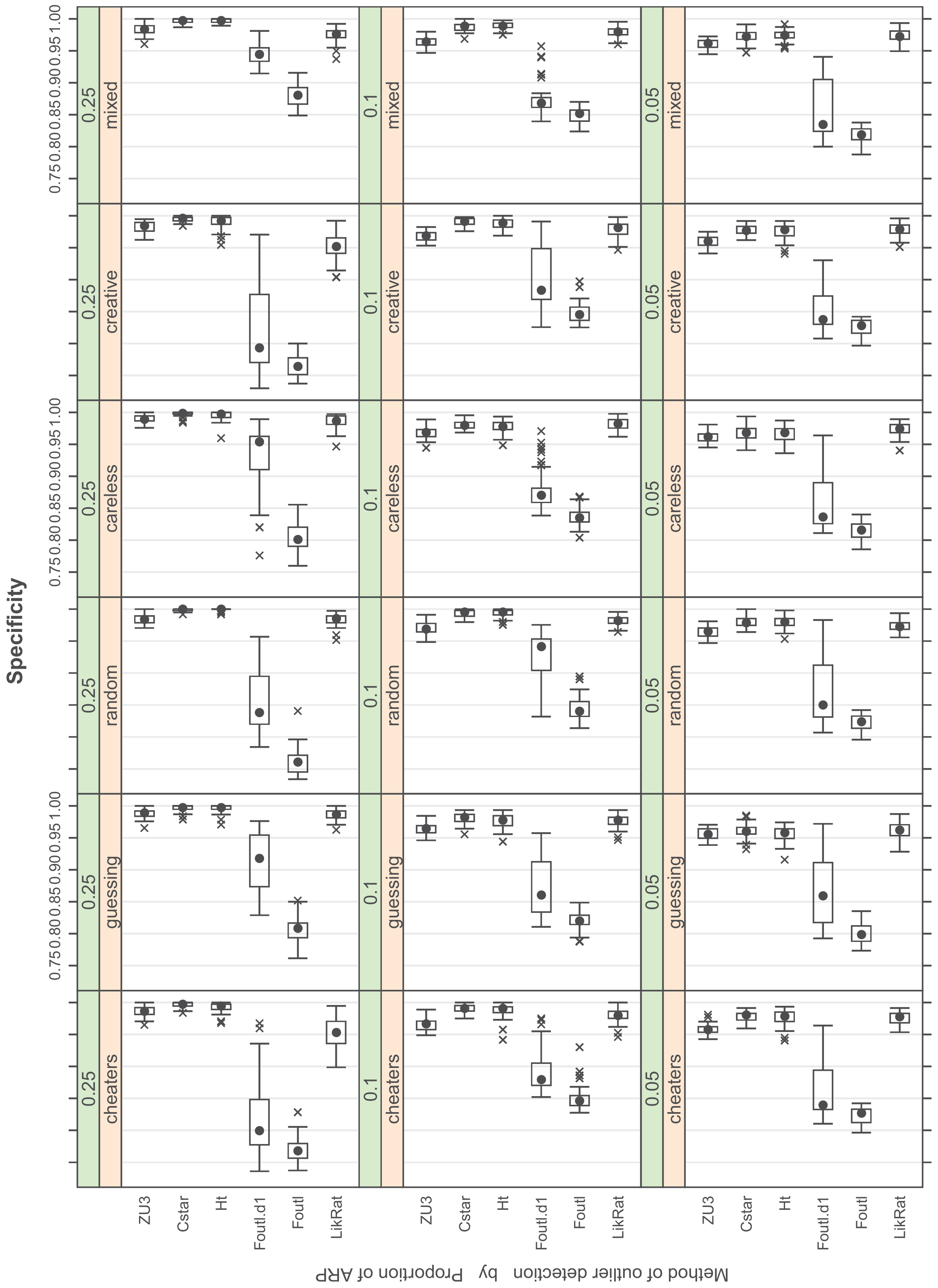

Figure 9: Simulation results for $n = 500$: Specificity (probability of correctly detecting a normal response pattern)

The main results from the simulation study can be summarized as follows.

1. The following general findings were valid for all the ARP detection methods:
    (a) The lower the proportion of ARP, the better its detection.
    (b) Cheaters and creative respondents were the easiest types of ARP to detect. This was particularly clear when looking at the sensitivity. This result was expected, given that cheaters and creative respondents were simulated with the largest deviations from normal patterns.
2. Regarding our suggested methods, the main results were the following:
    (a) In general, the method based on the log likelihood-ratio test statistic (`LikRat`) had the best performance over the proposed new methods.
    (b) For low proportion of ARP (5%), the methods based on functional depth (`Foutl` and `Foutl.d1`) were comparable to `LikRat` in sensitivity (even a little better when detecting careless respondents and mixed ARP), although the specificity was clearly greater for `LikRat`. When 10% were ARP, `Foutl.d1` still had good sensitivity performance for careless respondents and mixed ARP cases (once again, its specificity is worse than that of `LikRat`).
3. These findings allowed us to compare our methods with the person-fit statistics:
    (a) `LikRat`, `Ht`, `Cstar` and `ZU3` had high sensitivity when the proportion of ARP is 5% or 10%, except for guessers or careless respondents. The specificity was always high, even for a 25% proportion of ARP.
    (b) Comparison of these four methods (in both sensitivity and specificity) enabled us to state that `Ht` had the best performance, surpassing by a

narrow margin `Cstar` and `LikRat`, both of which were comparable in quality and outperform `ZU3`.

(c) The sensitivity of `LikRat` was slightly greater than that of `Ht` and `Cstar` when all ARP were either careless or creative respondents. Their specificities were similar. On the other hand, `Ht` and `Cstar` outperformed `LikRat` in the detection of random respondents.

Discussion. The proposed ARP identification methods present better detection rates in conditions with less presence of ARP. This is the usual result obtained with other indices: detection rates decrease as the percentage of ARP increases (e.g., Karabatsos, 2003). Thus, under the simulated condition with a 25% presence of ARP, the detection rate of all ARP types is low. In general, however, under simulated conditions with a relatively low percentage of ARP (5% or 10%) the detection rates increase with all methods and for any type of ARP. This result is in accordance with those reported by Rupp (2013).

Calculating the difference between the log-likelihoods presents the best performance among the proposed methods. Its good functioning is due not only to a high ARP detection rate against false negatives (in general sensitivities above .90), but also to a high detection rate of normal patterns against false positives (specificities above .95). The sensitivity of this index is lower when identifying characteristic patterns of guessers and careless respondents.

This result is not unexpected, since the previously defined characteristic patterns of guessers and careless respondents deviate less from the normal patterns than those of cheaters and creative respondents. However, both cheaters and guessers deviate from the representative function of normal responses in the most difficult items, just as the careless and creative respondents deviate from the representative function of normal responses in the easiest items.

Among the proposed methods, the one based on the log likelihood-ratio test statistic presents the best performance, and is globally comparable to that of the three non-parametric person-fit indices we have used. In particular, it performs similarly to `Cstar`, and both are slightly behind `Ht`.

5.2 Simulation of the identified ARP classification

Methods. A simulation study was conducted to evaluate the identified ARP classification proposal in Section 4.4. We simulated exams using $n = 500$ examinees and $m = 50$ items, with a 10% proportion of mixed type ARP (that is, there are 10 ARP of each type: cheaters, guessers, random, careless respondents, and creative respondents). ARP identification was performed by means of two procedures: our proposal based on the log likelihood-ratio test statistic (`LikRat`, the best one among our proposed methods) and the Ht person-fit statistic (`Ht`, the best one among the considered person-fit statistics). The processes for exam simulation and ARP identification were replicated $S = 500$ times.

We applied clustering methods to classify separately the two sets of identified ARP by either `LikRat` or `Ht` and followed the steps described in Section 4.4. Hierarchical clustering and k-means were used, the optimal number of clusters being chosen according to the within/between distance ratio.

When we analyzed all the simulated exams, it was not possible to replicate for each exam the detailed analysis conducted for the example in Section 4.4. Thus, automatic summaries of the clustering results were required. For each simulated exam, and for the sets of ARP identified by either `LikRat` or `Ht`, cross-tables like those in Tables 2 and 3 were created and taken as the output of the classification processes. The following statistics summarized such cross-tables:

1. *Number of columns K.* This was the optimal number of clusters when doing either k-means or hierarchical clustering. The range of possible values of K was constrained to between two and ten.

2. *Combined purity of each type of ARP.* In order to evaluate when a specific type of ARP, say t_i, was correctly allocated to one of the identified clusters while, at the same time, taking into account whether the cluster where it was mostly allocated was shared or not by other types of ARP, we computed the *ARP combined purity*:

   $$p_{t_i} = \max_{k=1\ldots K} \sqrt{\frac{|t_i \cap c_k|}{|t_i|}\frac{|t_i \cap c_k|}{|c_k|}},$$

   where $|t_i|$ was the number of identified ARP of type t_i, $|c_k|$ was the number of identified ARP allocated to Cluster k, and $|t_i \cap c_k|$ was the number of cases at the intersection. Observe that the first factor inside the square root can be understood as the purity of the cell (t_i, c_k) in its row, and the second factor as its purity in its column. So the combined purity p_{t_i} merged both purity measures by taking their geometric average.

   To fix ideas, the combined purity of *cheaters* in Table 2 (k-means) was

   $$p_{t_2}^{km} = \max\left\{0, 0, \sqrt{\frac{3}{10}\frac{3}{3}}, \sqrt{\frac{7}{10}\frac{7}{15}}, 0, 0\right\} = \sqrt{\frac{7}{10}\frac{7}{15}} = 0.57$$

   which was attained at the Cluster 4. On the other hand, the combined purity of *cheaters* in Table 3 (hierarchical clustering) was

   $$p_{t_2}^{hcl} = \max\left\{0, 0, \sqrt{\frac{10}{10}\frac{10}{26}}, 0\right\} = 0.62$$

   which was attained at Cluster 2. Thus, *cheaters* were classified better by hierarchical clustering than by k-means.

Table 5: Summary statistics of the $S = 500$ values of the optimal number of clusters K, for the four combinations of ARP identification and clustering methods.

Method for ARP identification	Method for clustering	Median	Mean	Mode	Maximum
`Ht`	k-means	3	3.174	2	9
`Ht`	Hierarchical	2	2.156	2	4
`LikRat`	k-means	4	4.288	3	10
`LikRat`	Hierarchical	3	2.792	3	5

3. *Chi-squared distance between types of ARP.* In order to determine which types of ARP tended to be classified together, we computed the chi-square distance between the rows of the cross-tables (e.g., Greenacre, 2016):

$$d_{\chi^2}(t_i, t_j) = \sqrt{\sum_{k=1}^{K} \frac{1}{|c_k|} \left(\frac{|t_i \cap c_k|}{|t_i|} - \frac{|t_j \cap c_k|}{|t_j|} \right)^2}.$$

Results. Regarding the optimal number of clusters, K, Table 5 shows the summary statistics of $S = 500$ values of K on the simulated exams for the four combinations of methods for ARP identification and clustering. It can be seen that, roughly speaking, the sets of ARP identified by `LikRat` required one more cluster than those identified by `Ht`. Otherwise, hierarchical clustering tended to give fewer clusters than k-means (approximately one less cluster).

Given that the results on combined purity were similar for both `LikRat` and `Ht` ARP identification methods, only those referred to `LikRat` are reported here. Figure 10 summarizes the combined purity for each type of ARP as box-plots of the $S = 500$ obtained values. The main findings are the following. Firstly, the non-ARP type (that is, those examinees wrongly identified as ARP) tended to be classified in one cluster where these kinds of cases were mainly allocated. This is even more clear when hierarchical clustering was used. Secondly, k-means

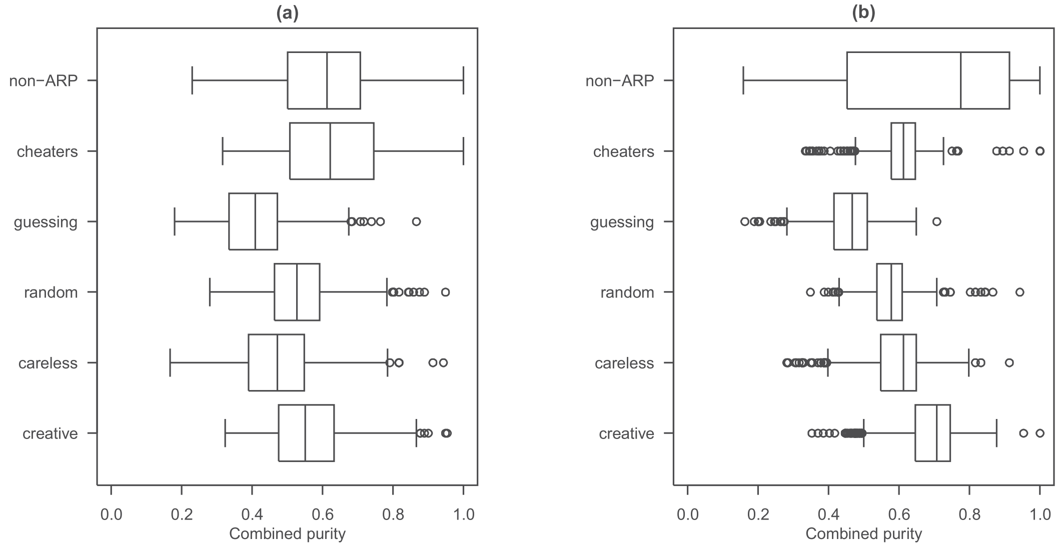

Figure 10: Average combined purity for each type of ARP in (a) k-means and (b) hierarchical clustering. ARP detection is conducted with the method based on log likelihood-ratio.

provided clusters with high combined purity for cheater respondents, whereas creative respondents were better classified by hierarchical clustering. Lastly, guesser respondents had the lowest representation quality in both clustering methods.

The matrices containing the average of chi-square distances between types of ARP throughout the $S = 500$ simulated exams are reported as color maps in Figure 11. Again, we show only the results from `LikRat` because those from `Ht` were similar. The main result was that hierarchical clustering provided fewer clusters, but they were separated better than by k-means. The observed distances for k-means seemed to indicate that there were four groups of ARP types (non-ARP, cheaters, guessing-random, careless-creative), while three groups weare clearly detected in the hierarchical clustering results (non-ARP, cheaters-guessing-random, careless-creative). These results corroborated the previous findings on the average number of clusters in Table 5. Finally, the non-ARP type was the furthest

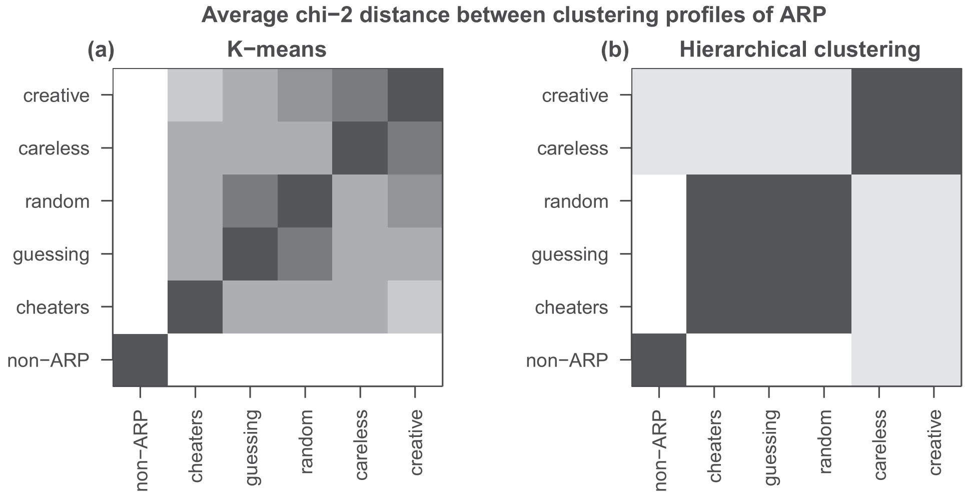

Figure 11: Average chi-squared distance between clustering profiles of ARP obtained by (a) k-means and by (b) hierarchical clustering. ARP detection was done with the method based on log likelihood-ratio. Black corresponds to the lowest distances and white to the greatest.

one (confirming the findings about the combined purity) although, in hierarchical clustering, non-ARP slightly resembled creative and careless respondents.

Discussion. No clustering method perfectly classifies the six types of responses considered. The k-means method tends to propose more (and less pure) clusters than the hierarchical method. In both cases, different types of ARP are combined in the same cluster in a logical way. The hierarchical clustering method is good at distinguishing clusters with response patterns that are mainly associated with (a) spuriously high scores (cheaters, guessers and random respondents), (b) spuriously low scores (careless and creative respondents), and (c) those with non-ARP (wrongly identified as such), which have a certain similarity to careless and creative respondents. The results obtained by the k-means method are less clear but still coherent. There is one pure cluster for non-ARP identified as such, another that is not so pure for cheaters, and two more that mainly group together guessing with random respondents, and careless with creative respondents, respectively. Once again, the responses associated with spuriously low scores tend to be grouped together, but now the responses associated with spuriously high scores are less drawn together than in hierarchical clustering.

6 Empirical example

We applied our proposed methods for identifying and classifying ARP to the responses of 600 students to 32 items on a Grade 12 science assessment test (SAT12) which measured their knowledge on the topics of chemistry, biology, and physics. These data are available at the `mirt` R package (Chalmers, 2012) as the data set named `SAT12`, and were obtained from the TESTFACT software manual (Wood et al., 2003). The original answers were transformed into binary, considering missing values to be wrong answers, and we followed the advice in the `mirt` R package about the potentially better scoring for Item 32. We removed the Cases 496 (with correct answers for all items) and 50 (a perfect Guttman pattern with only one wrong answer on the most difficult item).

Figure 12 describes this data set. It shows (a) the empirical cumulative distribution functions (ECDF) of the examinee proportion of correct answers (this variable seemed to be close to normality), and (b) the item proportion of examinees who answered it incorrectly (quite close to uniformity). These two features were used to estimate examinee abilities θ_i and item difficulties b_j, respectively, in accordance with Section 3. The PRF was then estimated for each examinee, as explained in Section 4.2.

In addition to the PRF, the IPRS for each examinee was also estimated as in Section 4.2. Both sets of functions were used for ARP detection by means of

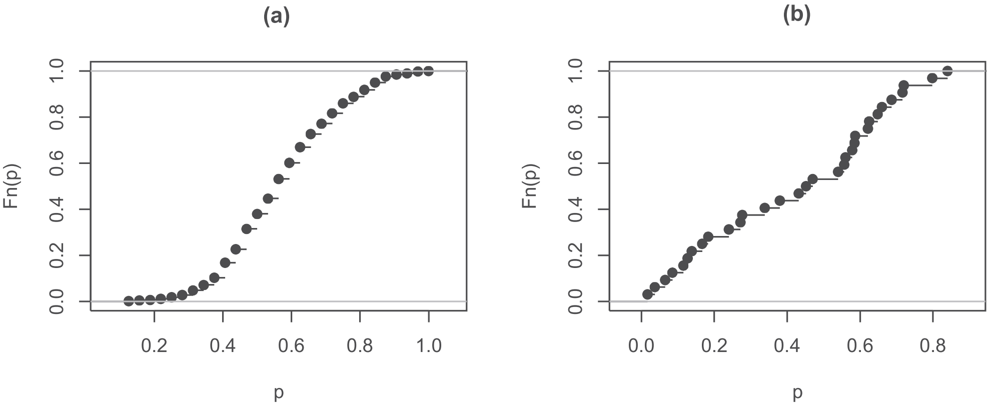

Figure 12: Description of the SAT12 dataset. ECDF of (a) examinee proportion of right answers and (b) item proportion of examinees who answered incorrectly.

the `LikRat` method, as explained in Section 4.3. The person-fit statistics Ht was also used for ARP identification. Among the 598 considered examinees, 86 were identified as ARP (30 only by `LikRat`, 39 only by Ht, and 17 by both methods).

The 86 cases identified as ARP were used in the classification step (as in Section 4.4). Both k-means and hierarchical clustering were applied, and the number of automatically determined clusters for both of them was three. The automatic selection for k-means yielded one very large cluster that was distributed nearly uniformly across the clusters of the hierarchical clusters. To break up this mixed cluster, we explored the k-means solution for $k = 4$, and the obtained solution strongly agreed with that of hierarchical clustering with three clusters, as can be seen in Table 6.

We now describe the ARP classification results corresponding to k-means with $k = 4$ and to hierarchical clustering with three clusters. Figure 14 shows the inverse logit transformation of the average PRF logit transformation for the individuals in each cluster. Table 6 and Figure 14 indicate that there were essentially three clusters in the SAT12 dataset:

Table 6: *Clustering of 86 curves identified as ARP in the SAT12 dataset. Crossing the assigned cluster by k-means and hierarchical clustering. The number of clusters for k-means is four. The cluster numbers correspond to those in Figure 14.*

	Hierarchical clustering		
k-means	1	2	3
1	18	0	0
2	2	0	0
3	1	39	0
4	0	11	15

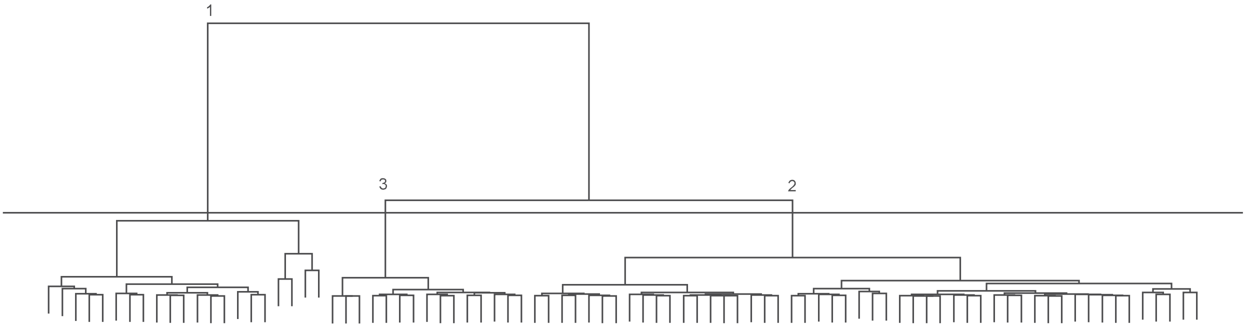

Figure 13: Dendogram for the hierarchical clustering applied to the 86 curves identified as ARP in the SAT12 dataset.

Cluster A is represented with dotted curves in Figure 14. These curves underwent a sudden decline (far more pronounced than the average) resembling perfect Guttman patterns. They were detected as ARP by `LikRat`, but not by Ht. This was a single group in hierarchical clustering (Cluster 1), but it became divided into two in k-means with $k = 4$ (Clusters 1 and 2).

Cluster B , represented with long-dashed curves in Figure 14, was composed by individuals with lower than average probabilities of giving the right answer to the items of low and medium difficulty, but the opposite happened for the most difficult items. A common characteristic was that the curves in this

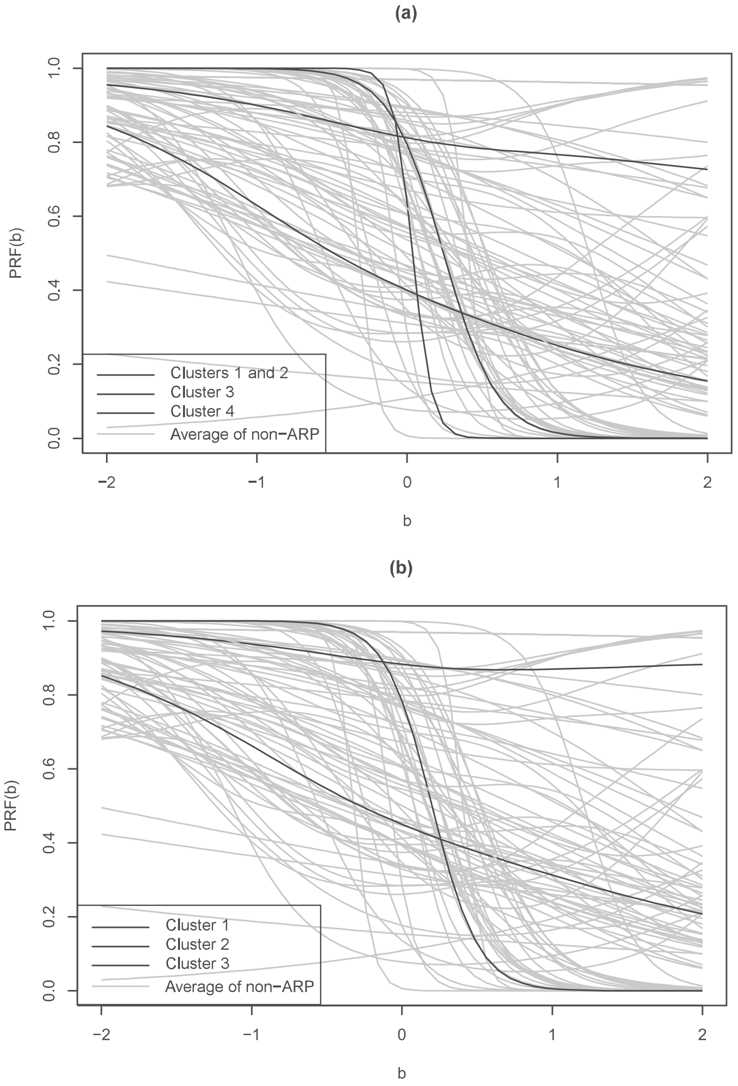

Figure 14: Clustering of 86 curves identified as ARP in the SAT12 dataset. (a) k-means results with $k = 4$ clusters. (b) Hierarchical clustering, and an automatically selected number of clusters equal to 3.

cluster decreased more slowly than the average. These examinees shared certain characteristics with careless and creative respondents.

Cluster C is represented with 2-dotted curves in Figure 14. It corresponded to individuals with greater than average probabilities of giving the right answer to the items with medium and high difficulty, but the opposite happened for the easiest items. As in the previous cluster, the curves in this cluster decreased more slowly than the average. These examinees shared certain characteristics with cheaters and guessers.

7 General discussion

An ARP detection and classification methodology is presented based on PRF. Regarding identification, our simulation experiments reveal that the approach denoted as `LikRat` outperforms the other two, and that its ability to identify ARP is comparable to other person-fit scalar statistics such as Sijtsma's Ht, Harnish Linn's Modified Caution Index and van der Flier's ZU3. This certainly does not mean that they provide similar lists of identified ARP. In fact Ht overlaps with `LikRat` sensibly less than with other person-fit statistics. We agree with Sijtsma and Meijer (2001), when they highlight that analyzing PRF from an FDA perspective constitutes a powerful tool for the diagnosis of aberrance, not only for ARP identification but also for their later classification. Although the functional approach is more complex than the computation of standard person-fit indices, examining the entire PRF is much more informative than summarizing it in scalar indices. Taking advantage of this fact, we propose a functional classification procedure in which the identified ARP are clustered. Despite its functional nature, our classification proposal can also be fed by ARP identified by any standard person-fit index.

Regarding classification of identified ARP, we have used functional distances to perform functional k-means and hierarchical clustering. In both cases, the number of clusters was chosen automatically. Our results indicate that k-means tends to identify more clusters than hierarchical clustering, and also that it is less stable than the latter because of its dependence on random initial centers for the k clusters.

We recommend professionals to follow the next steps (cf. empirical example):

1. Perform descriptive analyses (as in Figure 12) to visualize the distributions of abilities and difficulties, and to check if the 1PL model is plausible.
2. Estimate the IPRS and the PRF non-parametrically.
3. Flag all ARP cases identified by either `LikRat` or Ht.
4. Classify the flagged ARP using both functional k-means and hierarchical clustering, with automatic determination of the number of clusters.
5. Describe the final clusters representing the average curves as in Figure 14.

This strategy enables to distinguish patterns that are associated with spuriously low scores from those associated with spuriously high scores. It even allows different types of of ARP to be detected among the high-scoring examinees.

It is possible to extend this work in two main directions. On the one hand, the number of items and respondents could be expanded. In our study, we simulated only 50 items and both 200 and 500 respondents. Subsequent studies should be carried out to analyze the extent to which the results of the study can be generalized to other evaluation conditions. On the other hand, in addition to the difficulty of the items, future analyses should consider their discrimination and the probability of random responses. Our proposal can be generalized to the 2PL model as follows. Consider that items can vary in difficulty b and discrimination a, with $(b, a) \in \Omega \subseteq \mathbb{R}^2$. In this case the IPRS p depends on three arguments: $p(\theta, b, a)$. The (b_j, a_j) for item j, $j = 1, \ldots, m$, can be estimated non-parametrically using, respectively, the mean item score and the item scalability coefficient H_j (Sijtsma, 2005). Estimating $p(\theta, b, a)$ is possible by fitting a non-parametric binary regression model to the data $(\hat{\theta}_i, \hat{b}_j, \hat{a}_j, x_{ij})$, $i = 1, \ldots, n$, $j = 1, \ldots, m$. The PRF for individual i would then depend on b and

a, PRF(b, a), which can be estimated from the data $(\hat{b}_j, \hat{a}_j, x_{ij})$, $j = 1, \ldots, m$, using a non-parametric binary regression model. The differences, in the logit scale, of individual PRF and the IPRS profiles, would then depend on two parameters, $D_i(b, a)$. Detecting ARP based on the log-likelihood ratio test statistic would follow exactly the same structure as in our proposal for one parameter, because loglik.PRF$_i$ and loglik.IPRS$_i$ are defined in the same way. Generalizing it to the 3PL model would also be possible, provided that a non-parametric estimate of the pseudo-guessing parameter is available.

References

American Educational Research Association, American Psychological Association, National Council on Measurement in Education [AERA, APA, NCME]. (2014). *Standards for educational and psychological testing.* Washington, DC: American Psychological Association.

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*(6), 1–29. doi: 10.18637/jss.v048.i06

Cuevas, A., Febrero-Bande, M., & Fraiman, R. (2007). Robust estimation and classification for functional data via projection-based depth notions. *Computational Statistics*, *22*(3), 481–496.

Emons, W. H. M., Sijtsma, K., & Meijer, R. R. (2004). Testing hypotheses about the person-response function in person-fit analysis. *Multivariate Behavioral Research*, *39*(1), 1–35.

Emons, W. H. M., Sijtsma, K., & Meijer, R. R. (2005). Global, local and graphical person-fit analysis using person response functions. *Psychological Methods*, *10*(1), 101–119.

Febrero-Bande, M., Galeano, P., & González-Manteiga, W. (2008). Outlier detection in functional data by depth measures, with application to identify abnormal nox levels. *Environmetrics*, *19*(4), 331–345.

Febrero-Bande, M., & Oviedo de la Fuente, M. (2012). Statistical computing in functional data analysis: the R package `fda.usc`. *Journal of Statistical Software*, *51*(4), 1–28.

Greenacre, M. (2016). *Correspondence analysis in practice* (3rd ed.). New York: Chapman and Hall/CRC.

Hennig, C. (2018). fpc: Flexible procedures for clustering [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=fpc` (R package version 2.1-11.1)

Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty-six person-fit statistics. *Applied Measurement in Education*, *16*, 277–298.

Kokoszka, P., & Reimherr, M. (2017). *Introduction to functional data analysis.* Boca Raton, FL: CRC Press.

Meijer, R., & Sijtsma, K. (2001). Methodology review: Evaluating Person Fit. *Applied Psychological Measurement*, *25*, 107–135.

Nering, M., & Meijer, R. (1998). A comparison of the person response function and the lz person-fit statistic. *Applied Psychological Measurement*, *22*, 53–69.

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

Ramsay, J. O. (2000). TestGraf. a program for the graphical analysis of multiple choice test and questionnaire data [Computer software manual]. Retrieved from `http://www.psych.mcgill.ca/faculty/ramsay.html`

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

Rupp, A. A. (2013). A systematic review of the methodology for person fit research in item response theory: Lessons about generalizability of inferences from the design of simulation studies. *Psychological Test and Assessment Modeling*, *55*(1), 3.

Sijtsma, K. (2005). Nonparametric item response theory models. *Encyclopedia of social measurement*, *2*, 875–882.

Sijtsma, K., & Meijer, R. (2001). The person response function as a tool in person-fit research. *Psychometrika*, *66*(2), 191–208.

Tendeiro, J. N., Meijer, R. R., & Niessen, A. S. M. (2016). PerFit: An R package for person-fit analysis in IRT. *Journal of Statistical Software*, *74*(5), 1–27. doi: 10.18637/jss.v074.i05

Trabin, T., & Weiss, D. J. (1983). The person response curve: Fit of indivduals to item characteristic curve models. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 83–108). New York, NY: Academic Press.

Walker, A., Engelhard, G., Hedgpeth, M.-W., & Royal, K. (2016). Exploring aberrant response patterns using person fit and person response functions. *Journal of applied measurement*, *17*(2), 194–208.

Wood, R., Wilson, D., Gibbons, R., Schilling, S., Muraki, E., & Bock, R. (2003). Testfact 4 for windows: Test scoring, item statistics, and full-information item factor analysis [computer software] [Computer software manual].

Wood, S. N. (2017). *Generalized additive models: An introduction with r* (2nd ed.). Boca Raton, FL: Chapman and Hall/CRC.