PLOS ONE

Identifying Faking to Single Item Responses in Personality Tests: a New TF-IDF Based Method

--Manuscript Draft--

Manuscript Number:	
Article Type:	Research Article
Full Title:	Identifying Faking to Single Item Responses in Personality Tests: a New TF-IDF Based Method
Short Title:	Identifying Faking to Single Item Responses
Corresponding Author:	Giuseppe Sartori University of Padova Padova, ITALY
Keywords:	Faking detection; classification; Statistical Modeling
Abstract:	We wish to submit to the PlosOne journal a paper entitled: "Identifying Faking to Single Item Responses in Personality Tests: a New TF-IDF Based Method". Faking to a psychological test is often observed whenever an examinee may gain an advantage from it. While techniques are available to identify a faker, they cannot identify the specific questions distorted by faking. In this work, we evaluate the effectiveness of Term Frequency-Inverse Document Frequency (TF-IDF) { an information retrieval mathematical tool used in search engines and language representations { in identifying single-item faked responses. We validate the proposed technique on three different datasets containing responses to the 10 items Big Five questionnaire (total number of 694 participants) in different faking situations. Each participant responded twice, once honestly and once faking in order to achieve an objective in one of three different contexts (one to obtain a child custody and two to obtain a position at different companies). The proposed TF-IDF model was proven very effective in separating honest from faked responses { with honest having low TF-IDF values and fakers higher ones { and also in identifying which of the 10 responses to the questionnaire was distorted in the faking condition. Examples of the usage of the technique in single case evaluation are provided to show how the technique may be used for single case analysis in practical settings.
Order of Authors:	Alberto Purpura
	Dora Giorgianni
	Graziella Orrù
	Giulia Melis
	Giuseppe Sartori
Opposed Reviewers:	
Additional Information:	
Question	**Response**
Financial Disclosure Enter a financial disclosure statement that describes the sources of funding for the work included in this submission. Review the submission guidelines for detailed requirements. View published research articles from *PLOS ONE* for specific examples. This statement is required for submission	The authors received no specific funding for this work.

and **will appear in the published article** if the submission is accepted. Please make sure it is accurate. **Unfunded studies** Enter: *The author(s) received no specific funding for this work.* **Funded studies** Enter a statement with the following details: • Initials of the authors who received each award • Grant numbers awarded to each author • The full name of each funder • URL of each funder website • Did the sponsors or funders play any role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript? • **NO** - Include this sentence at the end of your statement: *The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.* • **YES** - Specify the role(s) played. * typeset	
Competing Interests Use the instructions below to enter a competing interest statement for this submission. On behalf of all authors, disclose any competing interests that could be perceived to bias this work—acknowledging all financial support and any other relevant financial or non-financial competing interests. This statement is required for submission and **will appear in the published article** if the submission is accepted. Please make sure it is accurate and that any funding sources listed in your Funding Information later in the submission form are also declared in your Financial Disclosure statement. View published research articles from *PLOS ONE* for specific examples.	The authors have declared that no competing interests exist.

NO authors have competing interests Enter: *The authors have declared that no competing interests exist.* **Authors with competing interests** Enter competing interest details beginning with this statement: *I have read the journal's policy and the authors of this manuscript have the following competing interests: [insert competing interests here]* * typeset	
Ethics Statement Enter an ethics statement for this submission. This statement is required if the study involved: • Human participants • Human specimens or tissue • Vertebrate animals or cephalopods • Vertebrate embryos or tissues • Field research Write "N/A" if the submission does not require an ethics statement. General guidance is provided below. Consult the submission guidelines for detailed instructions. **Make sure that all information entered here is included in the Methods section of the manuscript.**	All procedures performed in studies involving human participants were in accordance with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. Informed consent was obtained from all individual participants included in the study.

Format for specific study types **Human Subject Research (involving human participants and/or tissue)** • Give the name of the institutional review board or ethics committee that approved the study • Include the approval number and/or a statement indicating approval of this research • Indicate the form of consent obtained (written/oral) or the reason that consent was not obtained (e.g. the data were analyzed anonymously) **Animal Research (involving vertebrate animals, embryos or tissues)** • Provide the name of the Institutional Animal Care and Use Committee (IACUC) or other relevant ethics board that reviewed the study protocol, and indicate whether they approved this research or granted a formal waiver of ethical approval • Include an approval number if one was obtained • If the study involved *non-human primates*, add *additional details* about animal welfare and steps taken to ameliorate suffering • If anesthesia, euthanasia, or any kind of animal sacrifice is part of the study, include briefly which substances and/or methods were applied **Field Research** Include the following details if this study involves the collection of plant, animal, or other materials from a natural setting: • Field permit number • Name of the institution or relevant body that granted permission	
Data Availability Authors are required to make all data underlying the findings described fully available, without restriction, and from the time of publication. PLOS allows rare exceptions to address legal and ethical concerns. See the PLOS Data Policy and FAQ for detailed information.	Yes - all data are fully available without restriction

A Data Availability Statement describing where the data can be found is required at submission. Your answers to this question constitute the Data Availability Statement and **will be published in the article**, if accepted. **Important:** Stating 'data available on request from the author' is not sufficient. If your data are only available upon request, select 'No' for the first question and explain your exceptional situation in the text box. Do the authors confirm that all data underlying the findings described in their manuscript are fully available without restriction?	
Describe where the data may be found in full sentences. If you are copying our sample text, replace any instances of XXX with the appropriate details. • If the data are **held or will be held in a public repository**, include URLs, accession numbers or DOIs. If this information will only be available after acceptance, indicate this by ticking the box below. For example: *All XXX files are available from the XXX database (accession number(s) XXX, XXX.*). • If the data are all contained **within the manuscript and/or Supporting Information files**, enter the following: *All relevant data are within the manuscript and its Supporting Information files.* • If neither of these applies but you are able to provide **details of access elsewhere**, with or without limitations, please do so. For example: *Data cannot be shared publicly because of [XXX]. Data are available from the XXX Institutional Data Access / Ethics Committee (contact via XXX) for researchers who meet the criteria for access to confidential data.* *The data underlying the results presented in the study are available from (include the name of the third party*	Data and materials: The datasets generated during and/or analysed during the current study are available in our public repository: https://github.com/albpurpura/TFIDF-Faking Code availability: Available in our public repository: https://github.com/albpurpura/TFIDF-Faking

and contact information or URL). • This text is appropriate if the data are owned by a third party and authors do not have permission to share the data. * typeset	
Additional data availability information:	

Dear Editors and Reviewers,

We wish to submit to the *PlosOne* journal a paper entitled: "Identifying Faking to Single Item Responses in Personality Tests: a New TF-IDF Based Method".

We confirm that this work is original and has not been published elsewhere, nor is it currently under consideration for publication elsewhere.

In this paper, we evaluate the effectiveness of Term-Frequency-Inverse Document Frequency (TFIDF) - an information retrieval mathematical tool used in search engines and language representation - in identifying single-item faked responses, a previous unaddressed extremely important problem. We validate the proposed technique on three different datasets containing responses to the 10 items Big Five questionnaire (total number of 694 participants) in different faking situations. Each participant responded twice, once honestly and once faking to achieve an objective in one of three different contexts (one to obtain a child custody and two to obtain a position at different companies). The TF-IDF model efficiently separates honest from faked responses with honest having low TF-IDF values and fakers higher ones. The proposed TF-IDF technique was effective also in identifying which of the 10 responses to the questionnaire was distorted in the faking condition, obtaining an average precision of 69%. Examples of the usage of the technique in single case evaluation are provided to show how the technique may be used for single case analysis in practical settings.

We believe of *PlosOne* is the right venue for this work because of its focus on approaches employing computers for psychological research, as well as for its interest in theoretical and applied research.

Thank you for your consideration,

Best Regards,

Alberto Purpura, Dora Giorgianni, Graziella Orrù, Giulia Melis, Giuseppe Sartori

Identifying Faking to Single Item Responses in Personality Tests: a New TF-IDF Based Method

Alberto Purpura[1], Dora Giorgianni[2], Graziella Orrù[3], Giulia Melis[2], Giuseppe Sartori[2*]

1 Department of Information Engieering, University of Padua, Padua, Italy
2 Department of General Psychology, University of Padua, Padua, Italy
3 Department of Surgical, Molecular Critical Area Pathology, University of Pisa, Medical, Pisa, Italy

* giuseppe.sartori@unipd.it
Emails: purpuraa@dei.unipd.it (Alberto Purpura), dora.giorgianni1@gmail.com (Dora Giorgianni), graziella.orru@unipi.it (Graziella Orrù), giulia.melis@phd.unipd.it (Giulia Melis)

Abstract

Faking to a psychological test is often observed whenever an examinee may gain an advantage from it. While techniques are available to identify a faker, they cannot identify the specific questions distorted by faking. In this work, we evaluate the effectiveness of Term Frequency-Inverse Document Frequency (TF-IDF) – an information retrieval mathematical tool used in search engines and language representations – in identifying single-item faked responses. We validate the proposed technique on three different datasets containing responses to the 10 items Big Five questionnaire (total number of 694 participants) in different faking situations. Each participant responded twice, once honestly and once faking in order to achieve an objective in one of three different contexts (one to obtain a child custody and two to obtain a position at different companies). The proposed TF-IDF model was proven very effective in separating honest from faked responses – with honest having low TF-IDF values and fakers higher ones – and also in identifying which of the 10 responses to the questionnaire was distorted in the faking condition. Examples of the usage of the technique in single case evaluation are provided to show how the technique may be used for single case analysis in practical settings.

Keywords: Faking detection, Classification, Statistical Modeling

Introduction

In this work, we focus on a crucial problem in the interpretation of data collected through questionnaires, faking detection.

Within this context, responses to direct questions are easily faked when an examinee has a direct advantage in doing so. Distorsions in responding to direct questions may take two different forms: faking-bad and faking-good. Faking-bad characterizes some forensic settings (e.g., criminal, insurance claims) in which the examinee is likely to exaggerate or make up his/hers psychological disorder [1]. By contrast, faking-good is usually observed in a setting in which the respondent is expected to give highly desirable responses. In psychological questionnaires faking is usually controlled via the use of so-called control scales (e.g., MMPI II [2] and MCM III - Millon Clinical Multiaxial Inventory [3] that are among the most used

tests to evaluate psychiatric disorders) that tap the propensity of the respondent to depict a socially desirable profile and his propensity to report hyperbolic disorders. The abnormal score in these control scales is taken as evidence of an overall tendency of the subject to modulate his responses in the direction of socially desirability (fake good) or in the direction of none existent or hyperbolic psychopathology (fake bad). In general, faking is a continuous variable and the level of faking is modulated by the stake and by the strategy under implicit or explicit control by the respondent.

For this reason, efforts have been made to develop specific tests that flag the responder as a faker. While such procedures may spot the faker with a sufficient accuracy, to the best of our knowledge no procedure has been proposed to reconstruct the honest response profile once a faker has been identified and only the faked profile is available. In short, a non-depressed subject who wants to appear as depressed may be spotted as a faker, however there is no valid procedure that may be used to uncover his/her true level of depression resulting from honest responses. Suppose, for example, that the true depression level of the responder is 0.2 standard deviations above the mean of the depression scale. However, because of feigning, he appears to be 2.5 standard deviations above the mean. Available techniques today may flag that s/he is a faker but no technique is available that derives from the observed standard score of 2.5, the real score (unobserved 0.2).

In this paper, we have analyzed the potential of a standard information representation model developed within the information retrieval field, i.e. Term Frequency-Inverse Document Frequency (TF-IDF) [4], to identify the items that were faked by test takers. The proposed model has been validated on the 10 items Big Five Test [5]. We administered this questionnaire to a group of volunteers who were instructed to first answer it honestly, and then to intentionally fake their responses according to three different settings: i) to obtain a child custody in the context of a litigation case; ii) in the context of a job interview for a salesperson position, iii) in the context of a job interview for a position in a humanitarian organisation. Three faking contexts were taken into account since faking objectives are known to modulate how and where the respondent is faking.

The proposed approach to compute TF-IDF representations for questionnaires responses and the data employed for its evaluation is available in our public repository: `https://github.com/albpurpura/TFIDF-Faking`.

The original contributions of this work are:

- the proposal of a novel approach to precisely identify faking patterns in personality tests;
- a thorough evaluation of the proposed approach based on TF-IDF, with a comparison to similar techniques based on the distribution of raw response values;
- the creation of an experimental collection for the development of new item-level faking detection approaches.

Related Work

In both classical test theory and in Item Response Theory (IRT) the evaluation of the worth of an item is carried out with respect to the performance on that item of a sample of participants that constitute the validation sample. In both these theories, the overall *responding style* of a participant is not modelled. Responding style refers for example to the tendency of some participants to use the full range of response alternatives while others may have a bias toward high or low values of the scale.

TF-IDF – described in more detail below – has the potential for overcoming some limitations of both the classical test theory and the Item Response Theory. Indeed, neither classical test theory or IRT model the response style of test takers, focusing instead mainly on the relative position of a respondent specific response

with respect to the distribution of responses of subjects from the normative sample. By contrast, TF-IDF combines both the relative positioning of a response with respect to the validation sample, as well as the response style of a specific subject [6].

Proposed Approach

TF-IDF is a standard term-weighting method used in information retrieval to compute a numeric representation of textual data based on the distribution of the terms occurring in it. It is one of the mathematical tools used by search engines to retrieve web pages relevant to a user query, and rank them according to their relevance. It is also used as a text representation strategy in natural language processing. [6]

The TF-IDF representation strategy is based on the intuition that a certain term in a document d is representative for its general meaning proportionally to (i) its Term Frequency (TF) within d, and (ii) its inverse frequency in the collection (IDF), i.e. a term which frequently appears in d and at the same time is only rarely used in others, is more representative for the content of d than different ones which are either not frequently used in d or are very frequently used also in other documents.

Formally, given a collection C of documents, the TF-IDF value for each term t in a document $d \in C$ is calculated as:

$$\text{TF-IDF}(t, d, C) = TF(d, t) * IDF(t, C), \tag{1}$$

where TF(d, t) indicates the number of times a target term t appears in a document d, and IDF is equal to $\log(N/n)$ where N indicates the number of documents in C and n is the number of documents where t is used.

The application of TF-IDF however is not only limited to the Computer Science area and was also previously used for example in cognitive neuropsychology for modelling semantic memory and its disorders [7] as well as modelling brain responses to relevance [8].

TF-IDF for Detecting Faked Item Responses

When applied to item analysis, we compute the TF-IDF of each response in a questionnaire q considering as TF the number of times the current response is provided by the subject. For example, if a subject responded with the value 5 only to one item of our 10 items test questionnaire, then TF(5, q) = 1.

The IDF is computed as $\log(N/n)$ where N is the number of participants who responded to the test and n is the number of times a certain response value was used for the current item, i.e. the number of times participants responded using the value 5 to the specific item. The final score associated with each test item will be the product of the TF and IDF values. This adaptation of TF-IDF will have high values for responses to single questions which are highly atypical. Under this view, TF-IDF behaves as an anomaly detector indexing the very atypical responses at single item level (not at subject level).

In a practical setting, we could employ TF-IDF to spot a faker in the following way. Suppose we administered the same questionnaire, i.e. a 10 item personality test as reported below, to 100 participants and that each item response was represented by a value in a 5 point Likert scale.

Suppose also that we observe that subject number 1 answered with a 5 to questionnaire item number 3. S/he also answered using the value 5 in a total of 7 out of 10 responses. From our data collection experiment, we also observe that the distribution of the responses given to item number 3 of our questionnaire are the following: 300 people answered with a 2, 650 people with a 3 and 50 people with a 5. The TF-IDF score associated to item number 3 of subject 1 is therefore

computed as $7 * \log(1000/50) = 9.11$. This value will be higher for people that use the same – atypical – response many times in a questionnaire. This behavior is observed in faking situations, where a subject is altering his/her response in a consistent way to increase his/her social desirability.

Under this view, faked responses are expected to have an abnormally high TF-IDF when compared to the distribution of TF-IDF of honest responders.

The proposed approach based on TF-IDF is summarized in the pseudo-code below.

```
def tfidf(subject_responses, per_item_idfs,
    per_item_thresholds):
faked_responses_indices = [];
For response_index in range(len(subject_responses)):
current_response = subject_responses[response_index];
TF = count(response, subject_responses);
IDF = per_item_idfs(response_index, current_response);
If TF*IDF > per_item_thresholds[response_index]:
    faked_responses_indices.append(response_index);
Return faked_responses_indices;
```

Code Snippet 1. TF-IDF for single item faking detection

For each response given by a subject to a questionnaire, we compute its corresponding TF and IDF scores as described earlier. Then, if their product is higher than a certain threshold we flag the current response as faked. The IDF values are estimated based on the honest responses we collected from our volunteers. We select the faking detection thresholds as the percentile score in the TF-IDF values distribution for each questionnaire item that maximizes the Precision metric for faking detection on a separate validation set.

In practice, for our experiments we first consider only the subset of honest responses that we collected. Then, we estimate the IDF scores associated to each questionnaire item. Next, we consider a small subset of faked responses – as a validation set – and tune the faking detection threshold for each questionnaire item based on the distribution of these scores. This threshold value is not estimated on the raw TF-IDF scores but on the corresponding percentile in their distribution on a certain question. As an example, assume we observed that the value associated to the 75th percentile in the distribution of TF-IDF values in our validation set is a reliable threshold to distinguish between honest and faked responses for the questionnaire items. Then, we would use the value associated to the 75th percentile in our test set responses distribution to each item to detect fakers to each question.

When we receive a new questionnaire item to evaluate, we rely on the previously computed set of IDF scores on the respective honest responses and on the TF scores computed on the current responses to assign a TF-IDF score to each response. Then, if the TF-IDF value we obtain for a certain test item is higher than a certain threshold – that we validated beforehand as explained above – we report the item as faked. In other words, we rely on the TF-IDF score associated to the responses of a subject to a questionnaire as an indicator of his/her attitude and as a proxy to detect faking behaviors. For this purpose, TF-IDF, as proposed here, acts as an anomaly detector identifying suspicious differences between the to-be-evaluated case and the distribution of answers given by honest responders.

Furthermore, considering different faking contexts helps us exclude potential biases of the proposed TF-IDF approach to specific data distributions or situation-specific behaviors [9], [10].

Experimental Setup

To validate the proposed approach we conducted three experiments where participants were required to respond to the same questionnaire twice. First, they were asked to respond honestly to all of its questions. Then, we instructed them to fake their responses in three different contexts: a Job Interview for a Salesperson Position (JIS), one for a role in a Humanitarian Organization (JIHO) and to obtain a Child Custody (CC) in the context of a litigation. Having the same subject responding in an honest condition and in one of instructed faking is intended to have a ground truth useful for evaluating the accuracy level of the TF-IDF method when identifying faked responses. Indeed, in most practical applications only one set of responses per subject will be available and the goal of the proposed approach will be to spot faking patterns in his/her responses.

Participants

All participants to our data collection experiments were native Italian speakers and questionnaires were administered online. Information about age, gender and education level of the participants were also collected and are reported in Table 1.

Table 1. Number of participants, age and education of participants of the three groups

Group	Number (Women)	Age	Education	Children
CC	243 (185)	43.1	15.5(3)	Yes=137/243
JIS	230 (188)	40.8	15.8 (2.8)	–
JIHO	221 (176)	41.7	15.4 (2.9)	–

Notes. The participants of the three groups were required to fake their honest responses in order to achieve a result in three different contexts: a Child Custody litigation (CC) and two job interviews - for a Salesperson position (JIS) and for a Humanitarian Organization (JIHO). The three faking conditions are intended to capture the effect on response patterns of different faking objectives.

Materials and Methods

In the three experiments described above, participants were asked to respond to the same set of 10 items twice, first honestly and then altering their answers to obtain a personal gain in three different contexts described before.

All available responses were in multiple choice format where participants could choose only one option for each questionnaire item.

Each participant was required to respond estimating his/her agreement using a 5 point Likert scale (where 1 indicates a strong disagreement and 5 a strong agreement) to the Italian version of the following 10 items from the Big Five Inventory [11], [5]. The English version of the statements we considered is reported below:

"I see myself as someone who ...":

1. is outgoing, sociable (extroversion);
2. is reserved (extroversion reversed);
3. tends to find fault with others (agreeableness, reversed);
4. is generally trusting (agreeableness);
5. tends to be lazy (consciousness, reversed)
6. does a thorough job (consciousness);
7. gets nervous easily (emotional stability, reversed);
8. is relaxed, handles stress well (emotional stability);

9. has an active imagination (openness);
10. has few artistic interests (openness, reversed).

Before analyzing the collected data, the responses given to reversed items (signaled above) were complemented so that final scores of all the items were directly comparable.

Faking Instructions

Group 1, CC. Pretend you and your spouse are separating and fighting in court for the custody of your children (imagine to have a child if you don't have one). A court appointed psychologist will evaluate your parental skills and requires you to respond to the following personality questionnaire. Answer the test in a way that makes you emphasise your parental skills hiding behaviours or thoughts that are generally considered negative. Your goal is to get custody of your children.

Group 2, JIS. Pretend you are attending a job interview for the position of Sales Manager for a very prestigious company. It is your dream job and you want to give the best possible image of yourself. Your goal is to obtain the job and to be considered better than the other candidates. Answer in a way that makes you look good, hiding behaviours or thoughts that are generally considered negative for the position you aspire to. Try to give a positive image of yourself, even if it means to lie.

Group 3, JIHO. Similar to the above.

The procedure included, in all the three conditions, a final question aimed at verifying whether the participants read and fully comprehended the instructions before filling in the questionnaire. Participants who did not answer correctly to this final question were not taken into consideration in our analysis. The additional final question of Group 1 (Child custody litigation) was:
What instructions have you received when filling in the questionnaire?

- I always had to tell the truth;
- I received no specific instructions;
- First I had to tell the truth and then I had to lie to get the custody of the children;
- I always had to lie to get the custody of the children;
- First I had to lie to get the custody of the children and then I had to tell the truth.

This quality check was not satisfied by 44 out of 287 participants. For this reason, we included only 221 participants in our analysis, discarding the responses of all participants that answered incorrectly to the above question.

The additional final question for Group 2 and 3 was:
What instructions have you received when filling in the questionnaire?

- I always had to tell the truth;
- I received no specific instructions;
- First I had to tell the truth and then I had to lie to look good in a job interview;
- I always had to lie to make a good impression;
- First I had to lie to look good in a job interview and then I had to tell the truth.

This quality check for Group 2 (JIS), yielded to the exclusion of 44 out of 287 participants, with 243 participants included in the final analysis who were selected on the basis of their correct responses to the final questions. Regarding Group 3 (JIHO), we excluded 42 out of 272 participants including a total of 230 participants in the following evaluation.

Results

In this section, we present the results of the experiments we conducted to validate the proposed approach based on TF-IDF.

Descriptive statistics

We begin presenting a few characteristics of the data we collected. In Table 2, we report the mean and standard deviation of the responses submitted by our volunteers when asked to respond honestly and to fake to obtain a personal gain. If we observe these statistics, it appears that when faking responses in different contexts participants responded differently to different questions. For example, in the JIS dataset, faking does not significantly affect one of the two items related to the agreeableness of a person. We then report the percentage of responses in the faked test that were not changed, i.e. increased (e.g. from 3 to 5) or decreased (e.g. from 4 to 3). Results reported in Table 3 indicate a pattern that is similar in all the three faking conditions. Most interestingly, there is a 13-15% percent of cases in which the responses were decreased with respect to the honest condition. These could be explained by the presence of different faking strategies from deceivers, and by the test retest unreliability of every psychological test – indeed, test retest reliability of this specific 10 item Big Five Questionnaire is 0.72 [5].

As regards to responses correlations reported in the heatmaps of Fig 1, it is clear that correlations between honest and faked responses are ranging around zero. This indicates that there is no an immediate procedure for predicting honest responses from deceptive ones in all the three groups. Taken together, these data indicate that the identification of the specific response that is faked is not a trivial task. Indeed, correlations between faked and honest responses are very low and when changing their responses fakers alter them in the akin direction in only about 50% of their responses.

Fig 1. Correlation matrices between honest and faked items in the CC, JIS and JIHO datasets, respectively from left to right. Correlation between responses in th honest and Faking condition is virtually nonexistent rendering the identification of honest responses from faked responses a difficult task.

Table 2. Mean and Standard Deviation (SD) of the questions in each honest (H) and faked (F) questionnaire in different datasets.

Dataset	Measure		Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10
	Mean	(H)	3.87	2.43	2.96	3.42	3.17	4.48	2.94	3.01	3.54	3.32
CC	Mean	(F)	4.47*	2.20*	3.90*	3.69*	4.41*	4.58	4.35*	4.23*	3.40	3.98*
	SD	(H)	0.90	0.96	0.99	0.97	1.22	0.68	1.22	1.09	1.10	1.25
	SD	(F)	0.71*	0.93*	0.96*	0.90*	0.85*	0.81	0.80*	0.97*	1.24	0.93*
	Mean	(H)	3.67	2.44	2.95	3.42	3.21	4.47	2.74	2.92	3.63	3.45
JIS	Mean	(F)	4.55*	2.75*	3.46*	3.49	4.52*	4.67*	4.18*	4.40*	4.12*	3.84*
	SD	(H)	0.97	1.09	1.04	1.06	1.20	0.77	1.18	1.14	1.04	1.26
	SD	(F)	0.74*	1.25*	1.18*	1.04	0.82*	0.74*	1.05*	0.86*	1.06*	1.14*
	Mean	(H)	3.77	2.38	3.00	3.55	3.21	4.49	2.81	3.00	3.56	3.42
JIHO	Mean	(F)	4.53*	2.69*	3.94*	3.96*	4.47*	4.57	4.37*	4.35*	4.10*	4.08*
	SD	(H)	0.95	1.04	1.11	0.96	1.25	0.71	1.16	1.18	1.10	1.26
	SD	(F)	0.86*	1.25*	1.11*	0.94*	0.91*	0.86	0.84*	0.88*	1.02*	0.99*

Notes. Values in the faked questionnaires statistics marked with a * indicate a statistically significant difference. Student's t-test with p-value < 0.05 with the honest responses distribution to the same question. Different faking contexts emerge from different modulations of responses to each question. For example, Q2 was changed in one direction when faking Child Custody Litigation (CC) and in the opposite direction when faking a Job interview. Average scores to the 10 questions did not differ significantly in the three groups of honest (H) responders.

Table 3. Percentages of changed responses

Group	Total responses	% Unchanged	% Increased
CC	2210	38 %	48 %
JIS	2430	33 %	52 %
JHO	2300	35 %	52 %

Notes. Percentages of raw score responses in the two conditions (honest and faked) that were the same and changed in the direction as expected by the faking instructions. Instructions were such that increase in raw scores were expected

TF-IDF for Item Analysis

As a first step in our analysis, we evaluate whether the representation of responses induced by TF-IDF is more helpful in discriminating between honest and faked responses compared to what can be achieved using the raw scores. For this purpose, we relied on KL-Divergence which is an entropy-based measure to compare probability distributions. The values of this measure range from 0 to infinity, where 0 indicates that the two considered distribution – e.g. honest and faked – completely overlap. Higher values of KL-Divergence indicate a higher degree of separation between honest and faked responses distributions. The results reported in Tables 4, 5 and 6, show larger KL-Divergence scores between the distributions of honest and faked responses when represented with their TF-IDF value compared with the raw response score – i.e. a number from 1 to 5 – provided by test subjects. This validates our hypothesis of relying on TF-IDF for distinguishing faked from honest responses.

To further evaluate how the TF-IDF representation of honest responses differs from that of faked ones, we calculate the respective average for each participant over all the 10 items. We report these values in Fig 2.

These results indicate that the average TF-IDF for fakers is significantly higher that that for honest responders. For example, in the JIS dataset, the distribution of the average TF-IDF scores associated to each honest questionnaire is for the most part lower than 2 while for faked responses we often observe values on the opposite spectrum of the considered range. Overall, we observe that faking responders are assigned to higher average TF-IDF scores.

Fig 2. Average of TF-IDF response values over all the 10 items for honest and faked responses. Datasets from left to right: CC, JIS, JIHO. The odds for TF-IDF above 3 are the following: 1/3.8 (for every Honest with a TFIDF >3 there are 3.8 Fakers); 1/6.7; 1/4.5.

These results indicate that fakers can be distinguished from honest responders on the basis of the average of the TF-IDF over all the responses to the questionnaire items.

Table 4. KL-Divergence for CC dataset

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10
Raw scores	0.37	0.04	0.54	0.07	1.00	0.08	1.05	0.96	0.03	0.27
TF-IDF	2.93	3.66	3.32	4.51	7.21	0.44	8.21	5.46	4.76	7.30

Notes.KL-Divergence between the response values distributions on the CC dataset of honest and faked responses considering the raw scores representation and the TF-IDF one. The closer to 0 is KL-Divergence the less distinguishable are the two distributions(honest and faked).

Table 5. KL-Divergence for JIS dataset

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10
Raw scores	0.64	0.05	0.21	0.03	0.88	0.12	0.85	1.02	0.24	0.10
TF-IDF	1.29	2.23	1.08	4.38	2.86	0.44	2.65	3.38	2.95	3.28

Notes.KL-Divergence between the response values distributions on the JIS dataset of honest and faked responses considering the raw scores representation and the TF-IDF one.

Table 6. KL-Divergence for JIHO dataset

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10
Raw scores	0.58	0.06	0.44	0.11	0.71	0.13	1.09	0.84	0.16	0.23
TF-IDF	7.18	1.40	2.45	3.99	6.39	0.76	2.20	2.41	1.67	3.25

Notes.KL-Divergence between the response values distributions on the JIHO dataset of honest and faked responses considering the raw scores representation and the TF-IDF one.

Performance Evaluation

When used for item analysis, TF-IDF acts as a "novelty detector" where high values indicate a response that is used frequently by a subject but infrequently by all other ones in a validation sample. The index is participant-specific as well as item-specific. However, not only TF-IDF can be used to spot fakers – as concluded from the results reported in Fig 2 – but may be used also to identify which items have been faked. As already mentioned, this feature is particularly important as – at the time of writing – no procedure is available to address this problem. As described in more detail in Section 3, the proposed process for item-level faking detection can be summarized as:

1. calculate the TF-IDF score for each response of a target participant;
2. compare the obtained TF-IDF scores with a threshold specific to each questionnaire item;
3. if the TF-IDF values are outside of a certain range, then the response is categorized as an anomaly (faked).

We applied this procedure to the three aforementioned datasets – i.e. CC, JIS, JIHO – yielding the results summarized in Fig 3.

Given that each participant responded to the 10 items test twice, one in the honest condition and one in the faked condition, the comparison between items in the two conditions allowed us to identify which were the items with distorted responses for each participant. As indexed by the probability density function, fakers fake more frequently 7/10 items – with few participants faking 1, 2 or 10 items (see Fig 2). We evaluated the performance of the proposed approach first in terms of accuracy (Fig 3), then in terms of Precision, Recall and F1 Score (Table 7). We define the accuracy measure as the number of items in a questionnaire which were correctly identified as faked or not-faked. In Fig 2 we report the average accuracy achieved by our TF-IDF approach in different scenarios where the participants purposely altered a varying number of responses to obtain a personal gain in the previously described conditions.

Fig 3. Probability Density Function (PDF) of the number of faked responses per questionnaire (red line) and per-item accuracy for the faking detection task using the TF-IDF model on different datasets: [left panel] dataset CC; [central panel] dataset JIS; [right panel] dataset JIHO. The most frequent number of faked responses is 7/10. In each bar of the histograms we report the average accuracy obtained by the proposed model on different questionnaire subsets. Each subset contains all questionnaires where subjects faked a number of responses indicated on the x-axis.

In this context, we observe that the accuracy remains relatively stable in the $0.4 - 0.6$ range for all three conditions, indicating that faked items may be identified at the same level of accuracy independently from the specific faking strategy used by the subject.

In Table 7, we report the average performance achieved by the proposed approach in the three datasets. The results are compared with a simpler Distribution-based Model (DM) where fakers are identified following the same strategy designed for the proposed approach but relying on raw responses instead

of TF-IDF ones. Results indicate that there is a substantial increase in all the parameters when TF IDF is used instead of raw scores. In the same Table 7, we also report the Precision, Recall and F1 Score metrics. These are defined as follows:

$$Precision = \frac{TP}{TP + FP}, \tag{2}$$

$$Recall = \frac{TP}{TP + FN}, \tag{3}$$

$$F1Score = \frac{2 * Precision * Recall}{Precision + Recall}, \tag{4}$$

where TP, FP, and FN indicate respectively the True Positives – i.e. the number of faked responses correctly classified – False Positives – i.e. the number of responses wrongly classified as faked – and False Negatives – i.e. the number of faked responses which were not spotted by our model – in our classification task [12].

If we look in more detail at a few examples, we can improve our understanding of how the proposed approach works. In Fig 4, 5, and 6, we observe the TF-IDF scores associated to the responses of three of our subjects (one for each dataset). In particular, we report the sequence of responses of three faking subjects (blue line) to each of the 10 items in our questionnaire. We compare each response to the distribution of honest responses given by honest participants to the same questionnaire item. The distribution of honest responses to each question is indicated in the figures with a box plot indicating the mean of each distribution (orange line), the first quartiles (extremes of the boxes), and the 75th and 25th percentile value (upper and lower whisker). Finally, we highlight in red the responses where the participant altered his/her response – compared to the respective honest questionnaire he/she provided – when asked to fake his/her responses. For example, in the situation depicted in Fig 4, only Q6 had the same response in the honest and faked condition. In this specific case 7/9 of the faked responses are accurately detected by the algorithm when selecting the 75th percentile threshold for each question – i.e. when flagging as faked the responses that fall outside of the 75th percentile of the distribution of honest responses to the same question. In this case, the total accuracy is 8/10 since the algorithm correctly identified all not-faked responses (Q6) as honest but missed 2 faked ones (Q2 and Q9).

Fig 4. TF-IDF response values for participant n. 5 (faked responses) compared to the distributions of TF-IDF values of honest responders in the CC dataset. The red dot indexes the items to which the participant faked. In this specific case only Q6 had the same response in the honest and faked condition. In this specific case 7/9 of the faked responses were accurately signalled by the algorithm. The total accuracy is 8/10. Two faked responses were missed.

Similarly to what just observed, in Fig 5 we observe a different faking behavior in the context of a job interview for a salesperson position. In this case, the subject faked his/her responses in all but two questionnaire items – Q4 and Q6, which were correctly identified as honest – while faked his/her responses in the remaining ones. Out of the faked responses, the proposed approach was able to detect 6 of them, incorrectly flagging 2 of them – i.e. Q2 and Q10 – as honest with a global accuracy of 8/10.

In Fig 6 we observe another faking behavior, in the context of a job interview for a humanitarian organization. Here, the subject faked his/her responses in 6/10 cases and the proposed approach was able to correctly identifying 2 of them, obtaining a global classification accuracy of honest and faked responses of 6/10 – i.e. it correctly identified the 4 honest responses and 2 of the faked ones correctly while it incorrectly classified 4 of the 10 responses in the questionnaire.

Within this evaluation framework, we consider as faked all responses that are different from the respective honest ones provided by the same participant when

Fig 5. TF-IDF response values for participant n. 48 (faked responses) compared to the distributions of TF-IDF values of honest responders in the JIS dataset. The red dots index the items to which the participant faked. In this specific case Q4 and 6 had the same response in the honest and faked condition, i.e. the subject did not fake. Of the 8 faked responses, the TF-IDF algorithm spotted accurately 6/8 faked responses. The total accuracy is 8/10.

Fig 6. Here we report a case where the precision is below the reported average. TF-IDF response values for participant n. 4 (faked responses) compared to the distributions of TF-IDF values of honest responders in the JIHO dataset. The red dots index the items to which the participant faked. In this specific case Q1, 2, 5 and 6 had the same response in the honest and faked condition, i.e. the subject did not fake. Of the 6 faked responses 2/6 were correctly identified. Q4, Q7, Q9 and Q10 were not reported as faked. The total accuracy is 6/10.

taking the test the first time. However, as mentioned earlier, the reliability of a test is far from perfect [5]. Therefore, we believe that with a less strict identification of faked responses, the performance evaluation of our approach would yield higher values.

Table 7. Performance evaluation

	Group	Precision	Recall	F1 Score	Accuracy
TF-IDF	Child custody, litigation (CC)	0.6700	0.4597	0.5046	0.5566
	Job interview, salesperson (JIS)	0.7178	0.5279	0.5664	0.5642
	Job interview, humanitarian organization (JIS)	0.6808	0.5705	0.5877	0.5848
DM	Child custody, litigation (CC)	0.4735	0.4176	0.4057	0.3516
	Job interview, salesperson (JIS)	0.4692	0.3237	0.3413	0.3123
	Job interview, humanitarian organization (JIS)	0.4966	0.4295	0.4142	0.3422

Notes. Performance evaluation of the proposed approach (TF-IDF) and of a simpler Distribution-based Model (DM) relying on raw response values instead of TF-IDF ones. The use of TF IDF clearly outperforms the use of raw scores in all the evaluation metrics.

Conclusions

Psychological tests usually require participants to endorse a statement describing a personality feature or a behavior using a Likert scale (e.g., a 5 point scale). Such tests may easily be faked by modifying the original honest response in order to achieve an objective (e.g., in the context of a job application).

To handle this problem, faking propensity is usually detected using the so-called *control scales* which are subsidiary scales added to base scales (e.g., desirability scales such as BIRD [13]). However, these procedures help – at best – to spot fakers but do not allow to identify specific faked responses. In short, the identification of specific items that have been faked (e.g., to obtain a job) currently is a major unsolved issue.

The crucial factors responsible for the difficulty of this task are:

- the high variability of responses provided by each individual;
- the different degrees of intensity of faking;
- the influence of the final objective of faking which determines its manifestation: faking to land a position as a salesperson requires faking to

extroversion-related items while for landing a bank position faking on consciousness-related items may be a priority.

Currently, no technique is available which identifies the specific response that underwent faking and control scales identify, at best, only a general propensity of the participant to fake to a test but do not pinpoint where faking actually takes place.

For these reasons, we propose the use of a new technique, inspired by original research in information retrieval. The technique is known as Term Frequency-Inverse Document Frequency (TF-IDF for short) and is at the base of search engine technologies. This technique, when applied to Web search engines, for example, helps identifying the most relevant web page with respect to a query [14].

In traditional item analysis, identical responses of two subjects to the same item are evaluated in the same way. By contrast, according to the TF-IDF modeling strategy, two identical responses (of two different participants) to the same question may yield to different evaluations depending on the responses of other participant to the same items and the subject responses to other questions of the questionnaire. In other words, TF-IDF aggregates in a unique measure both the distribution of the group responses to a target item, as well as the subject style of response (e.g., how frequently s/he selects a certain response value also in other items).

TF-IDF, as presented here, may be considered a "novelty detector", as it compares new items with the distribution of the honest responses. If the TF IDF of a subject's response to an item is abnormally high this indicates that the participants does not belong to the sample of honest responders.

This approach has a great advantage over approaches that aim only to an overall classification of items. As a proof of concept of the effectiveness of TF-IDF in the identification of faked responses, we required about 700 participants, to respond to a short version of the Big Five test (10 items Big Five questionnaire [11]). Each participant was required to respond twice to the 10 items. The first time, honest responses were collected while the second time the participant was required to fake to achieve an advantage in three different scenarios: (i) to obtain a child custody, (ii) to obtain a salesperson position at a generic company, (iii) to obtain a position at a humanitarian organization.

The exploratory analyses, performed on raw scores, indicated that:

- Faking (good) resulted in an increase in average response values on all but one of the 10 items;
- Faking in different contexts, as expected, resulted in distortions of responses to different items of the test.

In evaluating the proposed TF-IDF approach, we observed the following:

- a TF-IDF based representation of item responses leads to a better separation between honest and the faked response values distributions compared to raw scores;
- the average (over all the 10 items of the considered questionnaire) TF-IDF response values of honest responders was lower than that of deceptive ones;
- lower TF-IDF values characterize honest responses to single items, while larger values are characteristic of faked ones;
- the proportion of faked responses associated to large TF-IDF representations (e.g., above the 80 percentile of the honest values distributions) was consistently larger than the honest counterparts;
- when employing the proposed approach based on TF-IDF values distributions for item-level faking detection, the achieved precision (the percentage of faked items correctly identified over all the faked items) was between 67% and 71%, indicating that TF-IDF correctly identifies most of the faked single

items responses. The same faking detection procedure, based on raw response values yielded, instead, to much lower precision values (see Table 7).

All the results indicated that TF-IDF could spot deceptive responses at single item-level with an unprecedented accuracy. Figs 4, 5, and 6 show, in practice, how single participants' responses can be analyzed to spot items which show abnormally high TF-IDF scores associated to them, and how these highly correlate with faking.

The identification of fakers and faking at single-item level is achieved without the use of any control scale, just by capitalising on one distinguishing feature of TF-IDF. It is worth noting that the use of TF-IDF as a novelty detector allows to spot a faker using only a comparable group of honest responders who previously answered the same questionnaire, independently from the context. This feature has interesting practical implications given that faking is context-specific and depends primarily on the strategy that the faker has implemented which, in turn, depends on the objective that faking is expected to achieve (e.g., the custody of a child or the attainment of a job as a salesperson).

While a full discussion of item analysis conducted using TF-IDF and that of Classical Test Theory (CTT) and of Item Response Theory (IRT) goes beyond the scope of this paper, it is worth noting that TF-IDF includes in the evaluation information about the response style of the test takers which, by contrast, is lacking both in CTT and IRT.

References

1. Sartori, G. and Zangrossi, A. and Orrù, G. and Monaro, M. Detection of malingering in psychic damage ascertainment. P5 medicine and justice. Springer, 2017 :330–341.
2. Butcher, J. N., Graham, J. R., Ben-Porath, Y. S., Tellegen, A., Dahlstrom, W. G. Multiphasic Personality Inventory–2 (MMPI-2): Manual for administration and scoring. Minneapolis, MN: University of Minnesota Press. 2001
3. Millon, T. Millon clinical multiaxial inventory: I & II. Journal of Counseling & Development. Wiley Online Library 1993, 70(3):421–426.
4. R. Baeza-Yates and B. Ribeiro-Neto. Modern information retrieval. volume 463. ACM press New York, 1999.
5. Rammstedt, B. and John, O.P. Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German. P5 medicine and justice. Elsevier, Journal of research in Personality. 2007 41(1): 203–212.
6. Zhang, W. and Yoshida, T. and Tang, X. A comparative study of TF-IDF, LSI and multi-words for text classification. Expert Systems with Applications. Elsevier. 2011 38(3): 2758–2765.
7. Sartori, G. and Lombardi, L. Semantic relevance and semantic disorders. Journal of Cognitive Neuroscience. MIT Press. 2004 16(3): 439–452.
8. Mechelli, A. and Sartori, G. and Orlandi, P. and Price, C.J. Semantic relevance explains category effects in medial fusiform gyri. Neuroimage. Elsevier. 2006 30(2): 992–1002.
9. Birkeland, S. and Manson, T. and Kisamore, J. and Brannick, M. and Smith, M. A meta-analytic investigation of job applicant faking on personality measures. International Journal of Selection and Assessment. Wiley Online Library. 2006 14(4):317–335.
10. Merckelbach, H. and Smeets, T. and Jelicic, M. Experimental simulation: Type of malingering scenario makes a difference. The Journal of Forensic Psychiatry & Psychology. Taylor & Francis. 2009 20(3):378–386.
11. Guido, G. and Peluso, A.M. and Capestro, M. and Miglietta, M. An Italian version of the 10-item Big Five Inventory: An application to hedonic and utilitarian shopping values. Personality and Individual Differences. Elsevier. 2015, 76: 135–140.
12. Powers, D. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. arXiv preprint arXiv:2010.16061. 2020.
13. Hart, C.M. and Ritchie, T.D. and Hepper, E.G. and Gebauer, J.E. The balanced inventory of desirable responding short form (BIDR-16). Sage Open. SAGE Publications Sage CA: Los Angeles, CA. 2015, 5(4): 2158244015621113.
14. Spärck, J.K.. IDF term weighting and IR research lessons. Journal of documentation. Emerald Group Publishing Limited. 2004.

Fig 1

Click here to access/download;Figure;Fig 1.tif

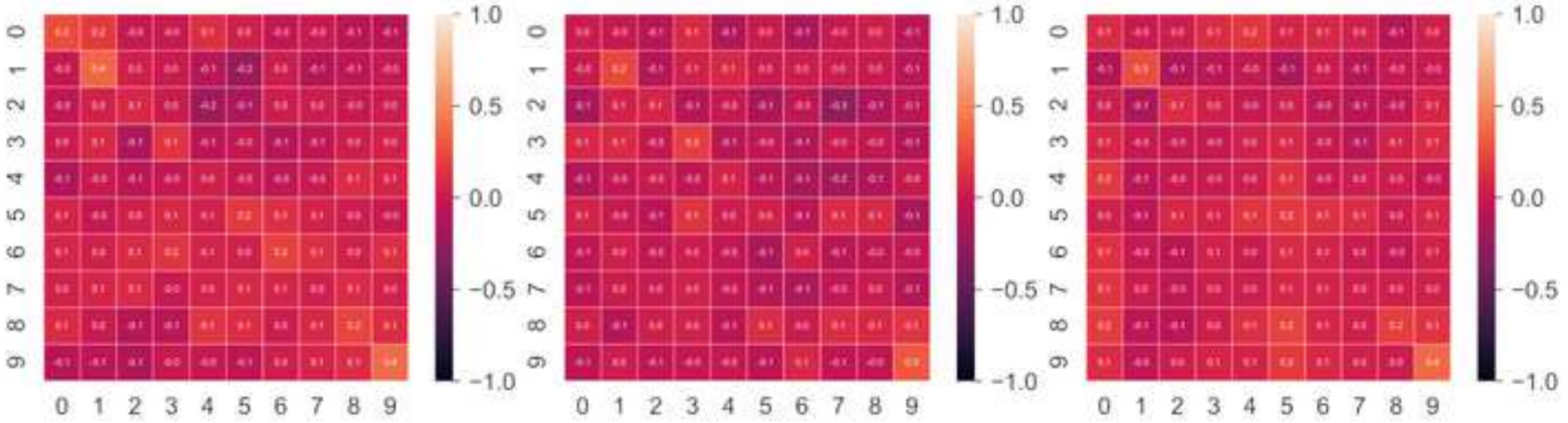

Fig 2

Click here to access/download;Figure;Fig 2.tif

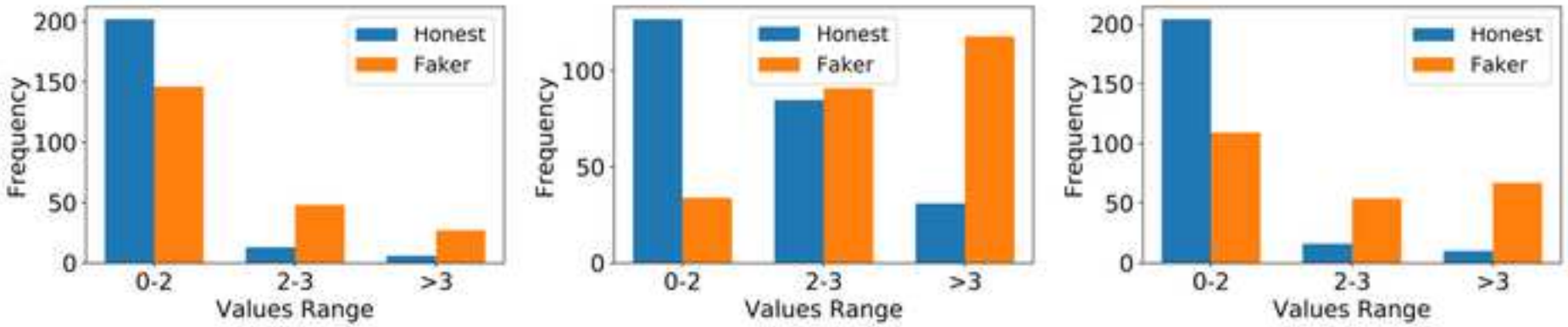

Fig 3

Click here to access/download;Figure;Fig 3.tif

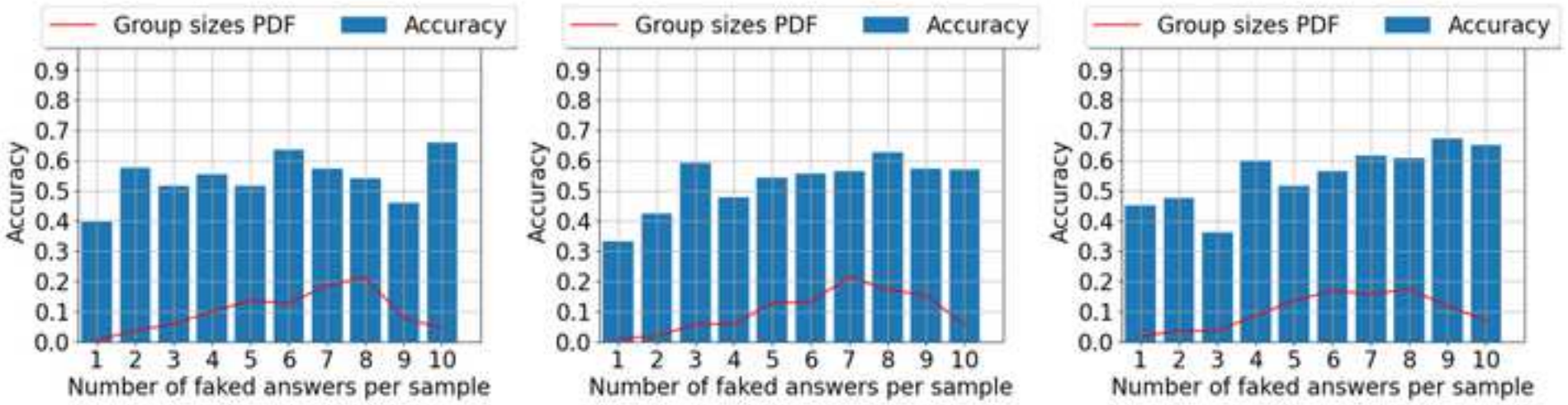

Fig 4

Click here to access/download;Figure;Fig 4.tif

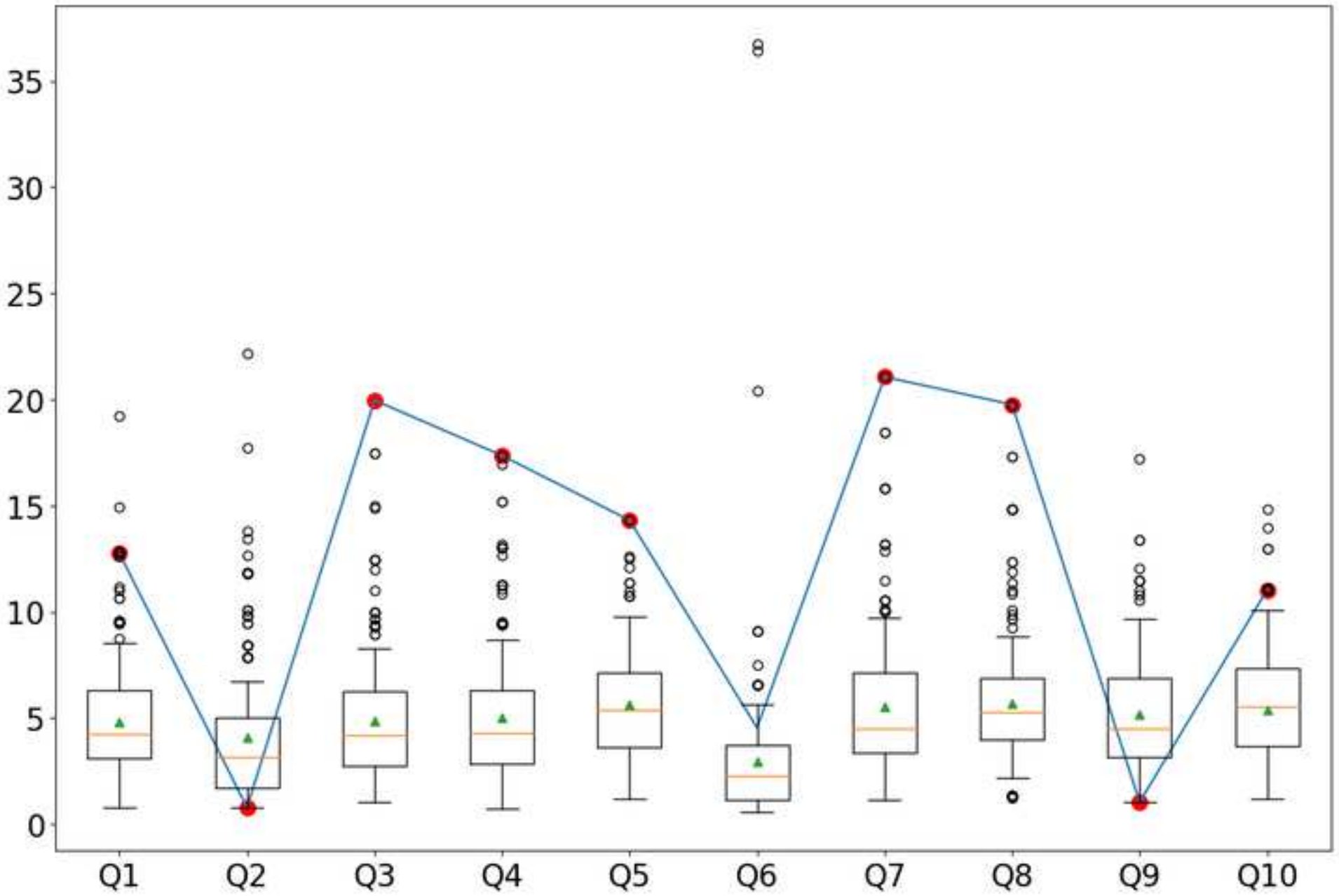

Fig 5

Click here to access/download;Figure;Fig 5.tif

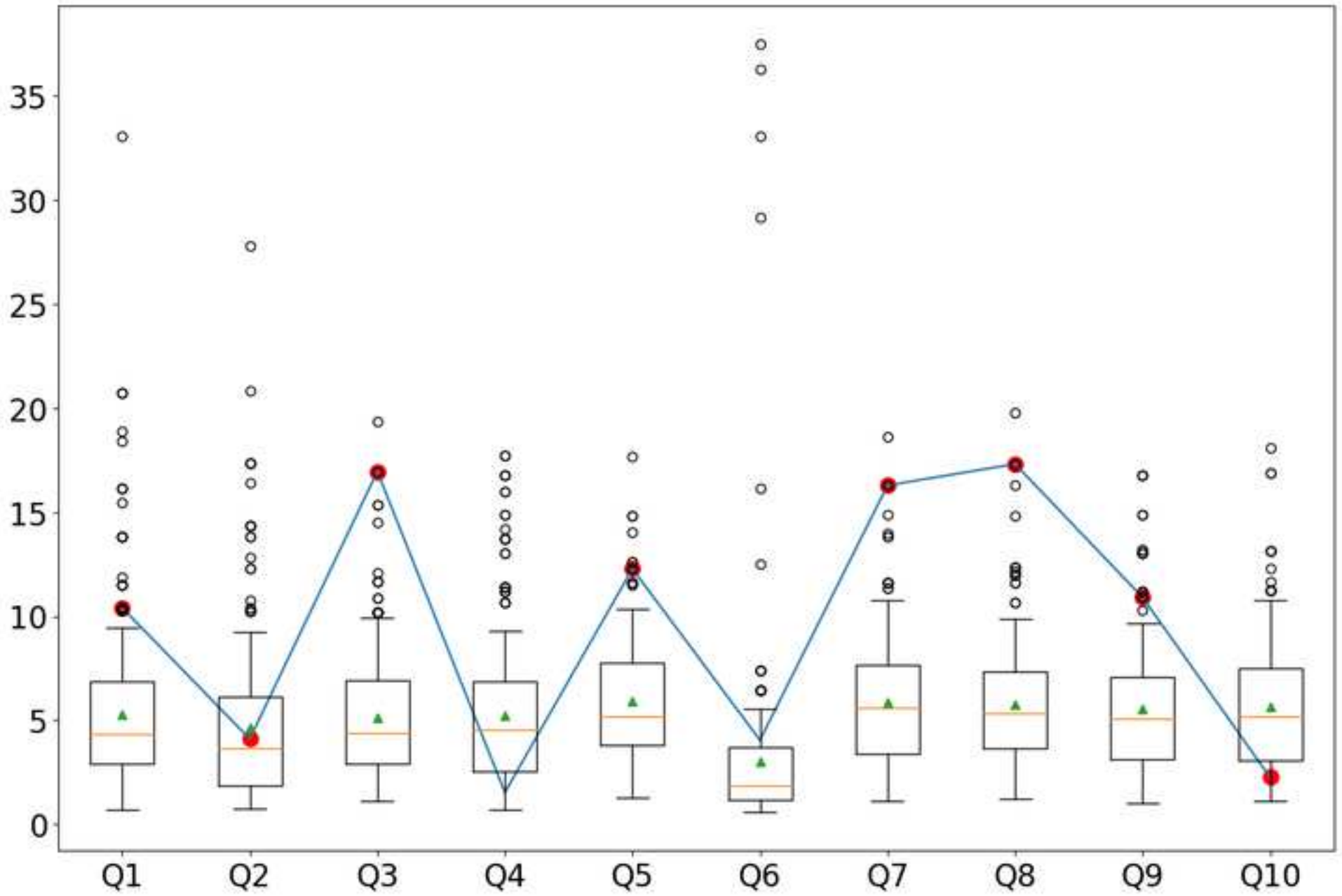

Fig 6

Click here to access/download;Figure;Fig 6.tif