NEW COMBINATORIAL & BAYESIAN UNCERTAINTY ESTIMATION OF TESTS & SURVEYS

Uncertainty of Tests/Surveys

Satyendra Nath Chakrabartty[1], Kangrui Wang[2], and Dalia Chakrabarty[3]
[1]indian ports association, india; alan turing institute, uk; department of mathematical sciences, loughborough university, uk

Corresponding author information: Dr. Dalia Chakrabarty; Department of Mathematical Sciences,Loughborough University, Loughborough LE11 3TU, U.K., (d.chakrabarty@lboro.ac.uk); http://www.lboro.ac.uk/departments/maths/staff/academic/dalia-chakrabarty/

Data Availability Statement: the data that support the findings of this study are openly available in (1) "Raw data from online personality tests" at https://openpsychometrics.org/_rawdata/, reference name "HSQ"; (2) "Yelp dataset challenge" at http://mondego.ics.uci.edu/projects/yelp/, under link named "Train.arff"; (3) and the 3rd data is uploaded with the manuscript as the file "DATAI.xls".

Abstract

We review reliability computation undertaken variously in the literature, to identify concerns with existing approaches; such concerns distort decisions about usability of a given test/survey, and are triggered with increased number of test items, and with increasing non-uniformity of the inter-item correlation structure. Another underlying assumption in these approaches is that the test/survey measures a single ability/trait, but a real test is likely to be multi-dimensional. To address such concerns, we advance new frequentist and Bayesian methods, that are not restricted by unrealistic assumptions about uni-dimensionality, homogeneity of inter-item correlations, or small to moderate number of test items. In our methods, uncertainty of test scores is parametrised by variance of the difference between the score obtained in an item in one subtest, and the corresponding item in the other subtest, where the two subtests arise from the splitting of the given test. Thus, our methods offer new ways of accomplishing such splitting. We illustrate our methods on three real datasets (with responses that are binary as well as on a Likert scale), and four different simulated datasets. We undertake thorough comparison of our results to those obtained by other techniques.

Key words: Mathematical psychology: 91Exx; Measurement and performance: 91E45; Partitions of sets: 05A18; Markov chains (discrete-time Markov processes on discrete state spaces): 60J10

New Combinatorial & Bayesian Uncertainty Estimation of Tests & Surveys

1. Introduction

Estimation of uncertainty of a test score data is extensively studied within Classical Test Theory, with the complementary test reliability defined as the proportion of observed score variance that is attributable to the true score. This theoretical definition then naturally poses the fundamentally difficult problem that the true score is itself unknown (Rudner & Schafes, 2002; Webb, Shavelson & Haertel, 2006). To circumvent this problem, various methods of obtaining the error variance have been advanced. In principle, the sought uncertainty in a set of test scores can be treated as a distance between two datasets, where the datasets could be the outputs achieved on administering the same test/survey to the given cohort, at two different time points, although such is expected to result in some learning during the inter-administration time, potentially driving the reliability to depend on the time gap as well as the homogeneity amongst the examinees (Gualtieri, Thomas & Lynda, 2006). Another possibility is to administer similar tests/surveys to a given cohort, though it is difficult to design such similar tests that maintain (quantified) sameness of quality. Then a better alternative is to avoid multiple administrations, and split the only test/survey administered to the cohort (Meadows & Billington, 2005), into two "subtests" comprising equal number of items, such that (s.t.) distance between score matrices obtained by the cohort in each of the subtests, is sought. Then the reliability of the whole test depends on how the test is dichotomised (Lord & Novick, 1968). Guttman (1945) suggests experimentally identifying the splitting of items, s.t. the split-half reliability is maximised, though a specific algorithmic protocol for finding this optimal splitting is not provided, and this reliability is shown by Ten Berge & Socan (2004) to be an overestimate when the number of test items is large, or the examinee sample size is small. Thompson, Green & Yan (2010) have shown that the maximal split-half coefficient obtained from the splitting method of Callendar & Osburn (1977), to be anything but robust – "badly" overestimating the reliability under some conditions and underestimating it, given other conditions. Increased non-uniformity in the distribution of true scores across items in the test/survey, implies increased inefficiency of an *ad hoc* splitting of

this test/survey; thus, splitting by including odd items in one subtest and even items in another (Murphy & Davidshofer, 1994; Gulliksen, 1987), fails if true scores in some items are likely to be higher than in others, owing to (for example), such items being easier than others. Indeed it is a hard problem to determine how to split a single test/survey into 2 subtests, where the aim is to ensure that the quality of each subtest is the same, and distance between the scores obtained in the 2 subtests by the cohort, is not an artefact of the splitting, but reflects only the uncertainty of the test/survey. There exists another school of thought though (Kaplan & Saccuzzo, 2001), in which reliability computation of the entire test is recommended, using the Spearman-Brown formula (Gulliksen, 1987; Suen, 1990; Eisinga, Te Grotenhuis & Pelzer, 2012).

Maintenance of quality between the 2 subtests is formalised by suggesting that the subtests be "parallel", where parallelity demands that all items of the test/survey, measure the same latent examinee ability, and the true score of each item is the same constant. Maintaining the very restrictive condition for parallelity in real-life tests/surveys is difficult, and often breached. The "tau-equivalent" model relaxes this restriction by allowing item-specific errors, though true scores of all items are held equal to each other still. The more relaxed, "essential tau-equivalent" model allows item scores to differ from each other by an item-specific additive constant. The congeneric model is the least restrictive in that it allows a linear relationship between scores s.t. true scores differ from each other by an additive constant, and a scale (Graham, 2006).

In this paper, we advance novel frequentist and Bayesian methods, that can be used to partition realistic, large to small tests/surveys that do not necessarily measure a single trait, nor manifest non-uniform inter-item correlation structures. The partitioning is done into *optimally-split* subtests that: minimise the absolute difference between the mean subtest item scores; or equivalently, maximise the inner product of subtest item score vectors; or comprise items with indices that are Bayesianly learnt, (with likelihood of these unknown indices, defined as a decreasing function of the distance between subtest item score vectors). Subsequent to splitting the test by our methods, we compute reliability as complementary to the test uncertainty that we define as proportional to the variance of the variable that is the difference between item scores in these *optimally-split* subtests.

If assumptions of the aforementioned essentially tau-equivalent model are violated (eg. items

measure the same latent variable in different scales), Cronbach alpha will underestimate the reliability of a given test score data, (Graham, 2006), leading to the test/survey instrument being criticised (and perhaps discarded) for not producing reliable results (Tavakol, 2011). An even greater worry regarding the applicability of Cronbach alpha – as well as interpretability of its computed value given a test/survey – is the fact that while increased internal consistency value (i.e. alpha) necessarily implies a higher measure of uni-dimensionality of the test/survey (i.e. the test/survey measures a unique latent variable), multi-dimensional tests do not necessarily imply a lower alpha than a uni-dimensional test (Cortina, 1993; Green, Lissitz & Mulak, 1977). Limitations of Cronbach's alpha have been discussed extensively by Panayides (2013); Eisinga, Te Grotenhuis & Pelzer (2012); Ritter (2010); Sijtsma (2009); Boyle (1991); Streiner (2003). Thus, alpha computed for the whole of heterogeneous test/survey, can distort our understanding of its reliability in a data-dependent way, s.t. a blind correction is not possible. Reliability computed using either of our frequentist splitting methods, is higher than reliability computed using any other splitting, including Cronbach alpha. Our Bayesian learning of the splitting, offers 95% Highest Probability Density credible regions on the computed reliability, and alpha may or may not be included within this credible region.

Our frequentist splitting by minimising difference of subtest item means, (Section 3.2) borrows from solutions advanced for the "knap-sack" problem in the literature by Hayes (2002); Borgs, Chayes & Pittel (2001); Mertens (2006); Garey & Johnson (1997, among others). Mertens (2006) defines the problem as partitioning a list of positive integers into a pair of partitions, while minimising the difference between the sum of entries in the 2 partitions. This method produces the same splitting, as partitioning by maximising the inner product of the two partitioned vectors (Section 3.3), though robustness to outliers of these methods of partitioning are not the same. Details of our Bayesian method of learning the partitions is discussed in Section 3.4. Our splitting techniques are compared to existing number partitioning methods in Section 8 of Supporting Documents.

Comparison of our reliability computation to Cronbach alpha is discussed in Section 4.1, contextualised to an applications made on a real test data comprising 50 items, administered to about 1000 examinees (Section 4), as well as on the real Yelp restaurant review data consisting of

676 items, collated from 8848 responders (Section 6). We demonstrate the efficacy of our Bayesian method, to compute the reliability of a dataset that comprises responses from 1022 responders, to a 32-item questionnaire that is on a 5-point Likert scale (Section 5).

2. Background

Let us consider a test such that (s.t.) the total number of test items is $P \in \mathbb{N}$, and number of examinees is $N \in \mathbb{N}$. Here we first consider multiple-choice tests, s.t. score obtained by the i-th examinee, in the j-th item is $X_i^{(j)} \in \{0, 1\}$. Item-score of the j-th item is $\tau_j = \sum_{i=1}^{n} X_i^{(j)}$. Here $i = 1, \ldots, n$, $j = 1, \ldots, p$. Let the item score vector of a given test be $\boldsymbol{\tau} = (\tau_1, \tau_2, \ldots, \tau_p)^T$. Let the p items be arranged so that half of these comprise one subtest (that we refer to as the g-th subtest) that the given test is split into, with the remaining $p/2$ items, comprising the h-th subtest. Thus, the methodology exposition that we undertake, is done by considering an even P; generalising applicability of our methods to odd P will be discussed in Section 3.1 and Section 8 of Supporting Documents. Item scores of items that are assigned to the m-th subtest are $\tau_1^{(m)}, \ldots, \tau_{p/2}^{(m)}$; $m = g, h$. Similarly, the score of the i-th examinee across all the items of the m-th subtest is $X_i^{(m)}$; $i = 1, \ldots, n$. The examinee score vector in the m-th subtest is $\boldsymbol{X}_m = (X_1^{(m)}, \ldots, X_n^{(m)})^T$. For the i-th examinee, the error ϵ_i in their score is defined as the difference between scores attained in the g-th and h-th subtests, i.e. $\epsilon_i := X_i^{(g)} - X_i^{(h)}$.

The methodologies that we advance below for attaining *optimal-splitting* of a given test into 2 subtests, effectively seek to minimise the absolute difference between sums of subtest item scores $\left| \sum_{j=1}^{p/2} (\tau_j^{(g)} - \tau_j^{(h)}) \right| = \left| \sum_{i=1}^{n} (X_i^{(g)} - X_i^{(h)}) \right| = \left| \sum_{i=1}^{n} \epsilon_i \right|$, where the first of these equalities stems from Theorem 3 below, and the second from the definition $\epsilon_i := X_i^{(g)} - X_i^{(h)}$. Our classically defined uncertainty is complementary to the reliability r_{tt} as

$$1 - r_{tt} = \frac{S_\epsilon^2}{S_X^2} = \frac{\sum_{i=1}^{n} \epsilon_i^2}{n} - \left(\frac{\sum_{i=1}^{n} \epsilon_i}{n} \right)^2, \tag{1}$$

$$\text{where uncertainty in the } i\text{-th examinee's response is} \quad \epsilon_i := X_i^{(g)} - X_i^{(h)}, \tag{2}$$

$$\text{and the test variance of the observed test scores is} \quad S_X^2 := \frac{\sum_{i=1}^{n} (X_i)^2}{n} - \left(\frac{\sum_{i=1}^{n} X_i}{n} \right)^2, \tag{3}$$

$$\implies r_{tt} = 1 - \frac{\| \boldsymbol{X}_g \|^2 + \| \boldsymbol{X}_h \|^2 - 2 \sum_{i=1}^{N} X_i^{(g)} X_i^{(h)} - \left[\sum_{i=1}^{n} \left(X_i^{(g)} - X_i^{(h)} \right) \right]^2 / n}{n S_X^2}. \tag{4}$$

We will discuss the connection of the data-driven reliability defined above, and the general, model-driven reliability measures, for example as discussed by Cho (2016) in Section 7 of the Supporting Documents. Also, in Section 5, the splitting of the test is extended to responses to a survey that is on a k-point Likert scale, where $k \in \mathbb{N}$.

3. Our methods

3.1. Splitting a test by exchanging items in the same row of the 2 subtests

We seek to split a given test (constituting p items), into the two subtests g and h, s.t. sum of absolute differences between the scores attained in the items of subtests g and h, is minimised, i.e. $\left| \sum_{j=1}^{p/2} (\tau_j^{(g)} - \tau_j^{(h)}) \right|$ is minimised, where the same number of items ($p/2$) constitute each subtest.

If we face a test with an odd number of items, we ignore the last item for the purposes of test dichotomisation. This is not to say that our splitting algorithm cannot deal with partitioning of an odd-number of elements into the two subtests; while our partitioning algorithm can deal with such a situation, it is our application-specific requirement of maintaining a same number of items in each subtest that drives us to work with even p values only.

Theorem 1 equates minimisation of the absolute difference between sums of item scores in the g-th and h-th subtests, with minimisation of absolute difference between sums of examinee scores attained in these 2 subtests. Theorem 2 discusses implication of this minimisation on the absolute difference between the sum of squares of examinee scores attained in these two subtests.

Theorem 1. Minimising the absolute sum $\mathcal{S}$ of differences between item scores attained in the $p/2$ items of the pair of subtests that are generated by splitting the given test into subtests g and h, implies minimising the absolute difference between means of scores attained by n examinees in

the g-th and h-th subtests. In other words,

$$\textit{minimising} \quad \sum_{j=1}^{p/2} |\tau_j^{(g)} - \tau_j^{(h)}| \Longrightarrow \quad \textit{minimising} \quad \left| \frac{\sum_{i=1}^{n} X_i^{(g)}}{n} - \frac{\sum_{i=1}^{n} X_i^{(h)}}{n} \right|.$$

The proof of this theorem is provided in Section 1 of the Supportig Documents.

Theorem 2. In a test with binary responses, absolute difference between sums of squares of examinee scores in the g-th and h-th subtests is of the order of $\epsilon^2 \mp 2T\epsilon \mp (p/2)^2 T_P \epsilon^{'}$, if absolute difference between sums of examinee scores is ϵ; difference between the sum of probabilities of correct examinee response to items in one subtest and another is $\epsilon^{'}$, where T_P is the sum of probabilities of correct examinee response to items in one subtest, and T is the sum of scores in one of the subtests, s.t. the total score in the other subtest if $T \pm \epsilon$.

The proof of this theorem is provided in Section 2 of the attached Supporting Documents.

3.2. Splitting using minimisation of absolute difference between sums of subtest item scores

Partitioning a set of positive integers into two groups, s.t. difference between sums of elements in the two groups is minimised, has been addressed before; (Section 8 of Supplementary Materials). Putting this into the context of our problem, one partition is the subtest g and the other h, which contains an equal number of elements as in g. Our method of splitting is akin to the differencing method (or the KK-heuristics method) presented by Karmakar & Karp (1982).

In Algorithm 1 of the Supporting Documents, we present our algorithm for identifying the 2 constituent subtests of a given test, by minimising the sum $\mathcal{S}$ of absolute difference between the scores obtained in these 2 subtests, i.e. by minimising $\mathcal{S} := \sum_{j=1}^{p/2} |\tau_j^{(g)} - \tau_j^{(h)}|$. We implement such splitting, by using an accept-reject idea based on differencing between the item-wise scores in the two subtests, over the N_{iter} iterations that we undertake, where the ℓ-th iteration comprises a total of $p/2$ "swaps". Here, a "swap" constitutes the exchange of the j-th item in the current g-th subtest, with the j-th item of the current h-th subtest; $j = 1, \ldots, p/2$; $\ell = 1, 2, \ldots, N_{iter}$. Value of $\mathcal{S}$ at the j-th swap during the ℓ-th iteration is $s_{(\ell-1)p/2+j}$.

Definition 1. In the 0-th iteration, the item-wise scores are sorted in an ascending order, resulting in the ordered sequence $\{\tau_1, \tau_2, \ldots, \tau_p\}$. Following this, the item with the highest total score is identified and allocated to the g-th subtest. The item with second highest total score is then allocated to the h-th test, while the item with the third highest score is assigned to h-th test and the fourth highest to the g-th test, and so on. Thus, initial allocation of items is as follows.

subtest g	subtest h	difference in subtest scores
τ_1	τ_2	$\tau_1 - \tau_2 \geq 0$
τ_4	τ_3	$\tau_4 - \tau_3 \leq 0$
$\vdots$	$\vdots$	$\vdots$

Subtests obtained after this very first dichotomisation of the sequence $\{\tau_j\}_{j=1}^{p}$, following this suggested pattern, are called the "seed subtests".

Definition 2. Once all N_{iter} iterations are undertaken, we identify values of $(\ell-1)p/2+j$ that minimise $\mathcal{S}$, using: $(\tilde{\ell}-1)p/2+\tilde{j} := \underset{(\ell-1)p/2+j}{\arg} \left[\min\left(s_{(\ell-1)p/2+j}\right)\right]$, and define $r_{tt}^{(min_{\mathcal{S}})} := r_{tt}^{(\tilde{\ell}-1)p/2+\tilde{j})}$, as the maximal reliability of the given test obtained by minimising $\mathcal{S}$.

3.3. Splitting a test by swapping items across rows

We have considered splitting of a given test, using other methods as well, namely, splitting of a given test, while maximising the correlation between the item scores of the resulting subtests, i.e. maximising $\mathcal{S}_\rho := \sum_{j=1}^{p/2} \tau_j^{(g)} \tau_j^{(h)}$. It is clear that swapping the j-th item of g-th subtest, with j-th item of h-th subtest will not produce any change in $\mathcal{S}_\rho$, for $j \in \{1, \ldots, p/2\}$, as $\mathcal{S}_\rho$ is symmetric in the j-th item of either subtest, by definition. However, swapping the j-th item of the g-th subtest with the $j^{/}$-th item of the h-th subtest, will induce a change in $\mathcal{S}_\rho$, if $j^{/} \neq j$, $j, j^{/} \in \{1, \ldots, p/2\}$. Thus, the maximisation of $\mathcal{S}_\rho$ is brought about by exchanging differently-indexed items between the 2 subtests. The algorithm for implementing splitting using the maximisation of the item score vector inner product, is given in Algorithm 2 of the Supporting Documents.

Definition 3. As in Definition 2, once the iterations are done, identify the $(\ell-1)p/2+j$ values that maximise $\mathcal{S}_\rho$, using: $(\tilde{\ell}-1)p/2+\tilde{j} := \underset{(\ell-1)p/2+j}{\arg}\left[\max\left(s_{(\ell-1)p/2+j}\right)\right]$, and define $r_{tt}^{(max_{\mathcal{S}_\rho})} := r_{tt}^{(\tilde{\ell}-1)p/2+\tilde{j})}$, as the maximal reliability of the given test obtained by maximising $\mathcal{S}_\rho$.

Theorem 3, holds minimisation of $\mathcal{S}$, equivalent to maximisation of $\mathcal{S}_\rho$.

Theorem 3. Splitting a given test into the g-th and h-th subtests by maximising the absolute of the inner product of the item score vectors $\boldsymbol{\tau}_g$ and $\boldsymbol{\tau}_h$ in these 2 subtests is equivalent to the splitting of the test by minimising the absolute sum of differences between the components of these item score vectors, where item score vector in the m-th subtest is $\boldsymbol{\tau}_m = (\tau_1^{(m)}, \ldots, \tau_{p/2}^{(m)})^T$, with $\tau_j^{(m)} := \sum_{i=1}^{n} X_i^{(m_j)}$; $m \in \{g, h\}$. In other words, maximising $\left|\langle \boldsymbol{\tau}_g, \boldsymbol{\tau}_h \rangle\right| = \left|\sum_{j=1}^{p/2} \tau_j^{(g)} \tau_j^{(h)}\right|$ is equivalent to minimising $\left|\sum_{j=1}^{n}\left(\tau_j^{(g)} - \tau_j^{(h)}\right)\right|$.

Proof of this theorem (using Cauchy Schwartz), is in Section 3 of the Supporting Documents.

It is to be noted that the $\mathcal{S}$-minimisation strategy, causes the same subtest-pair to be generated after every $(p/2+2)(p/2+1)/2$ swaps. This periodicity stems from the fact that the total number of possible splittings of a test with p items is $(p/2+1)+p/2+\ldots+1 = (p/2+2)(p/2+1)/2$. Thus, there is a repetition in the value of $\mathcal{S}$ (and reliabilities), with a maximal period of $(p/2+2)(p/2+1)/2$. We identify this as the maximal period, since it is possible even prior to the undertaking of all the $(p/2+2)(p/2+1)/2$ swaps, that 2 distinct subtest-pairs result in the same value of $\mathcal{S}$. A similar repetition is then noticed in results obtained using splitting by maximising $\mathcal{S}_\rho$.

3.4. Our new Bayesian splitting of a given test to attain minimum $\mathcal{S}$

In our Bayesian approach, we learn the indices $g_1, g_2, \ldots, g_{p/2}$ of items that comprise the g-th subtest that a given test of p items is split into, s.t. the remaining $p/2$ items constitute the h-th subtest. We learn indices $g_1, g_2, \ldots, g_{p/2}$, given the test score data $\{x_i^{(j)}\}_{i=1;j=1}^{n;p}$, using MCMC (Independent Sampler Metropolis Hastings). In any iteration, with indices of the items of the test

delineated in ascending order, the test item with the smallest value that is not identified as member of the g-th subtest, is the first item of the h-th subtest, (designated h_1), and so on, till all the left-over test items have been pulled into the h-th subtest.

We define the likelihood of these index parameters, given the data, as a smoothly declining function of the Euclidean norm of the difference between the item score vector of the current g-th and that of the current h-th subtests, s.t. likelihood of the index parameters given the data is a maximum when this distance is 0, and the likelihood is 0, when this distance approaches infinity. Given these constraints, we define the likelihood as $\mathcal{L} \propto \exp\left(-\frac{(\parallel \boldsymbol{\tau}_g - \boldsymbol{\tau}_h \parallel)^2}{2\sigma^2}\right)$, where $\parallel \cdot \parallel$ denotes the Euclidean norm, and σ^2 is the experimentally fixed variance of this Gaussian likelihood.

Thus, $\mathcal{L} \propto \exp\left[-\frac{\left(\tau_1^{(g)} - \tau_1^{(h)}\right)^2 + \ldots + \left(\tau_{p/2}^{(g)} - \tau_{p/2}^{(h)}\right)^2}{2\sigma^2}\right] = \prod_{j=1}^{p/2} \exp\left[-\frac{\left(\tau_j^{(g)} - \tau_j^{(h)}\right)^2}{2\sigma^2}\right]$. Here, $\tau_j^{(g)} := \sum_{i=1}^{n} x_i^{(g_j)}; \quad \forall j = 1, \ldots, p/2$, where g_j is an unknown parameter that we attempt to learn. $\tau_j^{(h)}$ is similarly defined. We place $Binomial(p, 0.5)$ priors on g_j, $\forall j = 1, \ldots, p/2$, in one set of chains, and $Uniform[1.p]$ priors in another set. The likelihood and the priors are used in Bayes rule to define the joint posterior probability $\pi(g_1, \ldots, g_{p/2} | \{x_i^{(j)}\}_{i=1;j=1}^{n;p})$, of the unknown indices $g_1, \ldots, g_{p/2}$, given the test score data. We generate posterior samples using Metropolis Hastings.

Then using the values of $g_1, \ldots, g_{p/2}$ that are current at the end of the k-th iteration, the $h_1, \ldots, h_{p/2}$ indices are identified. This is equivalent to identifying the g-th and h-th subtests in the k-th iteration. Here $k = 0, 1, \ldots, N_{iter}$. Having identified the items that comprise each of the 2 subtests, the score attained by the i-th examinee in each of the items in either of the current subtests, is identified in the k-th iteration, $\forall i = 1, \ldots, n$. This allows us to compute the reliability $r_{tt}^{(k)}$ in the k-th iteration, using our definition of the reliability, as per Equation 4.

In the k-th iteration, let the current value of the parameter g_j be $g_j^{(k-1)}$, and its value proposed in this iteration is $g_j^{(k\star)}$, where we propose $g_j^{(k\star)} \sim Binomial(p, \psi^{(k\star)})$, where the rate parameter of this Binomial proposal *pmf* is the parameter ψ, the current value of which is $\psi^{(k-1)}$ and the proposed value of $\psi^{(k\star)}$, where $\psi^{(k\star)} \sim Uniform[0.5 - a, 0.5 + a]$, with a fixed to 0.4 and 0.2 in two separate sets of experiments. Thus, the proposal density for all sought index

parameters, i.e. $\forall j \in \{1, \ldots, p/2\}$ is the same in a given iteration. Thus, for a given j, the acceptance ratio includes the ratio of the *Binomial pmf* with rate parameter $\psi^{(k-1)}$, to the *Binomial pmf* with rate parameter $\psi^{(k)}$. To compute (the logarithm of) these *Binomial pmf*s, we use Stirling's approximation. Thus, in the k-th iteration, the acceptance ratio of Metropolis Hastings includes this ratio of the proposal densities at the current values $(g_1^{(k-1)}, \ldots, g_{p/2}^{(k-1)})$, to the proposed values $(g_1^{(k\star)}, \ldots, g_{p/2}^{(k\star)})$, of the index parameters, as well as the posterior $\pi(g_1, \ldots, g_{p/2} | \{x_i^{(j)}\}_{i=1;j=1}^{n;p})$ of the proposed to the current values of the parameters.

As diagnostics, traces of the joint posterior, and of the current reliability are included. Tests are carried to check on results of varying a, σ^2 and the priors.

The algorithm for implementation of the Bayesian learning of indices of one of the subtests, and the resulting test reliability, is provided in Algorithm 3 of the Supporting Documents.

4. Empirical illustration on a real data set

In this section, we present results of applying our frequentist number partitioning methods, as well as the Bayesian method of splitting a real test into a pair of subtests, to then compute values of the reliability parameter. We undertake a direct comparison of our results with the Cronbach alpha reliability that is computed for the given test.

This real test data was obtained by examining 912 examinees in a multiple choice examination that was administered with the aim of achieving selection to a position. This test has 50 items, the response to which could be either correct or incorrect, and maximum time allowed for answering this test was 90 minutes. This test data has a mean score of about 10.99 and a variance of about 19.63. We refer to this dataset as DATA-I. For this real test data DATA-I, the results obtained by splitting the test via minimisation of the absolute difference $\mathcal{S}$ between the sum of components of the item score vectors in the resulting subtests, are shown in Figure 1. The results of splitting by maximisation of the inner product $\mathcal{S}_\rho$ of the item score are depicted in Figure 2. Again, results of splitting this real dataset using the Bayesian learning of the indices of the items of the g-th subtest, are depicted in Figure 3.

4.1. Comparing our results to Cronbach alpha

As discussed in Section 1, an underlying assumption for Cronbach alpha is uni-dimensionality of the test, i.e. the test measures one single latent ability/trait variable. We undertake a Principle Component Analysis (PCA) of the test dataset DATA-I to probe the relevance of a Cronbach alpha computation for the internal consistency of the real test data DATA-I. The results of this PCA are presented in Figure 4. These results demonstrate that for the dataset DATA-I, multiple components are relevant; in fact, the score of each of the 4th and 6th components, is in excess of half of that of the 3rd component, with other components also relevant (2nd, 5th, 7th, 8th). This indicates that this real test is not uni-dimensional. Equivalently, the figure indicates that the 20th centile of the variance in this dataset is explained by the first 3 to 4 eigenvalues, ranked by weight. Thus, the PCA of DATA-I helps us appreciate that the assumption of uni-dimensionality that underlies the correct usage of Cronbach alpha, is violated in this real-world example.

In Figure 5, we compare the Cronbach alpha value for test data DATA-I, with reliability obtained by minimising the absolute difference $\mathcal{S}$ between sums of components of the item score vectors of the subtests that result from the splitting of test data DATA-I. We also undertake such a comparison with reliabilities obtained from all other possible splittings of this test data. There are in fact, $(p/2+1)(p/2+2)/2$ number of splittings possible in total for a test with p number of items. For DATA-I then, $26 \times 27/2 = 351$ splittings are possible in total. We undertake each of these distinct 351 splittings of DATA-I into 2 subtests, and for each splitting – indexed by a "splitting index" – we compute values of $\mathcal{S}$; $\mathcal{S}_\rho$; and reliability r_{tt} (using Equation 4). Cronbach's alpha for this real test dataset is compared to such computed reliabilities in Figure 5.

One way of establishing the advantage of a method, is to seek its robustness to outliers. With the aim of identifying the robustness of reliability computed using our methods and Cronbach alpha to outliers in the test data, we undertook computation of – at each deletion of the q-th highest scoring pair of items from the test data DATA-I – reliability by minimising $\mathcal{S}$; reliability by maximising $\mathcal{S}_\rho$; reliability learnt Bayesianly; and Cronbach alpha. Thus, this exercise comprises $p/2 = 25$ steps for our real data DATA-I, s.t. in the q-th step, i.e. for "deletion index" q, the q-th highest scoring item pair is omitted from the data; $q = 1, \ldots, 25$. Thus, there are 48

items in the data DATA-I at any step. The reliability values computed using the 4 different methods, at each item-pair deletion, are plotted against deletion index q, in Figure 6.

5. Generalisation to reliability, with responses on a Likert scale

In this section, we generalise our methods for computing reliability, to a survey, responses to the items of which are on a k-point Likert scale. However, we will continue to refer to this instrument as a "test" and the responders as "examinees". That the Likert scale is not equidistant does not affect our reliability computation (defined in Equation 4), since our parametrisation of uncertainty of a test is the variance of the variable $X^{(g)} - X^{(h)}$. We demonstrate the Bayesian learning of the indices $g_1, \ldots, g_{p/2} \in \{1, \ldots, p/2\}$ of the g-th subtest, using the method discussed in Section 3.4, and the publicly available data that is reported by Martin et. al (2003), where this data comprises responses to an online questionnaire called the "Humour Styles Questionnaire" (or HSQ) that was formulated to collect responses (on a 5-point Likert scale) to questions on responders' attitudes towards humour in different contexts. The exact statements of the questions can be found in the file `codebook.txt` that is a component of the package submitted with the HSQ data, available at `https://openpsychometrics.org/_rawdata/`. The responses are assigned ranks 1,2,3,4,5 following this scheme: 1="Never or very rarely true", 2="Rarely true", 3="Sometimes true", 4="Often true", 5="Very often or always true". In the original dataset with 1037 responders, there was the rank -1 assigned to an item for which a responder did not select an answer. However, for our empirical demonstration, we deleted responses from any responder who left one or multiple items unanswered. This left us with n=1022 responders. There were 32 questions, i.e. 32 items in this dataset. Thus, for this application, $p = 32$, and the responses from the i-th responder is $x_i^{(j)} \in \{1, 2, 3, 4, 5\}$ $\forall j = 1, \ldots, p = 32$ and $\forall i = 1, \ldots, n = 1022$.

We use the generic term "test" to refer to this survey, and "examinees" as responders to this survey. In Figure 7, we depict the results obtained by splitting this real test dataset HSQ, using our Bayesian learning of $g_1, \ldots, g_{p/2}$, leaving the remaining test items to build up the h-th subtest. All parameters of the Metropolis Hastings chain are as used for the Bayesian learning given DATA-I (Section 3.4). As in Figure 3, in Figure 7, we depict traces of the likelihood, and the reliability that is computed at each iteration from the splitting of the full test into the g-th

and h-th subtests, done at each iteration. We also display the histograms of the examinee scores in the g-th and h-th subtests that are identified during the last iteration of this MCMC chain.

Ultimately, we compare the results we get for reliability for this test with the Cronbach alpha that can be computed even for tests, responses of which are on a k-point Likert scale. This computed value for the Cronbach alpha (of about 0.88) falls close to the left edge of the 95% Highest Probability Density credible region of about [0.847,0.915] on our Bayesianly learnt reliability; at about 0.88, alpha is less than the Bayesianly learnt modal reliability of about 0.907. Marginal posterior probability density of g_1, g_5, g_9, g_{13}, given the data HSQ are represented as histograms, and displayed in Figure 8.

5.1. Heterogeneous correlation of real test data DATA-I and HSQ

In this section we present Figure 9 that displays surface plots of inter-item variance-covariance values of the test data DATA-I (left panel of the figure) and HSQ (right panel), for the j-$j^{/}$-th item pair, where $j^{/} \leq j$, $j = 1, 2, \ldots, p$. $p = 32$ for HSQ and $p = 50$ for DATA-I. Thus, the figure displays the lower triangles of the variance-covariance matrices of these datasets. The two real datasets DATA-I and HSQ are differently heterogeneous in their inter-item covariance values.

One way that we choose to parametrise the non-uniformity of the sample covariance of two item scores, is to compute the sum C of frequencies of those inter-item covariance values that occur in the sample, with $\leq$0.05 times the frequency of the modal inter-item covariance in the test data – normalised by the sum of frequencies of all sample covariance values. Then the ratio C gives the normalised sum of covariances of the outlying items in the given test. The extent of heterogeneity in inter-item correlation structure of the HSQ data is manifest in outlier covariance values that contribute to about 20% of the weighted average of the inter-item covariance of the full test. The sample inter-item covariance in DATA-I corresponds to $C \approx$2.3 times lower in HSQ.

6. Reliability of a very large binary test data using minimisation of $\mathcal{S}$ and comparison to Cronbach alpha

With the aim of demonstrating our splitting method on a very large test dataset (or a survey) that comprises binary responses, we looked for such large real life test data in the literature. We

found this in an attempt by Sajnani et. al (2016), that is designed to address the problem of classifying reviews about restaurant businesses written on Yelp, which is a business directory and review service, enabled with social networking capacity. The ulterior aim of building this classifier is that an independent user can then use the categorised information that they are presented with, to make an informed decision about considered restaurants, without wading through wordy textual reviews. This addressed problem is an example of multi-label classification, since the aim in this work is to classify the Yelp restaurant reviews into the categories: "Food", "Service", "Ambience", "Deals/Discounts" and "Worthiness". Textual features of 10,000 Yelp reviews are extracted as 375 unigrams (that occur with frequency in excess of a pre-set threshold); 208 bigrams; 108 trigrams. Star ratings input by the reviewers were also extracted, into 3 binary features for the ratings: "1 to 2" stars; "3 stars"; "4 to 5" stars. In the training data that exists at http://mondego.ics.uci.edu/projects/ yelp/files/train.arff, the extracted features are used to define $p = 676$ binary attributes. Values of each such binary attribute, for $n = 8848$ reviews are included in the training data. We refer to this data that contains information about Yelp restaurant reviews, as DATA-YELP. A pdf of the technical report of the work exists at http://mondego.ics.uci.edu/projects/ yelp/files/technical_report.pdf.

Here we use the reference "test" to this dataset, in the general sense of referring to a test/survey data as "test data", as stated above in Section 2. For this real data DATA-YELP, the mean of the examinee scores is about 162415 and the sample variance of the examinee scores is 717.

We undertook a PCA of the test data DATA-YELP, to check for the correctness of Cronbach alpha for the computation of the internal consistency of such a very large real dataset. The results of this PCA are indicated in the lower panels of Figure 10. The histogram of the eigenvalue weights indicate that the 1st and 2nd eigenvalues are almost of comparable magnitudes, with the 3rd to the 6th eigenvalue not of negligible weights either. So this real test data DATA-YELP is not uni-dimensional. In fact, when we sort the eigenvalues by weights, we find that the first 3 eigenvalues contribute to about 20% of the total variance. The Cronbach alpha for this data is computed to be about 0.91.

From our splitting of the data DATA-YELP, using the minimisation of $\mathcal{S}$, we obtain results

in $r_{tt}^{(min_s)} \approx 0.9258$. The splitting that corresponds to the minimum $\mathcal{S}$, gives rise to the examinee score vectors in the 2 resulting subtests. Histograms of the examinee scores in the 2 subtests are overplotted in the upper left panel of Figure 10. Difference between the examinee score attained in the j-th item of the g-th subtest, and the j-th item of the h-th subtest, are plotted against the item pair index, in the upper right of this figure.

7. Conclusions

We have advanced 3 different methods of splitting a test, (or a survey) into 2 subtests, s.t. variance of the difference between examinee scores attained in the 2 subtests, normalised by the test variance, is defined as uncertainty of the test data; the test reliability is then complementary to this uncertainty. (Here, by "examinees", we include responders of a survey). The 3 methods are essentially equivalent, and operate by splitting a given test into 2 such subtests: by minimising the absolute difference $\mathcal{S}$ between the means of the subtest item score vectors, or; maximising the inner product of subtest item score vectors, or; by Bayesianly learning the positive-definite, integer-valued indices of the items in one of the identified subtests, with the likelihood defined as a smoothly declining function of the Euclidean distance between subtest item score vectors.

We conclude that the advanced splitting methods are not affected by messiness that typifies test data, and the practical limitations of test design, as evidenced by our implementation of the splitting of a very large real test data; of real-world multidimensional tests; and of real tests with non-uniformly correlated items. Tackling such existent problems, is however what limits implementation of existing reliability models, (including that of Cronbach alpha). In fact, we split a real test data in all ways possible, and illustrate our frequentist method of splitting to be such, that the computed reliability is the highest. We also illustrate that the Bayesian learning of the reliability of this test is more robust to outliers amongst the test items, when compared to Cronbach alpha, while splitting by minimisation of $\mathcal{S}$ is comparably robust.

We present these data-driven splitting methods that enable the computation of reliability of large/small, heterogeneous, multi-dimensional real-life test data, that is binary or on a Likert scale, without needing to invoke restrictive model assumptions that cannot be practically adhered to.

References

Borgs, C., Chayes, J. and Pittel, B., (2001). "Phase transition and finite-size scaling for the integer partitioning problem", *Random Structures & Algorithms*, *19*, 3-4, 247–288.

Boyle, G. J. (1991). "Does item homogeneity indicate internal consistency or item redundancy in psychometric scales?", in *Personality and Individual Differences*, 12(3), 3291–294.

Callendar, J.C. and Osburn, H.G. (1977), "A method for maximizing split-half reliability coefficients", in *Educational and Psychological Measurement*, 37, 819–825.

Chakrabartty, S. N, (2011). "Measurement of reliability as per definition", in *Proceedings of the Conference on Psychological Measurement: Strategies for the New Millennium*, School of Social Sciences, Indira Gandhi National Open University, New Delhi, 116–125.

Chakrabartty, S. N, (2013). Challenges of Education in India and Measurement of Overall Progress in *Redefining Education: Expanding Horizons*, ed. M. Sinha, Alfa Publication, New Delhi, ISBN: 978-93-82303-56-8, 98–109.

Cho, E., (2016), "Making Reliability Reliable: A Systematic Approach to Reliability Coefficients", *Organizational Research Methods*, *19*(4), 651-682.

Cortina, J., (1993). "What is coefficient alpha: an examination of theory and applications", *Intl Jl. of Applied Psychology*, *78*, 98–104.

Eisinga, R., Te Grotenhuis, M., Pelzer, B. (2012). "The reliability of a two-item scale: Pearson, Cronbach or Spearman-Brown?", *International Journal of Public Health*.

Garey, M. R. & Johnson, D. S. (1997). *Computers and In-tractability. A Guide to the Theory of NP-Completeness*, W.H. Freeman, New York.

Graham, J., M., (2006). "Congeneric and (Essentially) Tau-Equivalent Estimates of Score Reliability; What They Are and How to Use Them", *Educational and Psychological Measurement*, *66*, 6,930–944.

Green, S., Lissitz, R. and Mulak, S., (1977). "Limitations of coefficient alpha as an index of test unidimensionality", *Educational Psychological Measurement*, *37*, 3-4, 827–38.

Gualtieri, C. Thomas & Johnson, Lynda G. (2006). "Reliability and validity of a computerized neurocognitive test battery, CNS Vital Signs", *Archives of Clinical Neuropsychology*, *21*, 623–643.

Gulliksen, H. (1987). *Theory of Mental Tests*, Routledge, Taylor & Francis Group, New York, Oxford.

Guttman, L. (1945). "A basis for analysing test-retest reliability", in *Psychometrika*, 10, 255-282.

Hayes, Brian (2002). "The easiest hard problem", *American Scientist*, *90*, 2, 113.

Jacobs, Lucy C (1991). "Test Reliability", *IUB Valuation Services and Testing*, www.indiana.edu/best/test_reliability.shtml.

Kaplan, R.M. and Saccuzzo, D.P. (2001). *Psychological Testing: Principle, Applications and Issues (5th Edition)*, Belmont, CA: Wadsworth.

Karmarkar, N, and Karp, R. M. (1982), "An efficient approximation scheme for the one-dimensional bin packing problem", *Proc. FOCS*, pg. 312.

Kline, T. (2005). *Psychologocal Testing: A Practical Approach to Design and Evaluation*, Sage Publications, Thousand Oaks, London, New Delhi.

Lord, F. M, & Novick, M. R. (1968). *Statistical Theories of Mental test Scores*, Addison-Wesley Series in "Behavioral Science: Quantitative methods", Massachusetts, pages 568.

Martin, R. A., Puhlik-Doris, P., Larsen, G., Gray, J. and Weir, K., (2003). "Individual differences in uses of humor and their relation to psychological well-being: Development of the Humor Styles Questionnaire", *Journal of Research in Personality*, *37*, 48–75.

Meadows, M. & Billington, L. (2005). "A Review of The Literature on Marking Reliability", *National Assessment Agency, UK* (https://orderline.education.gov.uk/gempdf/1849625344/QCDA104983_review_of_the_literature_on_marking_reliability.pdf)

Mertens, S. (2006). "The Easiest Hard Problem", in *Computational Complexity & Statistical Physics*, Percus A., Istrate, G. & Moore, C. (eds.), Oxford University Press, Oxford, p. 125.

Murphy, K.R. & Davidshofer, C. O., (1994). *Psychological Testing*, Prentice Hall, Psychology, 548 pages.

Panayides, P. (2013). "Coefficient Alpha Interpret With Caution", in *Europe's Journal of Psychology*, 9(4).

Ritter, N. (2010). "Understanding a widely misunderstood statistic: Cronbach's alpha". *Paper presented at Southwestern Educational Research Association (SERA) Conference*, New Orleans, LA (ED526237).

Rudner, L. M & Schafes, W. (2002). "Reliability" in *ERIC Digest* ericdigests.org/2002-2/reliabillity/htm.

Sajnani, H., Saini, V., Kumar K., Gabrielova E., Choudary, P. and Lopes C., "Yelp Dataset Challenge", *http://mondego.ics.uci.edu/projects/yelp/*.

Satterly, D. (1994). "Quality in external assessment" in *Enhancing Quality in Assessment*, W. Harlen (Ed.), London: Paul Chapman.

Sijtsma, K. (1994). "On the use, the misuse, and the very limited usefulness of Cronbach's Alpha" in *Psychometrika*, 74(1), 107–120.

Stepniak, C. & Wasik, K. (2009). "When do plots of regressions of X on Y and of Y on X Coincide?" in *The Open Statistics and Probability Journal*, *1*, 52–54.

Streiner, D.L. (2003). "Starting at the Beginning : An introduction to co-efficient Alpha and Consistency", *Journal of Personality Assessment*, *80*, 99–103.

Suen, H. K. (1990). *Principles of Test Theories*, Routledge, Taylor & Francis, New York.

Tavakol M., (2011). "Making sense of Cronbach's alpha", *Intl Jl. of Medical Education*, *2*, 53–55.

Ten Berge, J. and Socan, G. (2004). "The greatest lower bound to the reliability of a test and the hypothesis of unidimensionality", in *Psychometrika*, 69, 613–625.

Thompson, B. L., Green, S. B. and Yang, Y. (2010), "Assessment of the Maximal Split-Half Coefficient to Estimate Reliability", in *Educational and Psychological Measurement*, 70(2), 232–251.

Webb, N. M., Shavelson R. J. & Haertel, E. H., (2006). "Reliability Coefficients and Generalizability Theory", *Handbook of Statistics*, *26*, ISSN: 0169-7161.

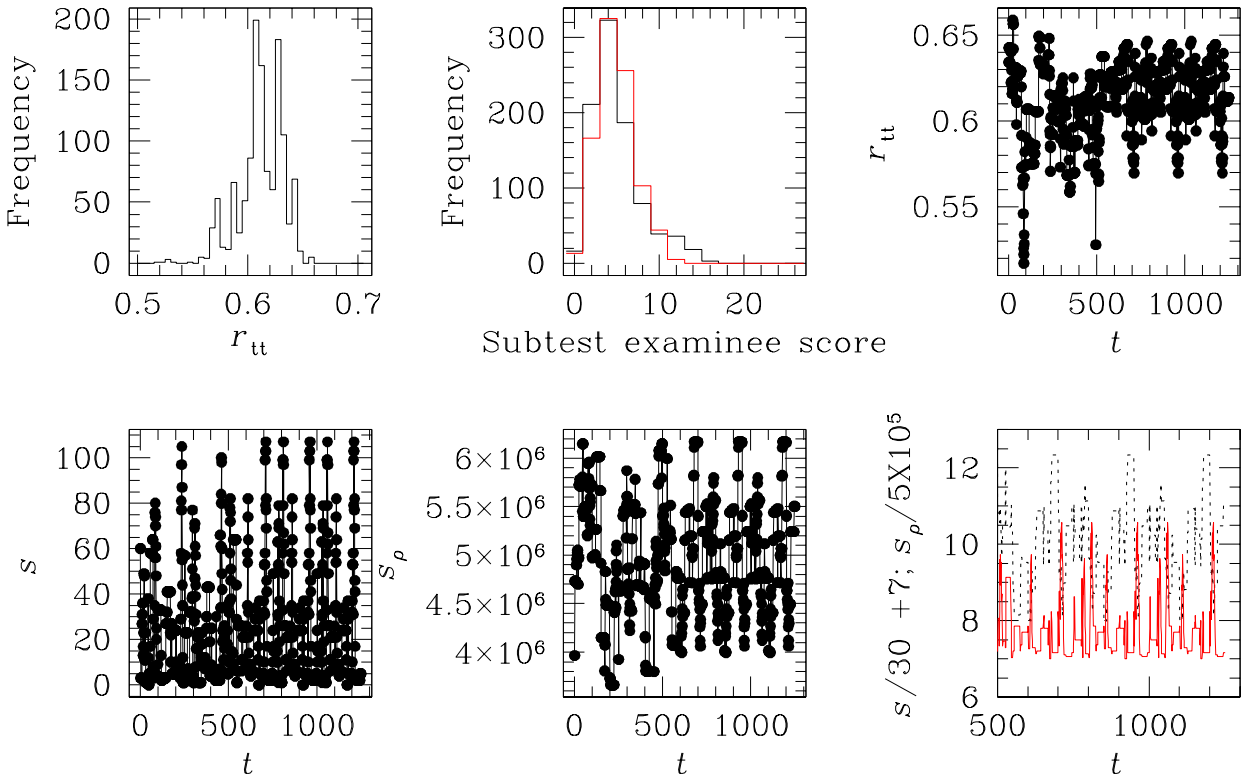

Figure 1: Results of splitting of the real test data DATA-I, into g-th and h-th subtests of equal (=25) number of items, using minimisation of $\mathcal{S}$, i.e. minimisation of the absolute difference between sum of components of item score vectors in the 2 subtests. *Lower left:* plot of value s of $\mathcal{S}$ at the t-th splitting of the test into the g-th and h-th subtests, where the current splitting index $t := 25(\ell - 1) + j$, where ℓ is the current iteration number, and j the current swap number; $\ell = 1, \ldots, 50$, $j = 1, \ldots, 25$. An iteration comprises 25 distinct swaps, where each swap is affected by exchanging the j-th item of the current g-th subtest with the j-th item of the current h-th subtest; a proposed swap may or may not be accepted depending on whether it results in a lower s or not (see Algorithm 1 in Supporting Documents). *Lower middle:* plot against t of value s_ρ of $\mathcal{S}_\rho$ which is the inner product of item score vectors of g-th and h-th subtests. *Lower right:* plot of linearly transformed $\mathcal{S}$ and $\mathcal{S}_\rho$ values, against splitting index t, to empirically verify the equivalence between maximisation of $\mathcal{S}_\rho$ and minimisation of $\mathcal{S}$; such is evident from the peaks of the linearly transformed $\mathcal{S}$ (in thin solid lines) that are noted to occur around the same t values, at which the scaled $\mathcal{S}_\rho$ values (in broken lines) are smallest. Here, the scaling and translation of $\mathcal{S}$ and $\mathcal{S}_\rho$ are undertaken to allow the transformed variables to be plotted within a given interval that allows for their easy visual comparison. Also, to enable such visualisation, we focus on a sub-interval of the values of t relevant to this run (≥ 500). *Upper right:* plot of reliability r_{tt} as computed by our definition (Equation 4), against splitting index t. *Upper middle:* histogram of the r_{tt} values obtained from this run that attains splitting of the given DATA-I test dataset, using minimisation of $\mathcal{S}$. Mean $\bar{r_{tt}}$ of this sample distribution of r_{tt} is about 0.6119 and its sample standard

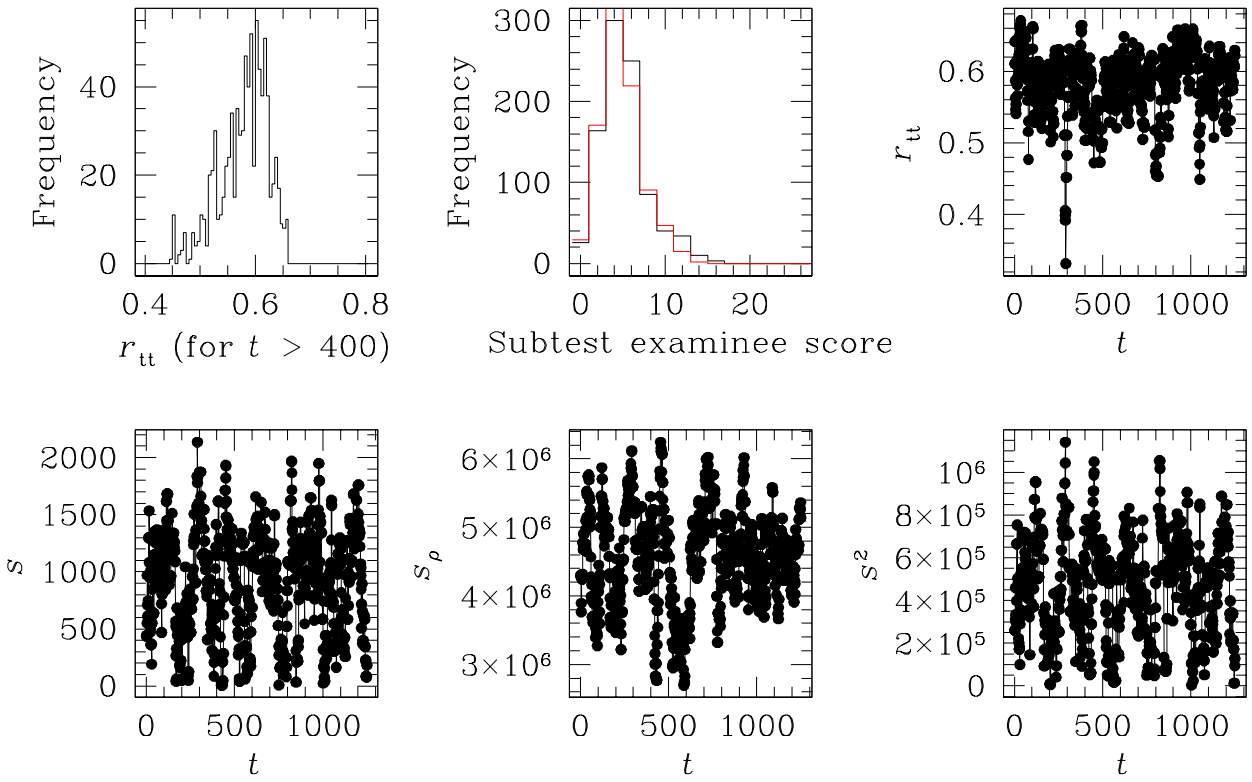

Figure 2: As in Figure 1, but in this run, splitting of real test data DATA-I is undertaken to maximise the value s_ρ of the inner product $\mathcal{S}_\rho$ of item score vectors in the g-th and h-th subtests. The sample mean reliability achieved by this method of splitting is about 0.5829 and the empirical standard deviation is about 0.0394. We identify $r_{tt}^{(max_{\mathcal{S}_\rho})} \approx 0.6596$. Here, the lower right panel displays a plot against the splitting index t, of the value s^2 of the absolute difference between sum of squares of components of the item score vectors in the current g-th and the current h-th subtests. N.B. Due to the permitted swapping of the j-th item of the current g-th subtest by the $j^{/}$-th item of the current h-th subtest, $(j \neq j^{/})$, under splitting by maximisation of $\mathcal{S}_\rho$, sum of components of the 2 subtest item score vectors, can be more different, than when swapping across rows of the 2 subtests is not permitted, as under splitting by minimising $\mathcal{S}$.

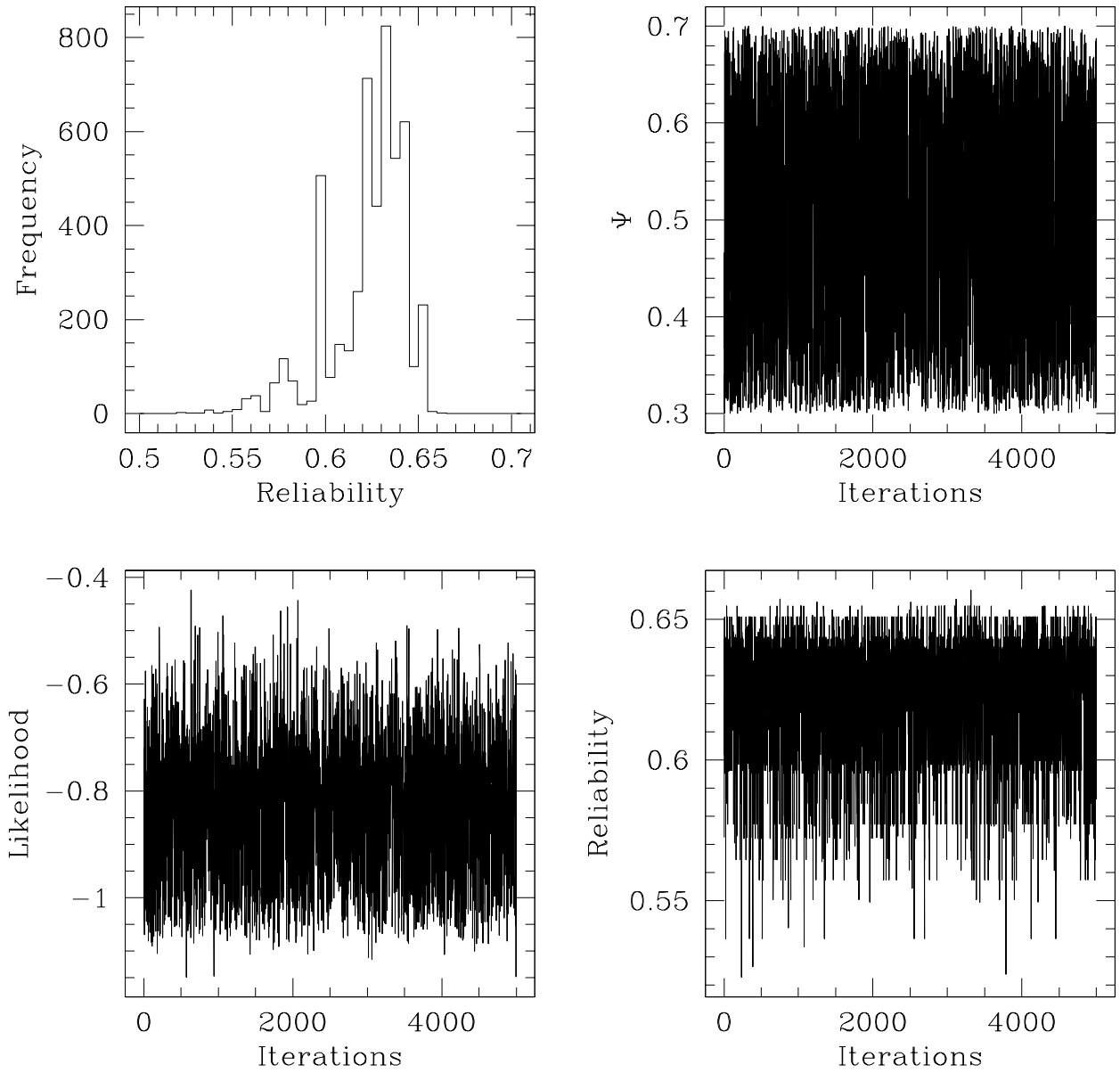

Figure 3: Figure representing results of splitting the real test dataset DATA-I that comprises responses of $n = 912$ examinees in 50 items, using Bayesian learning of the indices of the items in the g-th subtest. The remaining items constitute the h-th subtest. Likelihood is defined as a Gaussian in the Euclidean norm between the item score vectors of the 2 subtests, with a mean of 0 and a variance that is fixed. These results are obtained for $Binomial(50, 0.5)$ priors placed on the sought indices of the items of the g-th subtest. Posterior sampling is performed with Independent Sampler Metropolis Hastings, in which each item index of

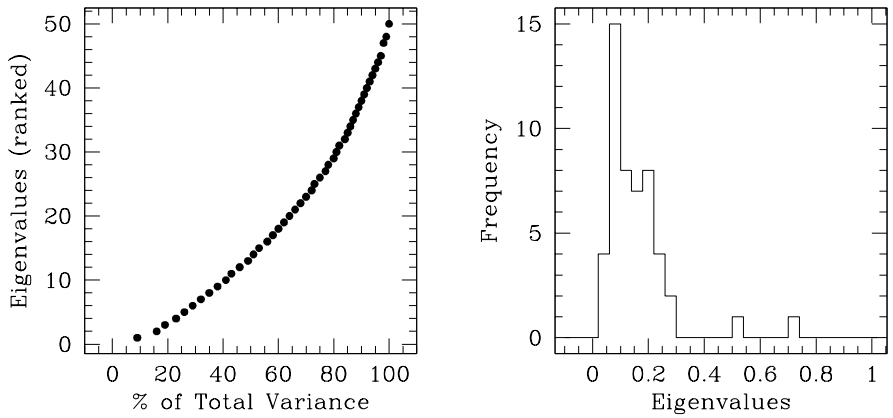

Figure 4: Figure showing results of a PCA done with the real test data DATA-I. The panel on the right displays the histogram of the eigenvalues, while the left panel depicts the eigenvalues (ranked by weights) needed to explain the fraction of the total variance.

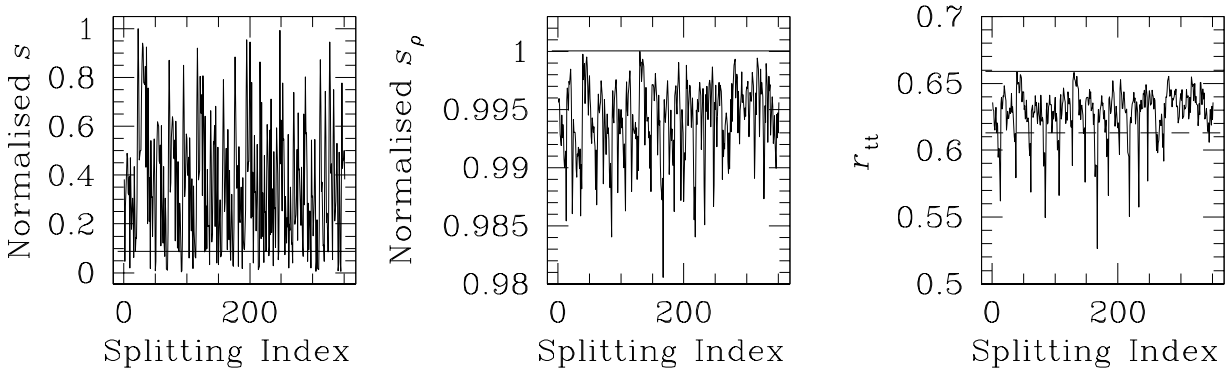

Figure 5: Figure showing results for each of the 351 possible splittings of the read test data DATA-I, where the said results include the absolute difference $\mathcal{S}$ between sums of components of the item score vectors in the 2 subtests that result from the splitting (left panel); inner product $\mathcal{S}_\rho$ of the subtest score vectors (middle panel); reliability r_{tt} computed using the examinee score vectors implied by the current splitting of the test data, in Equation 4 (right panel). These results are plotted against the splitting index, which takes values of $1, 2, \ldots, 351$ for DATA-I. Our results by minimising $\mathcal{S}$ are overplotted on these results, in solid line. Cronbach alpha for DATA-I is also computed and overplotted upon the computed reliability values in the right panel, in broken lines.

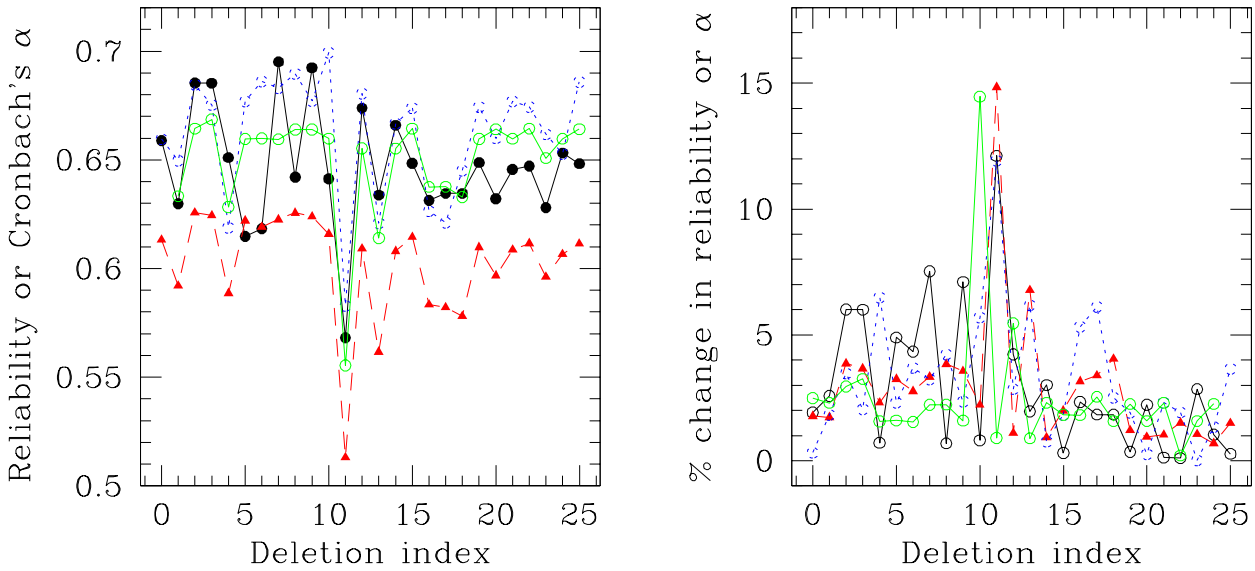

Figure 6: *Left*: figure showing reliability computed using our 3 methods, namely minimisation of $\mathcal{S}$ (in black filled circles, joined by a black solid line); maximisation of $\mathcal{S}_\rho$ (in open circles joined by a broken line – in blue in the electronic version); Bayesianly learning the indices of the items that comprise the g-th subtest (in filled circles joined by a grey solid line – in green in the electronic version), and Cronbach alpha (in filled triangles joined by a broken line – in red in the electronic version). For each case, reliability computed at a given deletion index is plotted against this index, where at the q-th deletion index, the q-th highest scoring pair of items is deleted from the test data, and Cronbach alpha as well as reliability of this data then computed, using our 3 different methods. The fractional change in reliability (over the reliability computed using a given method/definition for the whole test data DATA-I comprising 50 items), is plotted in the right panel, in corresponding line type and symbols (and colour). Variance of this fractional change (expressed as a percentage) is then computed for each of the 4 cases, and the Bayesianly identified reliability is the most robust, with a variance of about 2.45^2, while the reliability computed using splitting by maximising $\mathcal{S}_\rho$ is the least robust (with a variance of about 3.25^2 in the percentage change in reliability with sequential deletion of highest-scoring item pairs). The reliability computed by minimising $\mathcal{S}$ and Cronbach alpha are nearly equally robust, with variances of about 2.96^2 and 2.95^2 respectively.

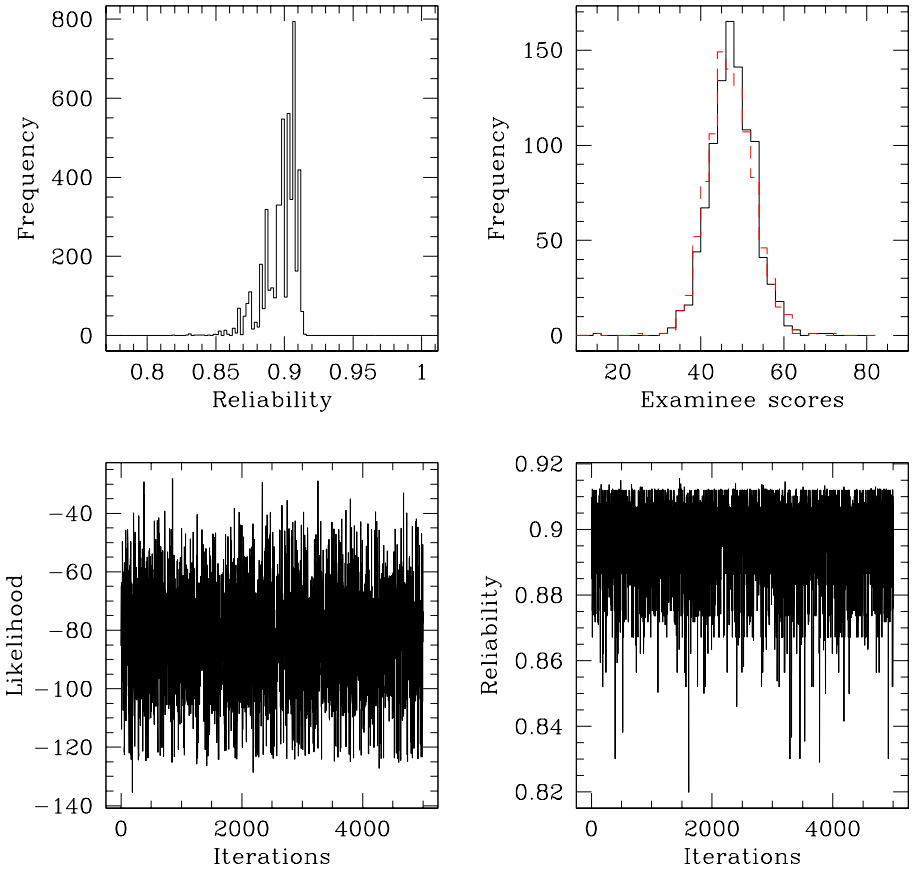

Figure 7: Figure representing results of Bayesian splitting the real survey (that we generically refer to as a "test") dataset HSQ that comprises responses on a 5-point Likert scale. We use responses from $n = 1022$ responders (who we refer to generically as "examinees") who answered every one of the 32 items of this test. Here we Bayesianly learn the indices of the items that comprise one of the subtests that the full test data is split into – we refer to this as the g-th subtest. The remaining items constitute the h-th subtest. Likelihood is defined as a Gaussian in the difference between the L^2 norms of the item score vectors of the 2 subtests, where this Gaussian is assigned a mean of 0 and a variance that is fixed. These results are obtained for $Binomial(32, 0.5)$ priors placed on the sought indices of the items of the g-th subtest. Posterior sampling is performed with Independent Sampler Metropolis Hastings, in which each item index of the g-th subtest is proposed from a $Binomial(32, \psi)$, with $\psi \sim Uniform[0.5 - a, 0.5 + a]$; in this run, $a = 0.2$. At every iteration, reliability is computed using (Equation 4). Traces of this reliability, and of the likelihood are presented in the lower right, and lower left panels respectively. Histogram of learnt reliability is presented in the top left, where the learnt 95% Highest Probability Density credible region is about $[0.847, 0.915]$, with the modal reliability of about 0.907. Cronbach alpha for this test is 0.879. Histograms of examinee scores in the 2 subtests identified in the last iteration of our Bayesian inference, are shown in solid and broken lines on the top right.

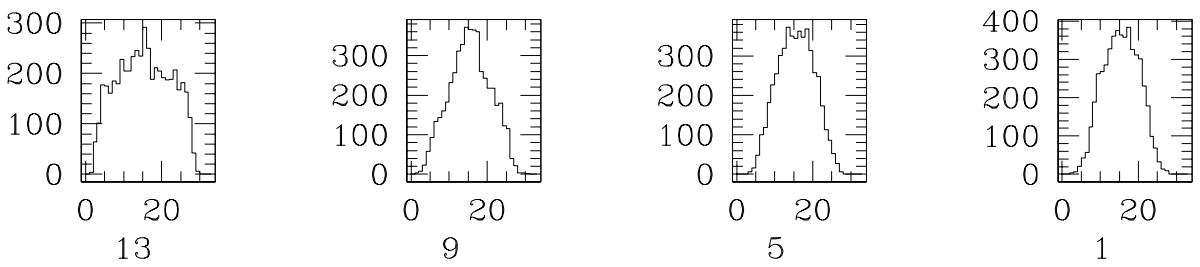

Figure 8: Marginal posterior probability density of the 1st, 5th, 9th and 13th item indices of an identified subtest between the subtest pair that real test data HSQ is split into. The marginals are represented as histograms here.

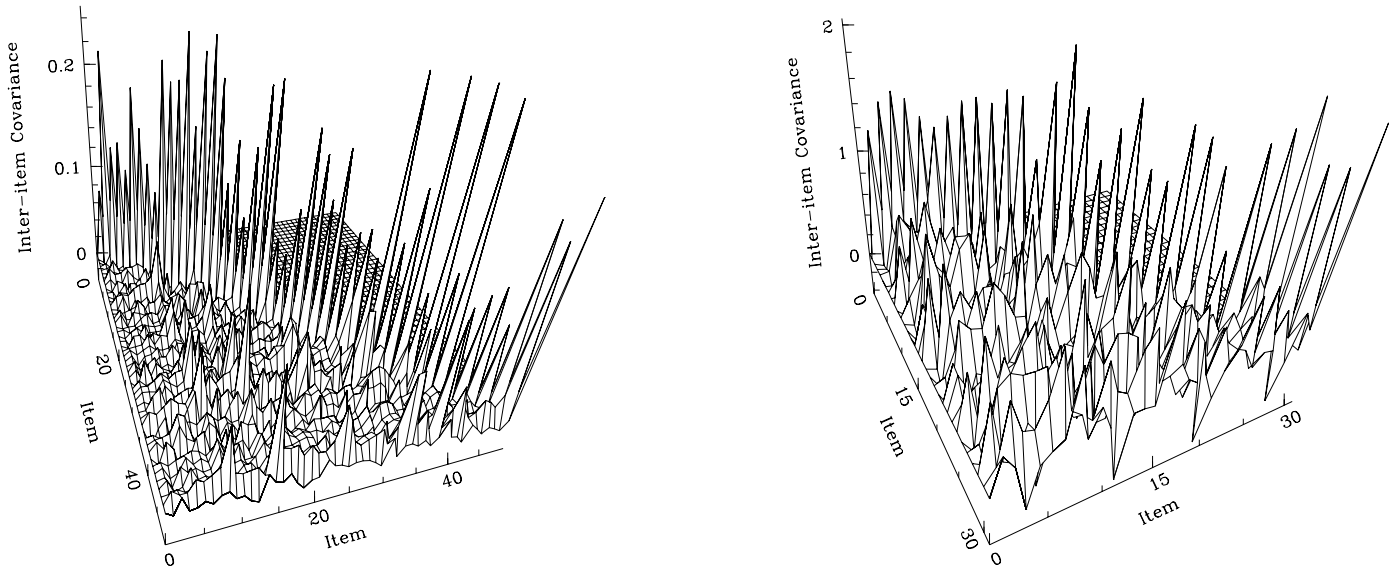

Figure 9: Surface plot of covariance between pairs of items in the given test data (DATA-I on the left, and HSQ in the right panel), plotted against item indices. Here only the lower-triangle of the inter-item variance-covariance matrix is plotted, i.e. covariance between the j-th and $j^{/}$-th item is plotted $\forall j^{/} \leq j; j = 1, 2, \ldots, p$; $p = 50$ for DATA-1 and $p = 32$ for HSQ. Non-uniformity in the covariance values are displayed in the plots. Outlying inter-item covariance values are parametrised by C, which gives the the normalised sum of frequencies of those (outlier) covariance values that occur with frequency ≤ 0.05 times the frequency of the modal covariance in the test data, with the normalisation given by the sum of frequencies of all inter-item covariance values in the given test data. For HSQ, $C \approx 20\%$, while the inter-item covariance sample of DATA-I, causes C to about 8.7%.

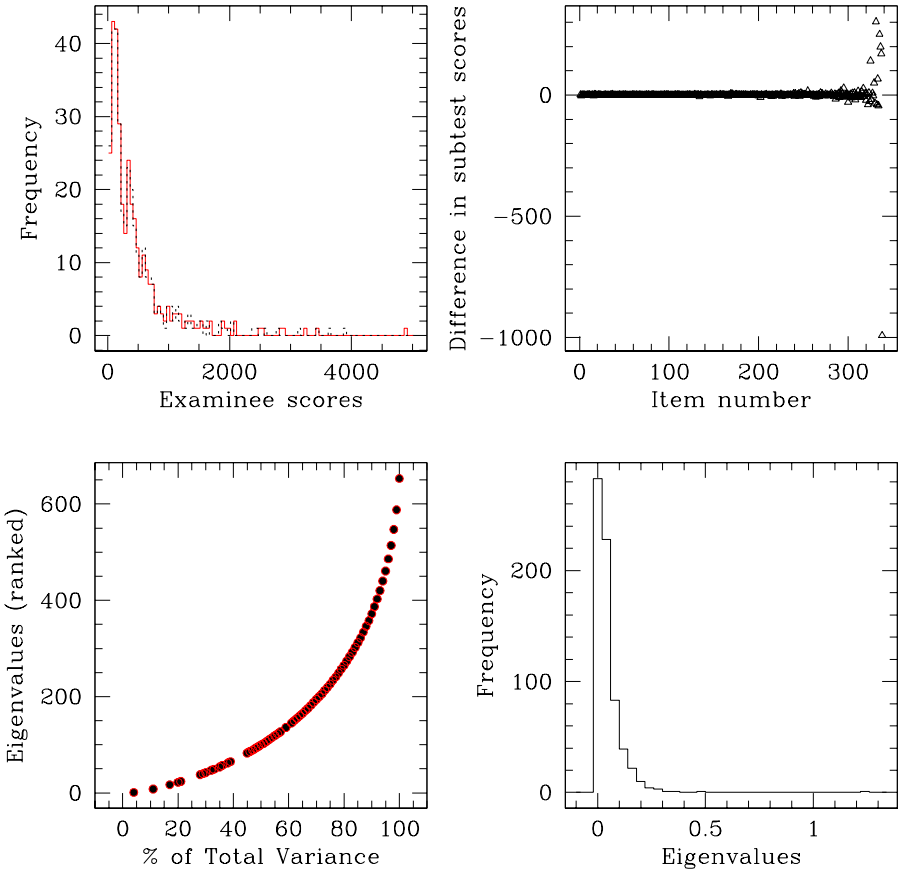

Figure 10: Figure representing results obtained using the very large real dataset DATA-YELP that comprises binary responses on 676 variables (or items), by 8848 responders (or examinees). The eigenvalue weight distribution is shown by the histogram plotted on the lower right. Relevance of at least 6 of the eigenvalues is indicated by this result. Indeed, when eigenvalues, ranked by their weights, are monitored, (lower left panel), it is found that to explain 20% of the total variation, about 3 eigenvalues need to be used. This plot of eigenvalues against fractional variation explained, is drawn by undertaking the PCA for the first half of the dataset, (i.e. for 4424 rows of the data), and then for the full dataset; results from the latter analysis is plotted in black full circles and results for half the dataset is then overplotted in open grey (or red in the electronic version) circles. The upper panels display results of the splitting done by minimising $\mathcal{S}$. In the upper left panel, histograms of the examinee score vectors in the 2 subtests that result from the splitting of DATA-YELP test data, are overplotted in black broken lines and grey (or red in the electronic version) solid lines. The upper right panel then displays the differences between the examinee scores in the j-th items of the g-th and h-th subtests, plotted against j; here $j = 1, \ldots, 676/2 = 338$. Reliability corresponding to the minimisation of $\mathcal{S}$ is about 0.93, while Cronbach alpha for this data is about 0.91.