Int. J. Metrol. Qual. Eng. **15**, 4 (2024)
© S.N. Chakrabartty et al., Published by EDP Sciences, 2024
https://doi.org/10.1051/ijmqe/2023018

International Journal of Metrology and Quality Engineering

Available online at:
www.metrology-journal.org

Research article

Reliable uncertainties of tests and surveys – a data-driven approach

Satyendra Nath Chakrabartty[1], Wang Kangrui[2], and Dalia Chakrabarty[3,*]

[1] Indian Institute of Social Welfare & Business Management, Kolkata, India
[2] University of Warwick, Warwick, UK
[3] Department of Mathematics, Brunel University London, Kingston Lane, London, Uxbridge UB8 3PH, Middlesex, UK

Received: 14 November 2022 / 9 December 2023

Abstract. Policy decisions are often motivated by results attained by a cohort of responders to a survey or a test. However, erroneous identification of the reliability or the complimentary uncertainty of the test/survey instrument, will distort the data that such policy decisions are based upon. Thus, robust learning of the uncertainty of such an instrument is sought. This uncertainty is parametrised by the departure from reproducibility of the data comprising responses to questions of this instrument, given the responders. Such departure is best modelled using the distance between the data on responses to questions that comprise the two similar subtests that the given test/survey can be split into. The paper presents three fast and robust ways for learning the optimal-subtests that a given test/survey instrument can be spilt into, to allow for reliable uncertainty of the given instrument, where the response to a question is either binary, or categorical – taking values at multiple levels – and the test/survey instrument is realistically heterogeneous in the correlation structure of the questions (or items); prone to measuring multiple traits; and built of small to a very large number of items. Our methods work in the presence of such messiness of real tests and surveys that typically violate applicability of conventional methods. We illustrate our new methods, by computing uncertainty of three real tests and surveys that are large to very-large in size, subsequent to learning the optimal subtests.

Keywords: Markov chains (discrete-time Markov processes on discrete state spaces) / mathematical psychology / measurement and performance / partitions of sets

1 Introduction

A test attempts the measurement of ability in a relevant subject, of examinees who are responding to the questions of this test, while a survey measures traits and preferences of the responders. Then it follows that as with all measurements, the measurement of abilities and traits of responders to a test or a survey (respectively), is accompanied by noise or uncertainties of measurement [1]. Such "uncertainty" of a test/survey is interpreted as the shortfall in reproducibility of the score data attained by a given set of responders, to the questions in the instrument that is the test/survey. Thus, one way to quantify this lack of reproducibility of an instrument, would be to administer the instrument a second time, although this practice is expected to result in some learning during the inter-administration time, potentially driving the uncertainty to depend on this time gap, (affected as well by the homogeneity of ability/traits amongst the responders [2]). 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. Such concerns render the usage of a single administration of the instrument, to a cohort, a more robust way for computing uncertainty of the test/survey, [3]. Instead of two "similar" copies of the administered test/survey, the instrument itself is split into two "subtests" – each comprising equal number of questions (or items), such that (s.t.) distance between score matrices obtained by the cohort in each of the subtests, is sought, in rder to parametrise the reliability of the whole test. Indeed, uncertainty in the test data then depends on how the test is dichotomised into these two subtests.

Policy decisions are often achieved subsequent to consulting the data that comprises responses or scores obtained by administering a survey or a test to chosen cohorts of responders. In the real-world, such test/survey instruments are typically designed s.t. they end up measuring not a single, but multiple abilities or traits, i.e. realistically, instruments are multidimensional, and not unidimensional. Characteristics – such as difficulty level, discriminatory power, etc. – of the different questions in a real instrument,

[*] Corresponding author: dalia.chakrabarty@brunel.ac.uk

are also, typically non-uniform across the whole test/survey, s.t. correlation between the vectors of scores obtained by the responders to a pair of questions, is typically distinct from the correlation between the score vectors for any other pair of questions in general. Lastly, real tests and surveys can comprise a very large number of questions – or items – and can be responded to, by a very large cohort.

On the other hand, unidimensionality and uniform inter-item correlaton structure, are requisite assumptions for the applicability of conventional techniques used for the computation of uncertainties. Additionally, success of implementation of such techniques is also often driven by the limited nature of the size of the score data of the instrument. Thus, in a real-world situation, reliability of a test or survey may be spuriously under-estimated or over-estimated. In fact, if the uncertainty of the instrument is wrongly computed, that is in general not discernible, s.t. it is possible that erroneously learnt/estimated reliability, misguides relevant policy decisions.

In this paper, we offer reliable and robust methods for fast estimation of uncertainty of a test/survey, without needing to make unrealistic assumptions about the design of the instrument. In particular, these methods are capable of estimation of such uncertainty, even when the instrument comprises a large number of items, answered by a very large cohort of responders. Thus, in Section 7 we illustrate our automated estimation of the uncertainty in the data comprising scores on 667 items related to restaurant reviews (on YELP), answered by 8848 responders. We show that neither this large data set, nor a moderately large response data to a real test, administered to about 1000 responders for personnel recruitment purposes, (Sect. 5), is unidimensional, i.e. each of these two real instruments, measures multiple abilities. Again, another real survey – responses to items of which are on a 5 point (Likert) scale – is multidimensional (Sect. 6). Heterogeneity of the inter-item correlation structure of this survey is also noted, as is the correlation structure of the test discussed in Section 5. In other words, our methods are robust to real-world departures from unrealistic assumptions about tests/surveys.

As motivated above, the best way to quantify the uncertainty of a test/survey instrument, is via a single administration of the instrument, in which the instrument is split into two subtests – each comprising half the number of items as in the full test – s.t. the 2 subtests are "similar" copies of each other. Then we parametrise the instrument uncertainty as the normalised variance of the random variable that is the distance between the datasets that comprise scores attained by a cohort of responders, to the items of the 2 subtests. In this paper, we advance novel frequentist and Bayesian methods, that can be used to split realistic, very large to small tests/surveys that do not necessarily measure a single trait, nor manifest uniform inter-item correlation structures, s.t. true scores of different items of the instrument are not necessarily equal, nor inter-related in any prescriptive and assumed fashion. The splitting of the test is done into *optimally-split* subtests by: optimising the inner product of the vectors of scores obtained in the items of each of the 2 subtests; or minimising the absolute difference between the mean item scores attained in the 2 subtests; or by learning the integer-valued indices of the items that make up one of the two subtests, using Bayesian inference by Markov Chain Monte Carlo (or MCMC) techniques, [4], s.t. likelihood of these unknown indices is defined as a decreasing function of the distance between score vectors of items in the 2 subtest items. Subsequent to our learning of the subtests of the test/survey, we compute instrument reliability as complementary to the uncertainty – parametrised as proportional to the variance of the difference between item scores in these *optimally-split* subtests.

Our frequentist splitting that works by minimising the mean difference between the subtest item scores, (Sect. 4.2) borrows from solutions advanced for the "knap-sack" problem in the literature [5–8], among others. [7] 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 (Sect. 4.3), though robustness to outliers in these two methods of partitioning, is not the same. Details of our Bayesian learning of the partitions is discussed in Section 4.4. Our splitting techniques are compared to existing number partitioning methods in Section 5 of the Supplement.

2 Background – test reliability

[9] suggests experimentally identifying the splitting of the questions or items of the test/survey instrument, 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 [10] to be an overestimate when the number of test items is large, or the examinee sample size is small. [11] have shown that the maximal split-half coefficient obtained from the splitting method of [12], 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 [13–14], 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[1].

Indeed it is a hard problem to determine how to split a single test/survey containing a given number of questions or items, into two subtests that contain equal number of items, where we need to ensure equality of quality of each subtest, and ensure that the distance between the scores obtained by the cohort in the two sets of items that the respective subtests comprise, is not an artefact of the splitting, but reflects only the uncertainty of the test/survey. Thus, given the score dataset attained by a cohort of Nresponders, to Pitems in the whole instrument, we seek to split this $N\times P$-dimensional score matrix by columns, such that the same number and quality of items characterise each of the two sought subtests.

[1] There exists another school of thought though in which reliability computation of the entire test is recommended,using the Spearman-Brown formula [15].

Here $N,\ P \in \mathbb{Z}_+$. Indeed, difficulty of retaining the same quality of questions/items in the two subtests will be driven by diversity of quality amongst the items in the original test, i.e. by the degree of inhomogeneity of the inter-item correlation structure relevant to the original test. Such difficulty of splitting will also increase, as the number of items of the original test increases.

This splitting is more demanding than stratified train-test split [16], i.e. the splitting relevant to a Machine Learning approach to a classification problem in which a dataset is desired to be split into training and test datasets in a way that preserves the same proportion of examples in each class, as is manifest in the original dataset.

Maintenance of quality between the 2 subtests is formalised by suggesting that the subtests be "parallel", i.e. all items of the test/survey measure the same latent ability/trait, and the true score of each item is the same constant. Maintaining this restrictive condition for parallelity in real-life tests/surveys is difficult, and often breached. The "tau-equivalent" model relaxes this by allowing item-specific errors or deviations from the true scores, though the true scores of all items are still held equal to each other. 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 [17].

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, [17], leading to the test/survey instrument being criticised (and perhaps discarded) for not producing reliable results [18]. 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 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 [19,20]. Limitations of Cronbach's alpha have been discussed extensively [15,21-24]. Thus, alpha computed for the whole of heterogeneous test/survey, can distort our understanding of its reliability, where such distortion is data-dependent, s.t. a universally-applicable correction is not possible for all instruments. Reliability computed using one of our frequentist splitting methods, is noted in general to be different from reliability computed using 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.

In Section 3 of the Supplement, we show that our definition of test/survey uncertainty, and thereby of instrument reliability, reduces to the congeneric definition of reliability advanced by [25].

3 Model setup

Hereon, we refer to the instrument simply as a test, though the methodology developed below applies to tests and surveys. Similarly, all responders, will be referred to as examinees. As per convention, the variable name is denoted a capital letter while its realisation, the corresponding small letter of the alphabet.

Let us consider a test 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 $\tau_j = \sum\limits_{i=1}^{n} X_i^{(j)}$. 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 number of 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 4.1 and Section 4 of Supporting Materials. 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 $\boldsymbol{X}_m = (X_1^{(m)}, \ldots, X_n^{(m)})^T$; $m = g,\ h$. 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. $\varepsilon_i := X_i^{(g)} - X_i^{(h)}$.

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

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

where the error in the i-th examinee's response is

$$\varepsilon_i := X_i^{(g)} - X_i^{(h)}, \tag{2}$$

and the test variance of the observed test scores is

$$S_X^2 := \frac{\sum_{i=1}^{n}(X_i)^2}{n} - \left(\frac{\sum_{i=1}^{n}X_i}{n}\right)^2, \tag{3}$$

Then it follows from equations (1) and (2) that reliability is

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

We discuss the connection of the data-driven reliability defined above in equation (4). Also, in Section 6, splitting of the test is extended to responses to a survey that is on a k-point Likert scale, where $k \in \mathbb{N}$.

4 Our methods

Estimation of uncertainty of a test score data is extensively studied within Classical Test Theory (CTT), with the complementary test reliability defined as that proportion of the variance of the score attained in a test, that is attributable to the "true score", where such true score differs from the attained or observed score by an additive error, in the CTT paradigm. This theoretical definition then naturally poses the fundamentally difficult problem with implementation, given that the true score is itself unknown [26]. In this paper, we address this conundrum by suggesting fast and reliable solutions for real tests and surveys.

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

We seek to split a given test (constituting ρ 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. $|\sum_{j=1}^{p/2}\left(\tau_j^{(g)} - \tau_j^{(h)}\right)|$ 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. While our splitting algorithm deals with partitioning of an odd-number of elements into the two subtests (Sect. 5 of the Supplement), 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.

In what follows, 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. On the othr hand, Theorem 2 below, 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 S of differences between item scores attained in the $\rho/2$ 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, i.e.*

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

The proof of this theorem is provided in Section 1 of the Supporting Materials.

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 $\varepsilon^2 \mp 2T\varepsilon \mp (p/2)^2 T_p \varepsilon'$, where: absolute difference between sums of examinee scores is ε; difference between the sum of probabilities of correct examinee response to items in the 2 subtests is ε'; Tp is the sum of probabilities of correct examinee response to items in one subtest, and Tis the sum of scores in one of the subtests.*

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

To summarise, the definition of reliability that we delineate in equation (4), is a model that treats the variance of the variable $X_g - X_h$, as the (unnormalised) uncertainty of the given test data, (with the normalisation provided by the test examinee score variance S_X^2).

4.2 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; (Sect. 6 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 [27].

In Algorithm 1 we present our algorithm for identifying the 2 constituent subtests of a given test, by minimising the sum S of absolute difference between the scores obtained in these 2 subtests, i.e. by minimising $S = \sum_{j=1}^{p/2} |\tau_j^{(g)} - \tau_j^{(h)}|$. 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 $\rho/2$ "swaps", where a swap is defined in definition 1.

Definition 1 *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;* $\mathrm{j} = 1, \ldots, \mathrm{p}/2\,; \ell = 1, 2, \ldots, \mathrm{N}_{\mathrm{iter}}$ *Value of S at the j-th swap during the ℓ-th iteration is* $\mathrm{S}\,(\mathrm{l} - 1)\;\mathrm{p}/2 + \mathrm{j}$. *A proposed swap may or may not be accepted depending on whether or not, it results in a lower value of S, which is the absolute difference between sum of components of item score vectors in the 2 subtests, (see Algorithm 1).*

Definition 2 *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-tah 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	$\tau_4 - \tau_3 \leq 0$
	$\vdots$	$\vdots$

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 3 *Once all N_{iter} iterations are undertaken, we identify values of $(\ell - 1)\, p/2 + j$ that minimise $S = \sum_{j=1}^{p/2} |\tau_j^{(g)} - \tau_j^{(h)}|$. Such values of $(\ell - 1)\, p/2 + j$ are:*

$$(\tilde{l} - 1)p/2 + \tilde{J} :=_{(l-1)p/2+j}^{\text{arg}} \quad [\min \quad \left(S_{(l-1)p/2+j}\right)].$$

Then the maximal reliability of the given test, obtained by minimising S, is defined as:

$$r_{tt}^{(\min_S)} := r_{tt}^{(\tilde{l}} - 1)\rho/2 + \tilde{j}.$$

[] 1. In the 0-th iteration, test is split into "seed subtests", (according to Definition 2). Compute $S = S_{seed}$2 (a). In the l-th iteration, the j-th swap, produces a proposed subtest g^{*} and h^{*}. 2(b). At the j-th swap within the l-th iteration, current value of S is$S_{(\ell-1)p/2+j}$.

After the j-th swap, compute proposed value s^* of S, where

$$S^* : |\left[\tau_1^{(g^*)} - \tau_1^{(h^*)}\right] + ... + \left[\tau_{p/2}^{(g^*)} - \tau_{p/2}^{(h^*)]|}.\right.$$

$if^{*} < S_{(\ell-1)p/2+j}$: then – update value of S from the current value $S_{(\ell-1)p/2+j}$, to S^*, – update current subtest g to g^{*}, – update current subtest h by h^{*}. – increment j by 1. – increment j by 1, – proceed to the j+1-th swap with current value $S_{(\ell-1)p/2+j}$ of S, and current subtests g and h. 2(c). At the l-th iteration, and j-th swap, identify examinee score vectors in the current g-th and current h-th subtests and implement in equation (4), to compute reliability $r_{tt}^{((\ell-1)p/2+j)}$. Continue till $p/2$ swaps have been undertaken within the l-th iteration. 3. Set iteration index l to l+1, and proceed till N_{iter} iterations are undertaken.

4.3 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 $s_\rho := \sum_{j=1}^{p/2} \tau_j^{(g)} \tau_j^{(h)}$. It is clear that swapping the j-th item of g-th subtest, with the j-th item of the h-th subtest, will not produce any change in S_p, for $j \in \{1, ..., p/2\}$, as S_p 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 S_p, if $j^{/} \neq j$, j, $j^{/} \in \{1, ..., p/2\}$ Thus, the maximisation of S_p 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 4.3.

Algorithm 1 Algorithm underlining our combinatorial splitting methodology that works by maximising inner product of the subtest item score vectors.

1. In the 0-th iteration, test is split according into the "seed subtests" (Definition 2). Compute$s_\rho = S_{seedp}$.

2(a). In the l-th iteration, the j-th swap, produces a proposed subtest g^{*} and h^{*} where j-th swap comprises: – exchanging j-th item of current g-th subtest with j-th item of current h-th subtests, with j a uniform random integer s.t. $j^{/} \in \{1, ..., j_1, j + 1, ...p/2\}$2(b). At the j-th swap within the l-th iteration, current value of S_ρ is $S_{\rho i\,(\ell-1)p/2+j}$.

After the j-th swap, compute proposed value s_p^* of S_p, where.

$$S_\rho^* := \left[\tau_1^{(g*)} \tau_1^{(h*)}\right] + ... + \left[\tau_{p/2}^{(g*)} \tau_{p/2}^{(h*)}\right].$$

if $S^* < S_{;(\ell-1)p/2+j}$: then – update value of S, from the current value $S_{;(\ell-1)p/2+j}$, to S^*, – update current subtest g to g^{*}, – update current subtest h by h^{*}. – increment j by 1, – proceed to the j+1-th swap with current value $S_{;(\ell-1)p/2+j}$ of S, and current subtests g and h.

2(c). At the l-th iteration, and j-th swap, identify examinee score vectors in the current g-th and current h-th subtests and implement in equation (4), to compute reliability $r_{tt}^{((\ell-1)p/2+j)}$. Continue till $p/2$ swaps have been undertaken within the l-th iteration.

3. Set iteration index l to l + 1, and proceed till N_{iter} iterations are undertaken.

Definition 4 *As in Definition 3, once the iterations are done, identify the $(\ell - 1)\, p/2 + j$ values that maximise S_ρ, using $(\tilde{l} - 1)\rho/2 + \tilde{I} :=_{(l-1)\rho/2+j}^{\text{arg}} \quad [\left(S_{(l-1)p/2+j}\right)],$, and define $r_{tt}^{(\quad \max \quad S_\rho)} : r_{tt}^{(\tilde{l}} - 1)p/2 + \tilde{J}$ as the maximal reliability of the given test obtained by maximising S_ρ.*

Theorem 3 states that the minimisation of S is equivalent to the maximisation of S_ρ.

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 $\tau_m = \left(\tau_1^{(m)}, ..., \tau_{p/2}^{(m)}\right)^T$,, with $\tau_j^{(m)} := \sum_{i=1}^{n} X_i^{(m_j)}; m \in \{g, h\}$. In other words, maximising $|\langle \tau_{g,} \tau_h \rangle| = |\sum_{j=1}^{p/2} \tau_j^{(g)} \tau_j^{(h)}|$ is equivalent to minimising $|\sum_{j=1}^{n} \left(\tau_j^{(g)} - \tau_j^{(h)}\right)|$.*

Proof of Theorem 3 (using Cauchy Schwartz), is in Section 4 of the Supplement.

It is to be noted that the 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$(p/2+1) + p/2 + ... + 1 = (p/2+2)$ $(p/2+1)/2$. Thus, there is a repetition in the value of 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 S. A similar repetition is then noticed in results obtained using splitting by maximising S_p.

Algorithm 2: Algorithm underlining our Bayesian test splitting methodology that works by learning indices of the items of one subtest, with likelihood defined as a smoothly declining function of squared Euclidean distance between subtest item score vectors.

1. In 0-th iteration, test is split into "seed subtests" (Definition 2).

2. In the k-th iteration, propose $g_j^{(k*)}$*Binomial* $\left(p, \Psi^{(k*)}\right), \forall j = 1, ...p/2, \quad \psi \sim \mathit{Uniform}\,[0.5 - a, \quad 0.5 + a]$, with fixed a; $0 < a \leq 0.4$.

3. Identify $h_1^{(*)}, ..., h_{p/2}^{(*)} \in \{1, 2, ..., p\}; h_1^{(*)}, ..., h_{p/2}^{(*)} \notin \{g_1^{(k*)}, ..., g_{p/2}^{(k*)}\};$; sort identified $h_1^{(*)}, ..., h_{p/2}^{(*)}$ as proposed item indices of h-th subtest. Let current parameters be $g_j^{(k-1)}; j = 1, ..., p/2$ and $\psi^{(k-1)}$.

4. Compute acceptance ratio as:

$$\alpha = \frac{\left[\Pi_{j=1}^{p/2}((k-1))\left(\psi^{(k-1)}\right)^{g_j^{(k-1)}}\left(1-\psi^{(k-1)}\right)^{p-g_j^{(k-1)}}\right]}{\left[\Pi_{j=1}^{p/2}((k-1))\left(\psi^{(k*)}\right)\left(1-\psi^{(k*)}\right)^{p-g_j^{(k*)}}\right]}$$

$$\times \frac{\left[\Pi_{j=1}^{p/2}\pi\left(g_1^{(k*)}...g_{p/2}^{(k*)}|\{x_i^{(\ell)}\}_{i=1;\ell=1}^{n;p}\right)\right]}{\left[\Pi_{j=1}^{p/2}\pi\left(g_1^{(k-1)}...g_{p/2}^{(k-1)}\right)|\{x_i^{(\ell)}\}_{i=1;\ell=1}^{n;p}\right]},$$

where for uniform priors,

$$\pi(g_i, \ldots, g_{p/2}|\{x_i^{(\ell)}\}_{i=1;\ell=1}^{n;p}) \propto \prod_{j=1}^{p/2} \exp\left[-\frac{\left(\tau_j^{(g)} - \tau_j^{(h)}\right)^2}{2\sigma^2}\right],$$

while usage of *Binomial* $(p,0.5)$ priors imply multiplying this likelihood to said priors.

5. $a \geq u \sim \mathit{Uniform}\,[0, \ 1] \ \ g_j^{(k)} \leftarrow g_j^{(k\star)}; \ \forall j = 1, \ldots, p/2.$ $g_j^{(k)} \leftarrow g_j^{(k-1)}$, $\forall j = 1, \ldots, p/2$.

6. Continue till $k = N_{iter}$.

4.4 Our new Bayesian splitting of a given test to attain minimum *S*

In our Bayesian approach, we learn the indices g_1, g_2, ..., $g_{p/2}$ 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, ..., $g_{p/2}$, given the test score data $\{x_i^{(j)}\}_{i=1;j=1}^{n;p}$, using Independent Sampler Metropolis Hastings algorithm, [4]. 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 $L \propto \exp\left(-\frac{(\|\tau_g - \tau_h\|)^2}{2\sigma^2}\right)$, where $\|\cdot\|$ denotes the Euclidean norm, and σ^2 is the experimentally fixed variance of this Gaussian likelihood. Thus,

$$L \propto \exp\left[-\frac{\left(\tau_1^{(g)} - \tau_1^{(h)}\right)^2 + \ldots + \left(\tau_{p/2}^{(g)} - \tau_{p/2}^{(h)}\right)^2}{2\sigma^2}\right]$$
$$= \prod_{j=1}^{p/2} \exp\left[-\frac{\left(\tau_j^{(g)} - \tau_j^{(h)}\right)^2}{2\sigma^2}\right].$$

Here, $\tau_j^{(g)} := \sum_{i=1}^{n} x_i^{(g_j)}; \forall j = 1, \ldots, p/2,$ 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, $\forall j = 1, \ldots, p/2$, 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,..., $g_{p/2}$, given the test score data. We generate posterior samples using Metropolis Hastings.

Then using the values of g_1,..., $g_{p/2}$ that are current at the end of the k-th iteration, the h_1,..., $h_{p/2}$ 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

$$g_j^{(k\star)} \sim \mathit{Binomial}\left(p, \psi^{(k\star)}\right),$$

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 $\psi^{(k\star)}$. We do not have any information that permits preference of some values of this rate parameter, to others, though we rule out a probability of $<0.5 - a$ and $>0.5 + a$ for any proposed item in the g-th subtest to be correctly answered in this k-th iteration, where $a = 0.4$ (or $a = 0.2$ in another set of experiments), i.e. rule out this rate parameter to be <0.5–a and $>0.5 + a$. This motivates the model that $\psi^{(k\star)} \sim \mathit{Uniform}\,[0.5 - a, 0.5 + a]$, with a fixed to 0.4 and 0.2 in two separate sets of experiments. 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 the posterior, we compute logarithm of these *Binomial pmf*s using Stirling's approximation. Thus, in the k-th iteration, the acceptance ratio of Metropolis Hastings includes this ratio of the proposal

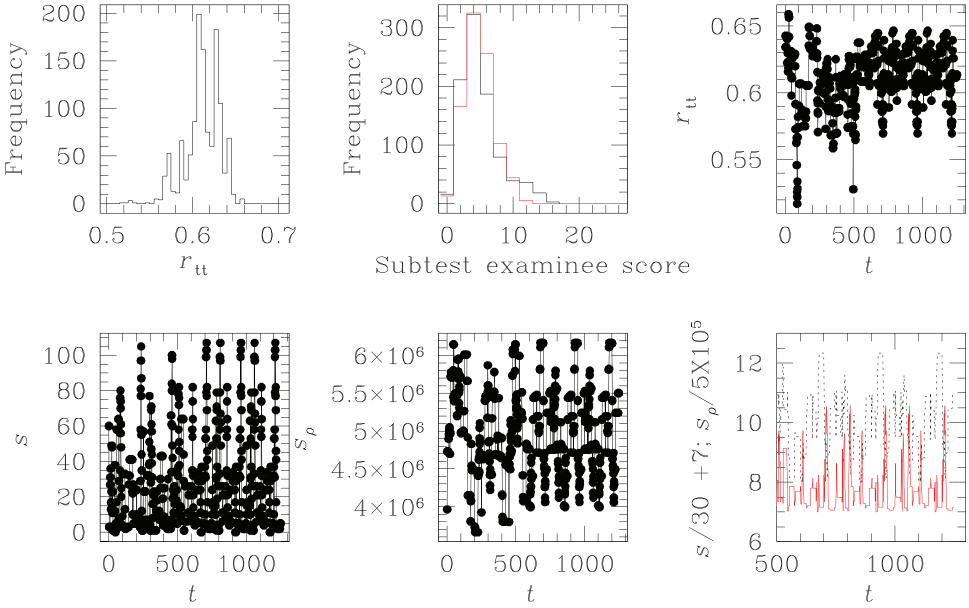

Fig. 1. Results of splitting of the real test data DATA-I, into g-th and h-th subtests of equal (=25) number of items, using minimisation of S, i.e. minimisation of the absolute difference between sum of components of item score vectors in the 2 subtests. *Lower left:* plot of value s of S at the t-th splitting of the test into the g-th and h-th subtests, where the current splitting index $t:=25(\ell-1)+j$, with l 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 a "swap" is defined for this method, in Definition 1. *Lower middle:* plot against t of value s_ρ of S_ρ which is the inner product of item score vectors of g-th and h-th subtests. *Lower right:* plot of linearly transformed S and S_ρ values, against splitting index t, to empirically verify the equivalence between maximisation of S_ρ and minimisation of S (in thin solid lines). scaling and translation of S and S_ρ 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 (Eq. (4)), against splitting index t. *Upper left:* histogram of the r_{tt} values obtained from this run that attains splitting of the given DATA-I test dataset, using minimisation of S. *Upper middle:* histograms of examinee scores obtained at the last accepted swap, in the g-th (in dark solid lines) and h-th subtests (in grey, or red in the electronic version of the paper), identified in this step.

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 current reliability are included. Tests are carried to check on results of varying a, σ^2 and 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 4.3.

5 Empirical illustration on a real data set DATA-I

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 min. This test data has a mean score of about 10.99 and a sample variance of about 19.63. We refer to this dataset as DATA-I.

For the real test data DATA-I, the results obtained by splitting the test via minimisation of the absolute difference S between the sum of components of the item score vectors in the resulting subtests, are shown in Figure 1. Sample mean, or $\tilde{r}_{tt}$ of values of r_{tt} depicted in the upper middle panel of this figure is about 0.6119 and its sample standard deviation (sd) is about 0.0203. We identify $r_{tt}^{(min_S)}\approx0.6589$. The results of splitting by maximisation of the inner product S_ρ of the item score are depicted in Figure 2. The mean reliability achieved by this method of splitting is about 0.5829 and the empirical standard deviation is about 0.0394. We identify $r_{tt}^{(max_{S_\rho})}\approx0.6596$. 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. Histogram of learnt reliability values is presented in the top left panel of this figure, where the learnt 95% Highest Probability Density credible region is [0.5475,0.6525] approximately, with a modal reliability of about 0.6325.

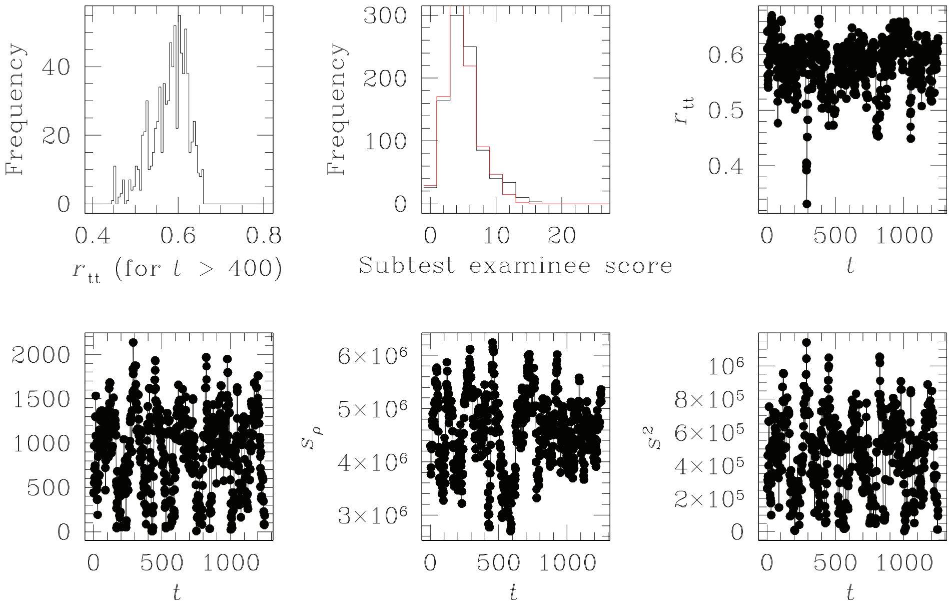

Fig. 2. As in Figure 1, but in this run, splitting of real test data DATA-I is undertaken to maximise the value S_ρ of the inner product S_ρ of item score vectors in the g-th and h-th subtests. Here, the lower right panel displays a plot against the splitting index t, of the value S^2 of the absolute 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 $j\prime$-th item of the current h-th subtest, $(j \neq j\prime)$, under splitting by maximisation of S_ρ, sum of components of the 2 subtests item score vectors, can be more different, than when swapping across rows of the 2 subtests is not permitted, as under splitting by minimising S.

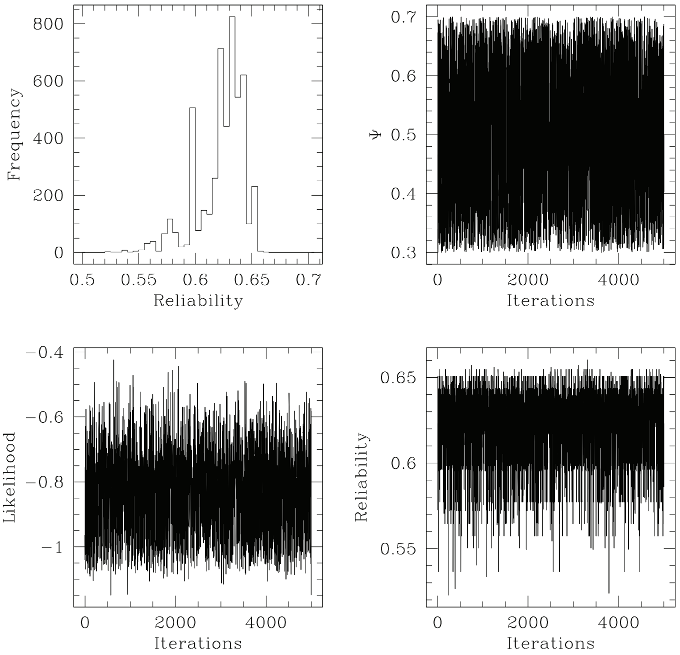

Fig. 3. Figure representing results of splitting the real test dataset DATA-I that comprises responses of n = 912 examinees in 50 items, using Bayesian learning of the indices of the items in the g-th subtest. The remaining items constitute the $h.$ −th subtest. Posterior sampling is performed with Independent Sampler Metropolis Hastings, in which each item index of the g-th subtest is proposed from a $Binomial(50, \psi)$, 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, of ψ, and of the likelihood are presented in the lower right, top right and lower left panels respectively; the traces indicate convergence.

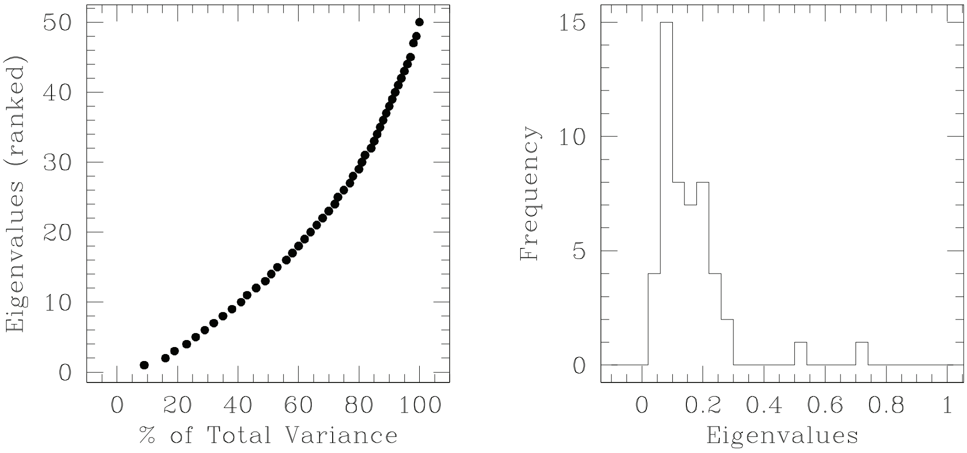

Fig. 4. Figure showing results of PCA of real test data DATA-I. On the right histogram of the eigenvalues is displayed, while the left panel depicts the eigenvalues (ranked by weights) needed to explain the fraction of the total variance.

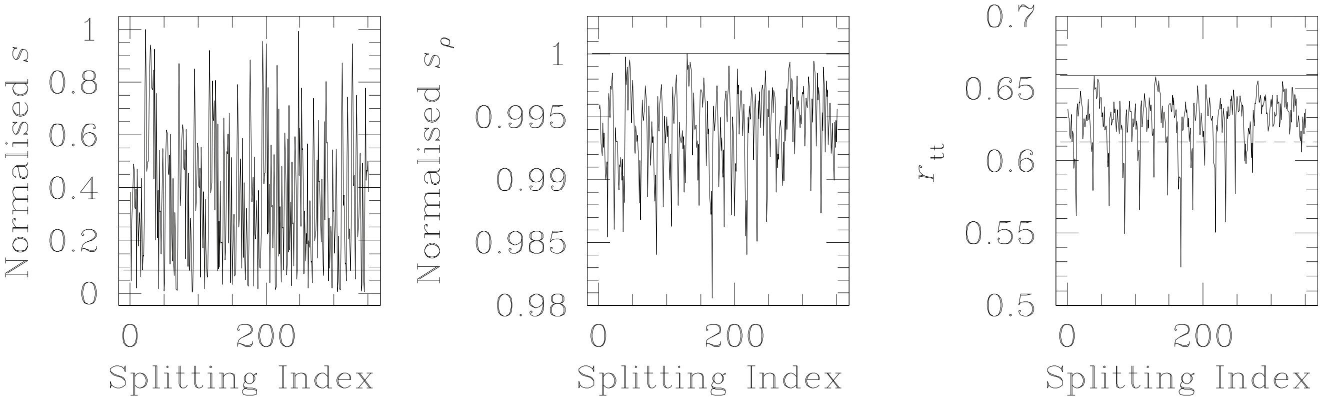

Fig. 5. Figure showing results for each of the 351 possible splittings of the read test data DATA-I, where the said results include the absolute difference S between sums of components of the item score vectors in the 2 subtests that result from the splitting, (*left panel*); inner product S_ρ of the 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,...351 for DATA-I. Our results by minimising S are overplotted in solid lines, on these results, in the right panel. Cronbach alpha for DATA-I is also computed and overplotted upon the computed reliability values in the right panel, in broken lines.

5.1 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 S between sums of components of item score vectors of the subtests that result from 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 S; S_ρ; and reliability r_{tt} (using Eq. (4)). Cronbach's alpha for this real test dataset is compared to such computed reliabilities in Figure 5.

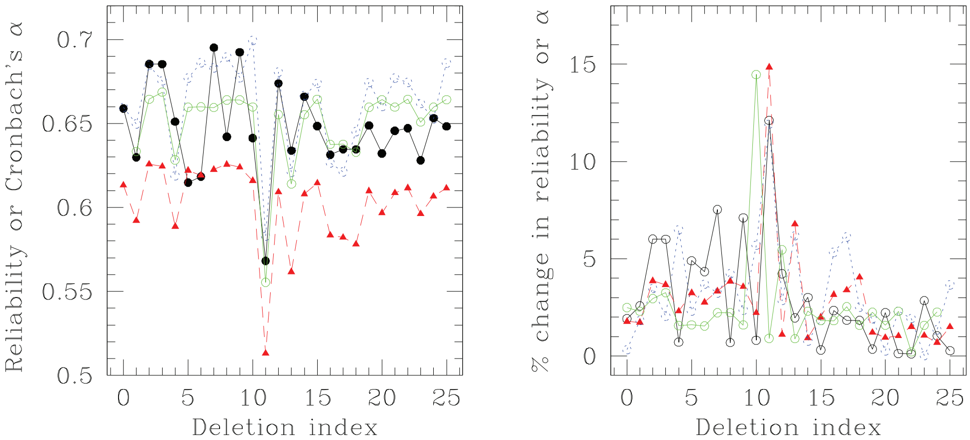

Fig. 6. *Left*: figure to bring out robustness to outliers, of the different techniques for computing reliability. Reliability computed by deleting the q-th highest scoring pair of items – from the data on responses to the learnt subtests of the test/survey – is plotted against the deletion index q. Results otained with subtests learnt via minimisation of S are in black filled circles, joined by a black solid line); via maximisation of S_ρ are in open circles joined by a broken line, (in blue in the electronic version); and via Bayesian inference on the indices of the items that comprise the g-th subtest are in filled circles joined by a grey solid line, (in green in the electronic version). Reliability computed by Cronbach alpha is also plotted in each case (in filled triangles joined by a broken line – in red in the electronic version). *Right:* 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 S_ρ is the least robust (with a variance of about 3.25_2) The reliability computed by minimising S and Cronbach alpha are nearly equally robust, with variances of about 2.96_2 and 2.95_2 respectively.

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, at each deletion of the q-th highest scoring pair of items from the test data DATA-I, we undertook computation of reliability by minimising S; reliability by maximising S_ρ; 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; $q = 1,...,25$. 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.

6 Reliability of a real survey with categorical responses – *HSQ*

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 Eq. (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 4.4, and the publicly available data that is reported by [28], 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. Figure 7 depicts results obtained by splitting this real test dataset *HSQ*, using our Bayesian learning of $g_1,...,g_{p/2}$, leaving the remaining test items to build up the g-th subtest. All parameters of the Metropolis Hastings chain are as used for the Bayesian learning given DATA-I (Sect. 4.4). As in Figures 3 and 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

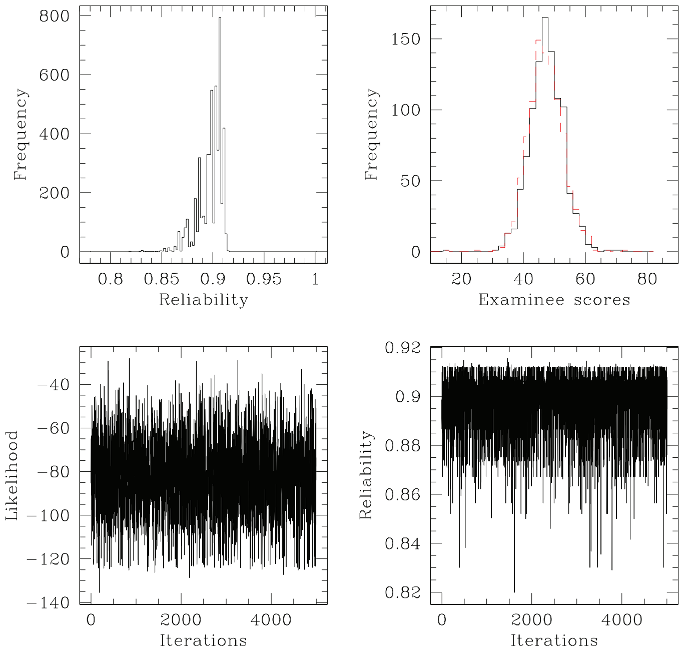

Fig. 7. Figure representing results of Bayesian splitting the real survey dataset *HSQ* that comprises responses on a 5-point Likert scale. We use responses from $n = 1022$ responders (who we refer to generically as "examinees") who answered every one of the 32 items of this test. Here we Bayesianly 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. 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 (Eq. (4)). Traces of this reliability, and of the likelihood are presented in the lower right, and lower left panels respectively. 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.

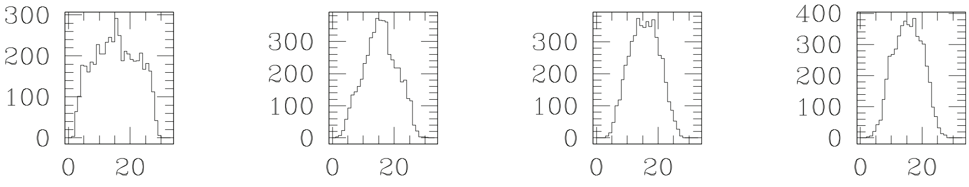

Fig. 8. Marginal posterior probability density of the 1st, 5th, 9th and 13th item indices of an identified subtest between the subtest pair that real test data *HSQ* is split into. The marginals are represented as histograms here.

iteration of this MCMC chain. Histogram of learnt reliability values is presented in the top left panel of this figure, where the learnt 95% Highest Probability Density credible region is about [0.847,0.915], with the modal reliability of about 0.907. 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.

6.1 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 $j - j\prime$-th item pair, where $j\prime \leq j$, $j = 1,2,...,p$. $p = 32$ 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 C that gives the number of inter-item covariance

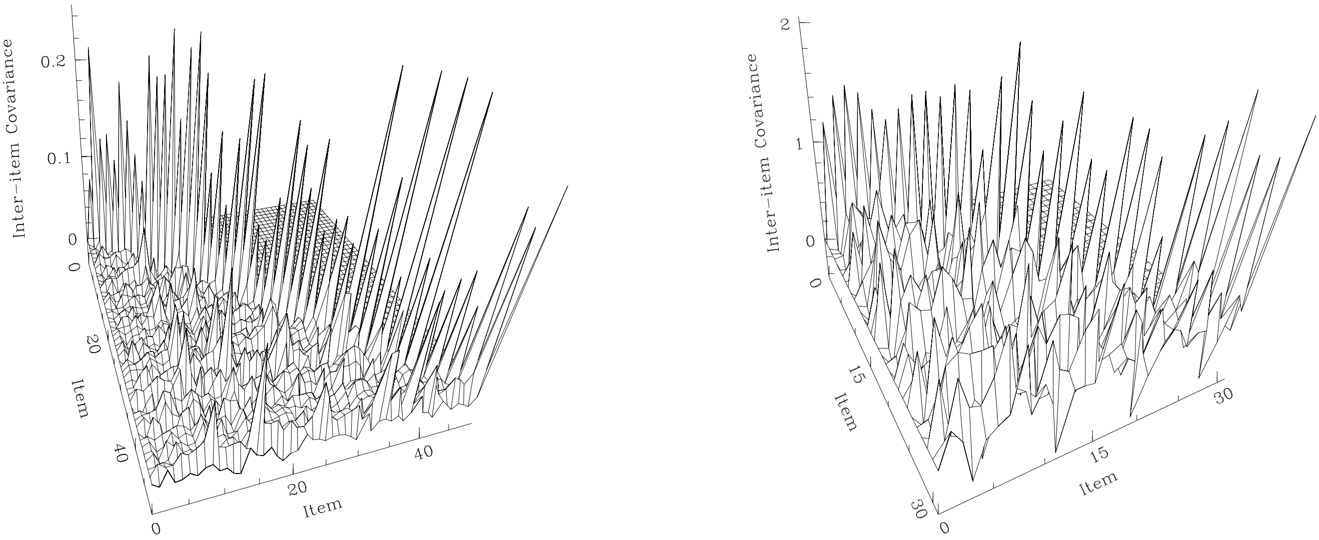

Fig. 9. Surface plot of covariance between pairs of items in the given test data (DATA-I on the left, and *HSQ* in the right panel), plotted against item indices. Here only the lower-triangle of the inter-item variance-covariance matrix is plotted, i.e. covariance between the j-th and j'-th item is plotted $\forall j' \leq j; j = 1, 2, \ldots, p$; $p = 50$ 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, (discussed in Sect. 6.1). For *HSQ*, $C \approx 20\%$, while the inter-item covariance sample of DATA-I, causes C to about 8.7%.

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 number 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*.

7 Reliability of a very large binary data using minimisation of 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 [29], 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 3. 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 unidimensional. 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 minimisation of S, we obtain results in $r_{tt}^{(min_s)} \approx 0.9258$. The splitting that corresponds to the minimum 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

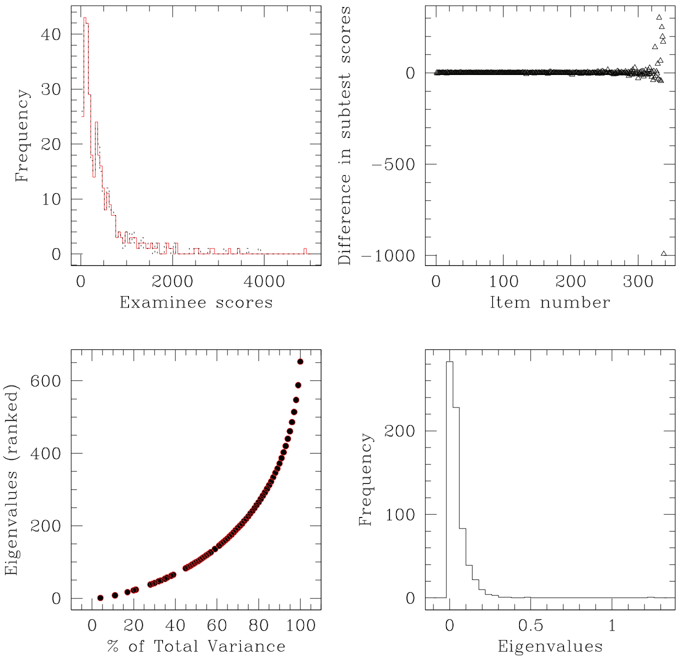

Fig. 10. Figure representing results obtained using the very large real dataset DATA-YELP that comprises binary responses on 676 variables (or items), by 8848 responders (or examinees). The eigenvalue weight distribution is shown by the histogram in the *lower right panel*. Relevance of at least 6 of 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 explained fractional variation, is drawn by undertaking the PCA 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 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 $j = 1,...,676/2 = 338$. Reliability corresponding to the minimisation of S is about 0.93, while Cronbach alpha for this data is about 0.91.

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 item pair index, in the upper right of this figure.

8 Conclusions

We present multiple data-driven methods for learning the optimally similar pair of subtests that comprise a given test/survey, in order to enable the computation of uncertainty in the data on responses to questions of large/small, heterogeneous, multi-dimensional real-life test/survey instrument – where the response is either binary or categorical – without needing to invoke restrictive model assumptions that cannot be practically adhered to, and are typically, not adhered to in reality.

The splitting methods that we advance, are not affected by messiness that typifies real test/survey response data, and by 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). We illustrate by splitting a real test data that our 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 S is comparably robust.

Above, we have advanced 3 different methods of splitting a test or a survey into 2 subtests, s.t. variance of the difference between responders' scores attained in the 2 subtests, normalised by the variance of the responses to he full test, is defined as uncertainty of the test data; the test reliability is then complementary to this uncertainty. The 3 methods are essentially equivalent, and operate by learning the 2 optimally-similar subtests that a given test/survey instrument is split into, where such optimal splitting is undertaken: by minimising the absolute difference 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 offer a fully objective protocol for implementation of each method, and illustrate these on real-world datasets comprising responses to a test relevant to recruitment of personnel for a real position; to a survey relevant to an undertaken psychometric task; and to a survey on restaurant reviews input by customers. We forward our methods towards fast, automated and reliable uncertainties of real test and surveys.

Declarations

- There is no funding to be reported.
- The authors declare that they have no competing interests.
- Ethics approval is not relevant for the work reported in this article.
- All authors give their consent for publication.
- Public availability of 2 of the data sets used in the work are mentioned within the text, and the anonymised version of the third data set that was used in the paper will be uploaded.
- All codes used for analysis in the text will be made available upon acceptance for publication.
- All authors have contributed equally to the work.

Supplementary material

The article is accompanied by supplementary information that includes proofs to the theorems that are stated within the text of the article; linking our advanced methods to extant congeneric methods in the literature; comparison of the methods discussed herein, for partitioning a set of integers into 2 subsets and presentation of results on simulated data.

The Supplementary Material is available at https://www.metrology-journal.org/10.1051/ijmqe/2023018/olm.

References

1. S. Bell, A beginner's guide to uncertainty of measurement, 2001
2. C. Borgs, J. Chayes, B. Pittel, Phase transition and finite-size scaling for the integer partitioning problem, Random Struct. Algorithms **19**, 247–288 (2001)
3. J. Brownlee, Train-test split for evaluating machine learning algorithms, 2020
4. J.C. Callendar, H.G. Osburn, A method for maximizing split-half reliability coefficients, Educ. Psychol. Meas. **37**, 819–825 (1977)
5. E. Cho, Making reliability reliable: a systematic approach to reliability coefficients, Organ. Res. Methods **19**, 651–682 (2016)
6. J. Cortina, What is coefficient alpha: an examination of theory and applications, Int. J. Appl. Psychol. **78**, 98–104 (1993)
7. R. Eisinga, M. Te Grotenhuis, B. Pelzer, The reliability of a two-item scale: Pearson, Cronbach or Spearman-Brown? Int. J. Public Health **58**, 637–642 (2012)
8. M.R. Garey, D.S. Johnson, Computers and intractability: a guide to the theory of np-completeness of mathematical sciences series (Freeman, 1979)
9. J.M. Graham, Congeneric and (Essentially) tau-equivalent estimates of score reliability; what they are and how to use them, Educ. Psychol. Meas. **66**, 930–944 (2006)
10. S. Green, R. Lissitz, S. Mulak, Limitations of coefficient alpha as an index of test unidimensionality, Educ. Psychol. Meas. **37**, 827–838 (1977)
11. C.T. Gualtieri, L.G. Johnson, Reliability and validity of a computerized neurocognitive test battery, CNS vital signs, Arch. Clin. Neuropsychol. **21**, 623–643 (2006)
12. H. Gulliksen, *Theory of mental tests* (Taylor & Francis Group, New York, Oxford, 1987)
13. L. Guttman, A basis for analysing test-retest reliability, Psychometrika **10**, 255–282 (1945)
14. B. Hayes, The easiest hard problem, Am. Sci. **90**, 113 (2002)
15. N. Karmarkar, R.M. Karp, An efficient approximation scheme for the one-dimensional bin-packing problem, 23rd Annual Symposium on Foundations of Computer Science (sfcs 1982), Chicago, IL, USA,1982, pp. 312–320, doi: 10.1109/SFCS.1982.61
16. M. Tavakol, Making sense of Cronbach's alpha, Int. J. Med. Educ. **2**, 53–55 (2011)
17. R.A. Martin, P. Puhlik-Doris, G. Larsen, J. Gray, K. Weir, Individual differences in uses of humor and their relation to psychological well-being: development of the humor styles questionnaire. J. Res. Pers. **37**, 48–75 (2003)
18. M. Meadows, L. Billington, A review of the literature on marking reliability, 2005
19. S. Mertens, The Easiest hard problem: number partitioning, in: A. Percus, G. Istrate, C. Moore (Eds.), *Computational Complexity & Statistical Physics*, Oxford University Press, Oxford, 2006, pp. 125–139
20. K.R. Murphy, C.O. Davidshofer, *Psychological testing: principles and applications* (Pearson/Prentice Hall, 2005)
21. P. Panayides, Coefficient alpha interpret with caution, Eur. J. Psychol. **9**, (2013)
22. N. Ritter, Understanding a widely misunderstood statistic: cronbach's alpha, 2010
23. C.P. Robert, G. Casella, *Monte Carlo statistical methods* (Springer-Verlag, New York, 2004)
24. H. Sajnani, V. Saini, K. Kumar, E. Gabrielova, P. Choudary, C. Lopes, *Yelp Dataset Challenge* http://mondego.ics.uci.edu/projects/yelp/
25. K. Sijtsma, On the use, the misuse, and the very limited usefulness of Cronbach's Alpha, Psychometrika **74**, 107–120 (1994)
26. D.L. Streiner, Starting at the beginning : an introduction to co-efficient alpha and consistency, J. Pers. Assess. **80**, 99–103 (2003)
27. J. Ten Berge, G. Socan, The greatest lower bound to the reliability of a test and the hypothesis of unidimensionality, Psychometrika **69**, 613–625 (2004)
28. B.L. Thompson, S.B. Green, Y. Yang, Assessment of the maximal split-half coefficient to estimate reliability, Educ. Psychol. Meas. **70**, 232–251 (2006)
29. N.M. Webb, R.J. Shavelson, E.H. Haertel, Reliability coefficients and generalizability theory, *Handbook of Statistics*, 2006

Cite this article as: Satyendra Nath Chakrabartty, Wang Kangrui, Dalia Chakrabarty, Reliable uncertainties of tests and surveys – a data-driven approach, Int. J. Metrol. Qual. Eng. **15**, 4 (2024)