Supplement to "Reliable Uncertainties of Tests & Surveys - a Data-driven Approach"

Satyendra Nath Chakrabartty, Wang Kangrui, Dalia Chakrabarty

November 15, 2022

In this document, we refer to the main paper as SNCKWDC.

1 Means of subtest examinee scores obtained by splitting a test by minimising absolute difference $\mathcal{S}$, between sums of components of subtest item score vectors

Theorem 1. *Minimising the absolute sum of differences between item scores attained in the $p/2$ items of the pair of subtests that are generated by splitting a test containing p items, 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,*

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

Proof. Let the score vectors comprising scores for each of the n examinees, in the g-th and h-th subtests be $\boldsymbol{X}^{(g)} := (X_1^{(g)}, \ldots, X_n^{(g)})^T$ and $\boldsymbol{X}^{(h)} := (X_1^{(h)}, \ldots, X_n^{(h)})^T$.

Let score attained by i-th examinee in j-th item of m-th subtest be $X_i^{(g_j)}$; $i = 1, \ldots, n$, $j = 1, \ldots, p/2$, $m = g, h$. This the index of the j-th item in the m-th subtest is m_j, where $m_j \in \{1, \ldots, p\}$. Then $X_i^{(m)} = \sum_{j=1}^{p/2} X_i^{(m_j)}$.

Now, mean of the g-th subtest is $\bar{X}_g = \frac{\sum_{i=1}^{n} X_i^{(g)}}{n}$, and mean of the h-th subtest is $\bar{X}_h = \frac{\sum_{i=1}^{n} X_i^{(h)}}{n}$.

Now, item score of the j-th item in the m-th subtest is $\tau_j^{(m)}$, $j = 1, \ldots, p/2$, $m = g, h$.

But, sum of item scores in all $p/2$ items in a subtest, is equal to sum of examinee scores achieved in these $p/2$ items, i.e.

$$\begin{aligned} \sum_{i=1}^{n} X_i^{(g)} &= \sum_{j=1}^{p/2} \tau_j^{(g)} \\ \sum_{i=1}^{n} X_i^{(h)} &= \sum_{j=1}^{p/2} \tau_j^{(h)} \end{aligned} \tag{1.1}$$

Now, we define

$$\begin{aligned} |\mathcal{S}| &:= \left| \left[\tau_1^{(g)} - \tau_1^{(h)}\right] + \ldots + \left[\tau_{p/2}^{(g)} - \tau_{p/2}^{(h)}\right] \right| \\ &= \left| \left[\tau_1^{(g)} + \ldots + \tau_{p/2}^{(g)}\right] - \left[\tau_1^{(h)} + \ldots + \tau_{p/2}^{(h)}\right] \right|. \end{aligned}$$

At the end of the splitting of the test into the g-th and h-th subtests, where we undertake this splitting to ensure minimum $\mathcal{S}$, let $|\mathcal{S}| = \epsilon$; $\epsilon \geq \mathbb{R}_{\geq 0}$. Then by definition of our method of splitting, ϵ is the minimum value of $|\mathcal{S}|$ by our method.

Then using

$$\mathcal{S} = \left| \left[\tau_1^{(g)} + \ldots + \tau_{p/2}^{(g)}\right] - \left[\tau_1^{(h)} + \ldots + \tau_{p/2}^{(h)}\right] \right| = \epsilon$$

in Equations 1.1, we get

$$\left| \sum_{i=1}^{n} X_i^{(g)} - \sum_{i=1}^{n} X_i^{(h)} \right| = \left| \sum_{j=1}^{p/2} \tau_j^{(g)} - \sum_{j=1}^{p/2} \tau_j^{(h)} \right| = \epsilon \tag{1.2}$$

Then

$$\left|\bar{X}_g - \bar{X}_h\right| = \left|\frac{\sum_{i=1}^{n} X_i^{(g)}}{n} - \frac{\sum_{i=1}^{n} X_i^{(h)}}{n}\right| = \epsilon/n,$$

using Equation 1.2.

Thus, (absolute) difference between means of g-th and h-th subtests is minimised, to ϵ/n, when (absolute) difference between sums of scores attained in items that comprise the 2 subtests, is minimised to ϵ.

2 Variance of subtests obtained by splitting a test by minimising absolute difference $\mathcal{S}$, between sums of components of subtest item score vectors

Theorem 2. *Absolute difference between sums of squares of examinee scores in the g-th and h-th sub-tests 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 $T \pm \epsilon$.*

Proof. Let the absolute difference between sums of scores in the g-th and h-th sub-tests that the given test is split into by our method, be $\epsilon \geq \mathbb{R}$. Then ϵ is the minimum value of this absolute difference (Theorem 3.1 of SNCKWDC). Thus,

$$\sum_{i=1}^{n} X_i^{(g)} = \sum_{i=1}^{n} X_i^{(h)} \pm \epsilon,$$

where this minimum value ϵ is small.

Then in the approximation that $\epsilon \approx 0$, we state:

$$\sum_{i=1}^{n} X_i^{(g)} \approx \sum_{i=1}^{n} X_i^{(h)},$$

where the order of this approximation is $\pm\epsilon$.

Then total of scores obtained in the g-th subtest is

$$T := \sum_{i=1}^{n} X_i^{(g)} \Longrightarrow \sum_{i=1}^{n} X_i^{(h)} = T \mp \epsilon. \tag{2.1}$$

Using the notation that, score obtained by i-th examinee in the j-th item of the g-th subtest is $X_i^{(g_j)}$, we define

$$X_i^{(g)} = \sum_{j=1}^{p/2} X_i^{(g_j)}, \quad \text{where} \quad X_i^{(g_j)} = 0 \quad \text{or} \quad 1,$$

and $g_j \in \{1, 2, \ldots, p\}$.

Similarly, we define $X_i^{(h)}$, $\forall i = 1, \ldots, n$.

Now,

$$\begin{aligned} \left(\sum_{i=1}^{n} X_i^{(g)}\right)^2 = & \quad \sum_{i=1}^{n} \left(X_i^{(g)}\right)^2 + \\ & X_1^{(g)} X_2^{(g)} + \ldots + X_1^{(g)} X_n^{(g)} + \\ & \ldots + \\ & X_n^{(g)} X_1^{(g)} + \ldots + X_n^{(g)} X_{n-1}^{(g)} \\ = & \sum_{i=1}^{n} \left(X_i^{(g)}\right)^2 + \sum_{i=1}^{n} \left[\sum_{k=1;k\neq i}^{n} X_i^{(g)} X_k^{(g)}\right] \end{aligned} \tag{2.2}$$

Now, for $k \neq i$, $X_i^{(g)} X_k^{(g)} =$

$$
\begin{aligned}
& \left(X_i^{(g_1)} + \ldots + X_i^{(g_{p/2})}\right)\left(X_k^{(g_1)} + \ldots + X_k^{(g_{p/2})}\right) \\
= \; & X_i^{(g_1)} X_k^{(g_1)} + X_i^{(g_1)} X_k^{(g_2)} + \ldots + X_i^{(g_1)} X_k^{(g_{p/2})} + \\
& \ldots + \\
& X_i^{(g_{p/2})} X_k^{(g_1)} + X_i^{(g_{p/2})} X_k^{(g_2)} + \ldots + X_i^{(g_{p/2})} X_k^{(g_{p/2})} \\
= \; & \sum_{j=1}^{p/2} \left[\sum_{j'=1}^{p/2} \left(\text{number of times } X_i^{(g_j)} = 1 \text{ and } X_k^{(g'_j)} = 1 \right) \right] \\
= \; & \left[\sum_{j=1}^{p/2} \left(\text{number of times } X_i^{(g_j)} = 1 \right) \right] \times \\
& \left[\sum_{j'=1}^{p/2} \left(\text{number of times } X_k^{(g'_j)} = 1 \right) \right] \\
\approx \; & (p/2)^2 \left[\sum_{j=1}^{p/2} \Pr(X_i^{(g_j)} = 1) \right] \left[\sum_{j'=1}^{p/2} \Pr(X_k^{(g'_j)} = 1) \right] \qquad (2.3)
\end{aligned}
$$

But $X_i^{(g_j)} \sim$ Bernoulli$(\pi_i^{(j)})$, i.e. $\Pr(X_i^{(g_j)} = 1) = \pi_i^{(g_j)}$. The approximation in Equation 2.3 stems from approximating the probability for an event with its relative frequency. Then following Equation 2.3, we get

$$
\begin{aligned}
\sum_{i=1}^{n} \sum_{k=1;k\neq i}^{n} X_i^{(g)} X_k^{(g)} \; & \approx \; (p/2)^2 \sum_{i=1}^{n} \sum_{k=1;k\neq i}^{n} \left[\sum_{j=1}^{p/2} \pi_i^{(g_j)} \sum_{j'=1}^{p/2} \pi_k^{(g_{j'})} \right] \\
& = \; (p/2)^2 \left[\sum_{i=1}^{n} \left(\sum_{j=1}^{p/2} \pi_i^{(g_j)} \right) \right] \left[\sum_{k=1;k\neq i}^{n} \left(\sum_{j'=1}^{p/2} \pi_k^{(g_{j'})} \right) \right] \qquad (2.4)
\end{aligned}
$$

Now, Equation 2.1 implies that

$$\sum_{i=1}^{n}\left[\sum_{j=1}^{p/2} X_i^{(g_j)}\right] \approx \sum_{i=1}^{n}\left[\sum_{j=1}^{p/2} X_i^{(h_j)}\right] \quad \text{i.e.}$$

$$\sum_{i=1}^{n}\left[\sum_{j=1}^{p/2}\left(\text{number of times } X_i^{(g_j)}=1\right)\right] \approx \sum_{i=1}^{n}\left[\sum_{j=1}^{p/2}\left(\text{number of times } X_i^{(h_j)}=1\right)\right] \quad \text{i.e.}$$

$$\sum_{i=1}^{n}\left[\sum_{j=1}^{p/2} \pi_i^{(g_j)}\right] \approx \sum_{i=1}^{n}\left[\sum_{j=1}^{p/2} \pi_i^{(h_j)\cdot}\right] \tag{2.5}$$

Then if we delete any 1 out of the n examinees, over which the outer summation on the RHS and LHS of the approximation 2.5 is carried out, it is expected that this approximation expressed in statement 2.5 would still be valid. This is especially the case if n is large. In other words, bigger the n, smaller is the distortion affected on the structure of the sub-tests generated by splitting the test data obtained after deleting the score of any 1 of the n examinees from the original full test data. Then using statement 2.5 for a large n, we can write

$$\sum_{k=1,k\neq i}^{n}\left[\sum_{j=1}^{p/2} \pi_k^{(g_j)}\right] \approx \sum_{k=1;k\neq i}^{n}\left[\sum_{j=1}^{p/2} \pi_k^{(h_j)}\right] \tag{2.6}$$

where

$$\sum_{i=1}^{n}\left[\sum_{j=1}^{p/2} \pi_i^{(g_j)}\right] = \sum_{i=1}^{n}\left[\sum_{j=1}^{p/2} \pi_i^{(h_j)}\right] \pm \epsilon',$$

from Equation 2.1, i.e. from $\sum_{i=1}^{n} X_i^{(g)} = \sum_{i=1}^{n} X_i^{(h)} + \epsilon$ or $\sum_{i=1}^{n}\left[\sum_{j=1}^{p/2} X_i^{(g_j)}\right] = \sum_{i=1}^{n}\left[\sum_{j=1}^{p/2} X_i^{(h_j)}\right] + \epsilon$. Here, expectation $\mathbb{E}X_i^{(g_j)} = \pi_i^{(g_j)}$ (or $\mathbb{E}X_i^{(h_j)} = \pi_i^{(h_j)}$), is used in place of the variable. Then $\epsilon' \in \mathbb{R}_{\geq 0}$ is s.t. $\epsilon \geq \epsilon'$, given that ϵ' is the absolute difference between sums of probability of correct response in the g-th sub-test and that in the h-th sub-test while ϵ is the absolute difference between sums of scores in the two sub-tests.

In other words, if we define the sum of probabilities of correct response in the g-th sub-test, T_P, as

$$T_P := \sum_{k=1,k\neq i}^{n} \left[\sum_{j=1}^{p/2} \pi_k^{(g_j)} \right],$$

then following Equation 2.1,

$$\sum_{k=1,k\neq i}^{n} \left[\sum_{j=1}^{p/2} \pi_k^{(h_j)} \right] \approx T_P \mp \epsilon'. \tag{2.7}$$

Using this, for sub-test g, in the last line of Equations 2.4 we get

$$\sum_{i=1}^{n} \sum_{k=1;k\neq i}^{n} X_i^{(g)} X_k^{(g)} \approx (p/2)^2 T_P \sum_{i=1}^{n} \left[\sum_{j=1}^{p/2} \pi_i^{(g_j)} \right], \tag{2.8}$$

where the approximation above stems from the approximation in statement 2.6. For sub-test h,

$$\sum_{i=1}^{n} \sum_{k=1;k\neq i}^{n} X_i^{(h)} X_k^{(h)} \approx (p/2)^2 T_P \sum_{i=1}^{n} \left[\sum_{j=1}^{p/2} \pi_i^{(h_j)} \right],$$

where the last approximation deviates from an equality, by a term that is of the order as in statement 2.8, enhanced by $\mp(p/2)^2 T_P \epsilon'$, following statement 2.7. Then

$$\sum_{i=1}^{n} \sum_{k=1;k\neq i}^{n} X_i^{(g)} X_k^{(g)} \approx \sum_{i=1}^{n} \sum_{k=1;k\neq i}^{n} X_i^{(h)} X_k^{(h)}, \tag{2.9}$$

where the error in this approximation is of the order of $(p/2)^2 T_P \epsilon'$.

But, from last line of Equations 2.2, $\left(\sum_{i=1}^{n} X_i^{(g)} \right)^2 - \sum_{i=1}^{n} \left(X_i^{(g)} \right)^2 = \sum_{i=1}^{n} \sum_{k=1;k\neq i}^{n} X_i^{(g)} X_k^{(g)}$. Then using statement 2.9 in this, we get that for the g-th and the h-th sub-tests,

$$\left(\sum_{i=1}^{n} X_i^{(g)} \right)^2 - \sum_{i=1}^{n} \left(X_i^{(g)} \right)^2 \approx \left(\sum_{i=1}^{n} X_i^{(h)} \right)^2 - \sum_{i=1}^{n} \left(X_i^{(h)} \right)^2, \tag{2.10}$$

where–as for statement 2.9–the error in this approximation is of the order of $(p/2)^2 T_P \epsilon'$.

But $\left(\sum_{i=1}^{n} X_i^{(g)}\right)^2 \approx \left(\sum_{i=1}^{n} X_i^{(h)}\right)^2$ (by squaring both sides of Equation 2.1), where the error in this approximation is of the order of $\epsilon^2 \mp 2T\epsilon$. Then in statement 2.10 we get

$$\sum_{i=1}^{n} \left(X_i^{(g)}\right)^2 \approx \sum_{i=1}^{n} \left(X_i^{(h)}\right)^2, \tag{2.11}$$

where the error in this approximation is of the order of $-\epsilon^2 \pm 2T\epsilon \mp (p/2)^2 T_P \epsilon'$. In other words, if absolute difference between total scores attained in the 2 subtests is ϵ, then absolute difference between total squared scores in the 2 subtests is of the order of $\epsilon^2 \mp 2T\epsilon \pm (p/2)^2 T_P \epsilon'$.

3 Linking our defined reliability to general models for reliability, and strength of our splitting methods

To summarise, the definition of reliability that we delineate in Equation 4 of SNCKWDC 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 being provided by the test examinee score variance S_X^2.

While such a measure of test uncertainty as ours, can be held equivalent to existing general models of reliability under certain conditions (discussed below), it is important to appreciate that the flexibility of our models lies in the advanced methods for splitting the given full test into the g-th and h-th subtests - scores in which can then be input into our definition of reliability, namely Equation 1 of SNCKWDC. The splitting methods that we have advanced satisfy the following desirable properties:

1. splitting is not affected if the test inter-item correlations are not equal or comparable; all three of our advanced splitting methods work in the presence of a non-uniform test correlation structure.

2. splitting is also unaffected if the test is not strictly uni-dimensional, while Cronbach alpha is affected by multi-dimensionality.

3. splitting is unaffected by size of the test dataset, i.e. methods can handle large, as well as small datasets.

A general (congeneric) model for reliability that is available in the literature is exemplified by the model advanced by [2] who defines reliability as $(\sum_{j=1}^{p} \lambda_j^2)/S_X^2$, where λ_j is the loading on the j-th item, s.t. variance of the j-th item's score variable is given by $\lambda_j^2 + S_{\nu_j}^2$, with $S_{\nu_j}^2$ representing the error variance of this j-th item's score.

Now, recalling Equation 1 of SNCKWDC, with the the error variable ϵ defined as in Equation 2 of SNCKWDC (as difference between an examinee's scores in the 2 subtests), our definition of reliability in Equation 4 of SNCKWDC suggests that

$$r_{tt} \;=\; 1 - \frac{S_\epsilon^2}{S_X^2} = 1 - \frac{S_{X_g - X_h}^2}{S_X^2} = \frac{S_X^2 - S_\epsilon^2}{S_X^2},$$

Thus our definition of reliability reduces to the general (congeneric) definition of reliability advanced by [2], if we set $S_X^2 - S_\epsilon^2 \equiv \sum_{j=1}^{p} \lambda_j^2$. In other words, our definition of reliability reduces to the general congeneric reliability, in the paradigm that the variance S_X^2 of the examinee scores in the full test, is equal to the sum of variances of all p item scores, given that $\sum_{j=1}^{p} S_{\nu_j}^2 = S_\epsilon^2$. In other words, our definition of reliability equates the congeneric one, given equality of the sum $\sum_{j=1}^{p} S_{\nu_j}^2$ of error variances of all item scores, and error variance S_ϵ^2 of examinee scores in the full test.

4 Minimising absolute difference $\mathcal{S}$ between sums of components of subtest item score vectors, is equivalent to maximising their inner product $\mathcal{S}_\rho$

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. Let $\mathcal{S}_\rho := \left|\langle \boldsymbol{\tau}_g, \boldsymbol{\tau}_h \rangle\right|$. By Cauchy Schwartz: $\left|\langle \boldsymbol{\tau}_g, \boldsymbol{\tau}_h \rangle\right| \leq \| \boldsymbol{\tau}_g \| \| \boldsymbol{\tau}_h \|$, where in our notation, $\| \cdot \|$ is the L^2 norm. Now for any vector $\boldsymbol{A} \in \mathbb{R}^d$, for any $d \in \mathbb{N}$, $\| \boldsymbol{A} \| \leq \| \boldsymbol{A} \|_1$, where $\| \boldsymbol{A} \|_1$ is the L^1 norm of this arbitrary vector $\boldsymbol{A}$. Therefore,

$$\begin{aligned} \mathcal{S}_\rho = \left|\langle \boldsymbol{\tau}_g, \boldsymbol{\tau}_h \rangle\right| &\leq \| \boldsymbol{\tau}_g \| \| \boldsymbol{\tau}_h \|_1, \\ \text{i.e.} \quad \left|\langle \boldsymbol{\tau}_g, \boldsymbol{\tau}_h \rangle\right| &\leq \left[\sum_{j=1}^{p/2} |\tau_j^{(g)}|\right] \left[\sum_{j=1}^{p/2} |\tau_j^{(h)}|\right] \end{aligned} \tag{4.1}$$

$\Longrightarrow$ If lowest value that $\left[\sum_{j=1}^{p/2} |\tau_j^{(g)}|\right] \left[\sum_{j=1}^{p/2} |\tau_j^{(h)}|\right]$ can take is maximised, $\mathcal{S}_\rho = \left|\langle \boldsymbol{\tau}_g, \boldsymbol{\tau}_h \rangle\right|$ is maximised.

Now in a test/survey, item score in j-th item is non-negative i.e. $\tau_j^{(g)} \geq 0$, $\tau_j^{(h)} \geq 0 \; \forall j = 1, \ldots, p/2$, Thus, we rephrase the statement in the last paragraph to state that if lowest value that $\left[\sum_{j=1}^{p/2} \tau_j^{(g)}\right] \left[\sum_{j=1}^{p/2} \tau_j^{(h)}\right]$ can take, is maximised, $\mathcal{S}_\rho$ is maximised.

Now, by minimising $\mathcal{S} = \left| \sum_{j=1}^{p/2} \tau_j^{(g)} - \sum_{j=1}^{p/2} \tau_j^{(h)} \right|$, we are minimising

$$\left| \sum_{j=1}^{p/2} \tau_j^{(g)} - \sum_{j=1}^{p/2} \tau_j^{(h)} \right|^2 =$$

$$\left[\sum_{j=1}^{p/2} \tau_j^{(g)} \right]^2 + \left[\sum_{j=1}^{p/2} \tau_j^{(h)} \right]^2 - 2 \left[\sum_{j=1}^{p/2} \tau_j^{(g)} \right] \left[\sum_{j=1}^{p/2} \tau_j^{(h)} \right] =$$

$$T^2 + (T \pm \epsilon)^2 - 2 \left[\sum_{j=1}^{p/2} \tau_j^{(g)} \right] \left[\sum_{j=1}^{p/2} \tau_j^{(h)} \right], \tag{4.2}$$

where sum of item scores in g-th subtest is T, sum of item scores in h-th subtest is $T \pm \epsilon$, with $\epsilon \geq 0$ small. Minimisation of $\mathcal{S}$ minimises ϵ.

So Equation 4.2 states that minimising $\mathcal{S}$, for a given T, is equivalent to maximising $2 \left(\sum_{j=1}^{p/2} \tau_j^{(g)} \right) \left(\sum_{j=1}^{p/2} \tau_j^{(h)} \right)$.

But we recall that $\mathcal{S}_\rho$ is maximised if the smallest value that $2 \left(\sum_{j=1}^{p/2} \tau_j^{(g)} \right) \left(\sum_{j=1}^{p/2} \tau_j^{(h)} \right)$ attains, is maximised. But for a given T, this attained value is $(T(T - \epsilon))$ as computed at the minimum ϵ attained by minimising $\mathcal{S}$. Thus, $\mathcal{S}_\rho$ is maximised if $\mathcal{S}$ is minimised.

5 Comparing our splitting methods to existing number partitioning methods

As we said above, our methods may be considered reminiscent of differencing algorithms, such as [3], but unlike in these algorithms, we do not sequentially replace the two largest numbers in the current iteration, by their absolute difference. In our binary swaps, at every iteration, a new subtest-pair is generated. Importantly, the demand of our test-based application drives us to partitions (subtests) of equal sizes, though our splitting strategy that splits by minimising the absolute difference between the sums of item scores in the 2 subtests, can in principle work with an odd

number of positive integers that we partition – since we can sum scores of $p/2$ items in one subtest, as well as scores of $p/2 + 1$ items in another. (However, the splitting strategy that works by maximising the inner product of the item score vectors of the 2 subtests will not work, unless these score vectors are of the same dimensionality).

To illustrate this, we use the same fiduciary example list $\{8, 7, 6, 5, 4\}$ that [4] uses. We use maximal iteration number N_{iter}=1 for this illustration, and implement the splitting by minimising $\mathcal{S}$, as delineated in Algorithm 1 of SNCKWDC. Here the j-th generated subtest-pair, is due to the j-th swap. Our method will

- begin with one (g-th) seed-subtest: $\{8, 5, 4\}$, and the other (h-th) that is $\{7, 6\}$. This will lead to $\mathcal{S} = s_{seed}$=4 at the 0-th swap. Next we proceed to

- the 1-th swap, in which swapping of the elements in the first row leads to the g-th subtest $\{7, 5, 4\}$ and the h-th: $\{8, 6\}$, yielding proposed $\mathcal{S}$ value of 2 at the 1-th swap. As this is a reduction from the earlier iteration, we accept the subtest-pair proposed by this swap as our current subtest-pair; update the proposed $\mathcal{S}$ value to 2, and proceed to

- the 2-th swap, in which we swap elements in the 2nd row, yielding a proposed $\mathcal{S}$ value of 4, which being higher than the current value of 2 of $\mathcal{S}$, is rejected, i.e. the current subtest-pair is as they were at the end of the 1-th swap. So current g-th subtest: $\{7, 5, 4\}$; h-th: $\{8, 6\}$; current $\mathcal{S}$ value is 2. As the swap index is still less than number of rows (3) in our data, we proceed to

- the 3-th swap, in which, given the currently accepted subtests, we swap elements of the 3rd row yielding the proposed g-th subtest $\{7, 5\}$ and the proposed h-th: $\{8, 6, 4\}$ implying a proposed $\mathcal{S}$ value of 6 at the end of swapping elements of all possible rows. So the algorithm stops at the end of this 3-th swap. Again, the proposed $\mathcal{S}$ value of 6 is higher than the current $\mathcal{S}$ of 2. So we reject the proposed subtest-pairs achieved in this 3-th swap, and retain the subtest-pair that were current in the 2-th swap, namely, g-th subtest: $\{7, 5, 4\}$ and h-th: $\{8, 6\}$.

Thus, the perfect solution that yields an $\mathcal{S}$ of 0, is missed by our splitting algorithm Algorithm 1 of SNCKWDC that aims to split using minimisation of absolute difference between sums of item scores in the 2 subtests.

The differencing algorithm of [3] would also miss the prefect solution, resulting in the partitions of $\{7, 5, 4\}$ and $\{8, 6\}$. Greedy heuristics would yield $\{7, 6\}$ and $\{8, 5, 4\}$ (see [4]).

[4] suggests that beyond a size of 23 elements in the list, the problem gets easier. The idea of a "phase change" in the easiness of achieving a perfect partition ($\mathcal{S}$=0 for us) is discussed by [1]. In fact, in applications to real tests, it is situations with p $\sim$50 that are of relevance to us. Thus, given our choice of the splitting strategy, a salient feature of our method is that it is a fast method, the order of which depends on the number of items in the test and not the examinee number, (see Remark 1 of Supplementary Materials), thus allowing for fast splitting of the test score data and consequently, fast computation of reliability.

6 Application to simulated data

In order to validate our method of computing the classically defined reliability following dichotomisation of a test into parallel groups, we use our method to find the value of r_{rt} of 4 toy tests, the scores of which are simulated from chosen models, as described below. We simulate the 4 test data sets D_1, D_2, D_3, D_4 under 4 distinct model choices; the underlying standard model is that the score $X_i^{(j)}$ is obtained by the i-th examinee to the j-th item in a test, is a Bernoulli variate, i.e.

$$X_i^{(j)} \sim \text{Bernoulli}(p) \quad \text{implying}$$

$$\Pr(X_i^{(j)} = 1) = p, \;\; \Pr(X_i^{(j)} = 0) = 1 - p$$

where the probability $Pr(X_i^{(j)} = 1)$, of answering the j-th item correctly by the i-th examinee

- is held a constant p_i for the i-th examinee $\forall\, j = 1, \ldots, n$, with p_i sampled from a uniform distribution in [0,1], $\forall\, i = 1, \ldots, N$, in data D_1.

- is held a constant p_i for the i-th examinee $\forall\ j = 1, \ldots, n$, with p_i sampled from a Normal distribution $\mathcal{N}(0.5, 0.2)$, $\forall\, i = 1, \ldots, N$, in data D_3.

- is held a constant p_j for the j-th item for all examinees, with p_j sampled from a uniform distribution in [0,1], $\forall\, j = 1, \ldots, n$, in data D_2.

- is held a constant p_j for the j-th item for all examinees, with p_j sampled from a Normal distribution $\mathcal{N}(0.5, 0.2)$, $\forall\, j = 1, \ldots, n$, in data D_4.

We use n=50 and N=999 in our simulations.

Thus we realise that data sets D_1 and D_3 resemble test data in reality, with the ability of the i-th examinee represented by p_i. Our simulation models are restrictive though in the sense that variation with items is ignored. Such a test, if administered, will not be expected to have a low reliability. On the other hand, the data sets D_2 and D_4 are toy data sets that are utterly unlike real test data, in which the probability of correct response to a given item is a constant, irrespective of examinee ability, and examinee ability varies with item–equally for all examinees. A toy test data generated under such an unrealistic model would manifest low test variance and therefore low reliability. Given this background, we proceed to analyse these simulated tests with our method of computing r_{tt}.

We implement our method to compute reliabilities of all 4 data sets (using Equation 3.6 of Paper I). The results are given in the following table that showing results of using our method of dichotomisation of 4 simulated test data sets, $D_1, \ldots, D_4$, into 2 parallel sub-tests, g and h each, resulting in the computation of classically-defined reliability r_{tt} of the test

The reliabilities of tests with data D_1 and D_3 are as expected high, while the same for the infeasible tests D_2 and D_4 are low, as per prediction. We take these results as a form of validation our method of computing r_{tt} based on the splitting of a test.

In Figure 1, we present the histograms of the 2 sub-tests that result from splitting the realistic data D_3 as well as histograms of the 2 sub-tests that result from the splitting of the unrealistic test

Data set	Test variance	$\sum_{i=1}^{N} x_i^{(g)}$	$\sum_{i=1}^{N} x_i^{(h)}$	$\sum_{i=1}^{N} (x_i^{(g)})^2$	$\sum_{i=1}^{N} (x_i^{(g)})^2$	$\sum_{i=1}^{N} x_i^{(g)} x_i^{(h)}$	Reliability r_{tt}
D_1	222.89	12014	12013	202506	201523	198257	0.9662
D_2	7.96	10440	10439	113212	112843	109133	0.02050
D_3	110.06	12597	12597	188727	189465	183564	0.8994
D_4	10.86	12683	12684	166199	166520	161130	0.0361

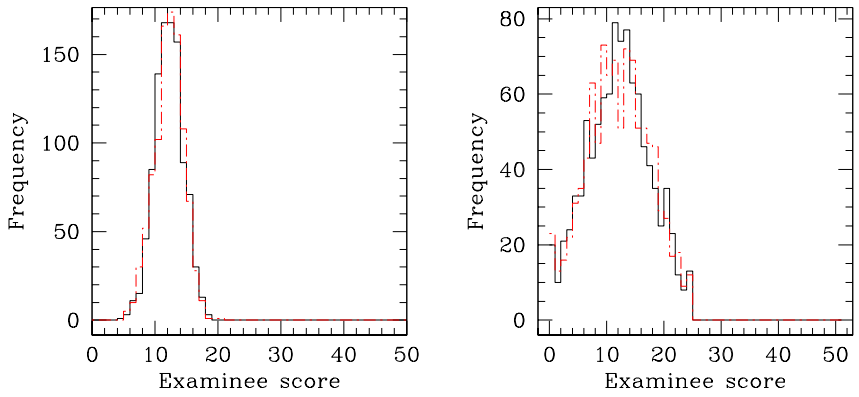

Figure 6.1: Figure showing histograms of the scores of 999 examinees in the 2 sub-tests (in solid and broken lines respectively), obtained by splitting the simulated test data set D_3 which has been generated under the choice that examinee ability is normally distributed (right panel). The left panel includes histograms of the 2 sub-tests that result from splitting the test data D_4 that was simulated using examinee ability as item-dependent, with probability for correct response to the j-th item given as a normal variate, unrealistically fixed for all examinees, $\forall\, j = 1, \ldots, 50$.

data D_4.

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.

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

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

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.