מרכז ארצי לבחינות והערכה (ע"ר)
NATIONAL INSTITUTE FOR TESTING & EVALUATION
المركز القطري للامتحانات والتّقييم
מיסודן של האוניברסיטאות בישראל

Estimating the intra-rater reliability of essay raters

Report
RR
16-05

April 2016

Yoav Cohen

NITE REPORT RR-16-05

ISBN:978-965-502-197-4

Estimating the intra-rater reliability of essay raters

Yoav Cohen

April 2016

This paper benefited from fruitful discussions with and comments from A. Allalouf, R. Fortus, N. Gafni and T. Kennet-Cohen. I thank them all.

Table of Contents

List of Figures

List of Tables

Abstract

The intra-rater reliability in rating essays is usually indexed by the inter-rater correlation. We suggest an alternative method for estimating intra-rater reliability, in the framework of classical test theory, by using the dis-attenuation formula for inter-test correlations. The validity of the method is demonstrated by extensive simulations, and by applying it to an empirical dataset. It is recommended to use this estimation method whenever the emphasis is not on the average intra-reliability of a group of raters, but when the intra-rater reliability of a specific rater is of interest.

Introduction

The rating of essays written as a response to a given prompt is a complex cognitive task that encompasses many subtasks. Reading is of course the main task, but also: understanding and interpreting the written essay; relating to its assessment context; relating to its cultural context; constructing a theory-of-mind of the writer; conducting comparison processes – with other essays and other writers; and engaging in numerical estimation and decision processes. Each of the subtasks is a source of variability between raters, either due to genuine differences between the raters or to the error inherent in each of the subtasks.

Hence the great diversity among raters, even after they have undergone a long training period; a diversity that is reflected in the final numerical ratings. Raters differ in their leniency/strictness, in their tendency to use (or not) the full range of the rating scale, and in the consistency in which they rate the essays.

In the present paper I concentrate on the question of the consistency of ratings within each rater – known as intra-rater reliability. Ideally, intra-rater reliability is estimated by having the rater read and evaluate each paper more than once. In practice, however, this is seldom implemented, since the two readings of the same essay by the same rater are not independent. The approach taken in the present work is based on the conception of reliability in classical test theory. In the first section I present a short discussion of essays as parallel and equivalent tests, and then suggest a way to estimate the intra-rater reliability by basing it on the long-standing formula for dis-attenuating inter-test correlations. The section concludes with a discussion of the numerical relation between two ways for estimating intra-rater reliability.

In the second section of the paper some simulation studies are presented, showing the validity of the proposed approach, and its dependence on the assumptions of classical test theory. One of the requirements for accurate estimation of intra-rater reliability is unidimensionality, or homogeneity, of the rating process across different raters. The last part of this section discusses the problems that heterogeneity poses, and sketches a possible solution.

In the third section of the paper, the method of estimating intra-rater reliability is demonstrated by applying it to empirical data.

The paper concludes with a short discussion of the estimation method and its utility in the analysis of essay ratings and in the scoring of the essays.

Raters as parallel or equivalent forms

In high-stakes testing programs which include writing essays among the various tasks that are measured, a standard procedure is to have two or more raters read and evaluate each of the essays. Thus, each rater can be considered as a parallel form on which the examinee is measured.

The correlation between ratings given by any two raters who read the same group of essays is usually considered as one type of inter-rater reliability estimate (called *standardized interrater coefficient* by Brennan, 2001), very much like the parallel-form reliability which is calculated for multiple-choice (MC) test forms.

The following, however, is a discussion of intra-rater reliability. This type of reliability can be thought of as the parallel test reliability of two test forms. Note that in classical test theory, if two forms are considered genuinely parallel, i.e. the correlation between the true scores obtained on the forms is perfect, then their observed inter-correlation is a good estimate of their reliability. In the absence of additional information, this estimate is symmetric, meaning that the two parallel forms have the same reliability. In fact, when defining genuinely parallel forms the requirement is that the two forms have the same first and second moments for both the true scores and the observed scores. Hence, genuinely parallel forms have the same error variance, which implies that they also are equally reliable.

This equivalence of reliabilities, however, cannot be assumed when considering raters. Raters differ in the consistency, accuracy and precision with which they execute their work. A plausible assumption is that some raters are more reliable than others. The raters can therefore be considered as essentially tau-equivalent (Lord & Novick, 1968) and may differ in their mean rating and reliability, or even as congeneric forms (Joreskog, 1971), in which case they can also differ in terms of their true-score variance.

One way to examine the quality of raters is by looking at their correlations with other raters. Raters who are less consistent or reliable would produce lower correlations with their peers than would those who are more reliable. But note that, in analogy with the two-form situation, the estimate is symmetric; i.e., the same estimate of reliability is given to the two raters, even if one of them is, in fact, more reliable than the other. Similarly, the mean inter-rater correlation of a specific rater with her peers is an estimate both of the reliability of the specific rater and that of a hypothetical "mean rater" who represents the peers. Thus, the reliability coefficient of a non-reliable rater estimated in this way will be biased upwards, while the reliability of a 'good' rater will be biased downwards, exhibiting a regression to the mean. This kind of reliability estimate is useful for determining the ratio of the reliabilities of two raters, or it can be used to rank-order the raters for purpose of quality control, e.g. in order to exclude or replace the scores given by the lowest ranking raters, but it is not an accurate and direct estimate of the intra-rater reliabilities.

A direct estimate of intra-rater reliability

The issue of rater errors has been given excellent treatment in the framework of Generalizability Theory (Brennan, 2001), where the effects of multiple sources of rating errors are simultaneously investigated. In the present work we limit the investigation only to one source of measurement error, that which is caused by the inconsistency of each rater by him/herself. The error components of essay topic, genre or prompt, and the inconsistency of the examinee in producing responses, are not investigated in the present work. Thus, the study is limited to assessment designs in which a collection of essays, all in response to a single prompt, are rated by a group of raters. We use classical test theory to estimate the intra-rater reliability, by looking at the inter-rater correlations. Note, that the inter-rater correlations are insensitive to variation of scales among the raters as long as the raters use scales that are linearly related.

We accomplish the estimation by using the method for correcting for attenuation. As stated in almost every book dealing with classical test theory (e.g. Guilford, 1954 Eq. 14.35; Haertel, 2006, Eq. 42; and Lord & Novick, 1968, Eq. 3.9.6) the correlation between **true** scores on two measures can be estimated by dividing the observed correlation between these two measures by the square root of the product of the reliabilities of these two measures; following the notation and formulation of Lord and Novick (1968):

Equation 1

$$\rho(T_x , T_y) = \frac{\rho_{xy}}{\sqrt{\rho_{xx'}\rho_{yy'}}}$$

Where, T_x and T_y are the true scores of x and y, $\boldsymbol{\rho}$ is the correlation coefficient, ρ_{xy} is the correlation between the observed scores x and y, and $\rho_{xx'}$, $\rho_{yy'}$ are the reliabilities of x and y.

There are various ways in which a true score can be defined (cf. Lord & Novick, 1968). For the following discussion we will follow Haertel (2006) in defining a true score of an essay as "the rating dictated by the rubric", and thus assume that the correlation between the true ratings given by any two raters is perfect (i.e. $\rho\ (T_x,T_y) = 1.0$). Then Equation 1 can be rewritten as:

Equation 2

$$\rho_{xy} = \sqrt{\rho_{xx'}\rho_{yy'}}$$

In practice we observe only ρ_{xy} , so the equation involves two unknowns and cannot be solved. But if we look at three correlations: r_{12} , r_{13} and r_{23} (adopting from now on a notation of r's instead of $\boldsymbol{\rho}$'s) we can write a system of three equations with three unknowns (r_{xx} denotes the reliability of measure x, while r_{xy} denotes the correlation between x and y):

Equation 3

$$r_{12}^2 = r_{11}\, r_{22}\,, \quad r_{13}^2 = r_{11}\, r_{33} \quad \text{and} \quad r_{23}^2 = r_{22}\, r_{33}$$

The positive-valued solutions for the three unknowns are:

Equation 4

$$r_{11} = r_{12}\, r_{13} / r_{23}$$

$$r_{22} = r_{12}\, r_{23} / r_{13}$$

and $$r_{33} = r_{13}\, r_{23} / r_{12}\,,$$

Or, in general, for any $i \neq j \neq k$:

$$r_{ii} = r_{ij}\, r_{ik} / r_{jk}\,.$$

As mentioned already, we assumed here that the correlation between the true scores of two raters is perfect. We could also assume that each rater has a unique and specific component in his/her rating, akin to the specific factors in Guilford's theory of intelligence. But had we assumed a lower correlation *v*, among the raters, the estimates of the intra-rater reliabilities, spelled out in Equation 4, would have been higher. For example instead of $r_{11}=r_{12}\, r_{13} / r_{23}$ it would equal $r_{12}\, r_{13} / (v\, r_{23})$ where *v* is smaller than 1. So assuming that the true inter-correlation between raters is 1.0, leads to conservative estimates of the intra-rater reliability.

The next section discusses the relation between two types of intra-reliability estimates. The first type is based on the dis-attenuation formula as described above, termed "dis-attenuation formula", DAF for short; the second is the traditional reliability estimate – based on the "mean inter-rater correlation", or MIC for short. The relation between these two estimates is shown with two goals in mind: first, to demonstrate the bias inherent in MIC estimates and, second, to provide a quick and easy numerical formula for estimating DAF reliabilities on the basis of MIC estimates when the original inter-rater correlation matrix is not available.

Then, a section is presented which describes the simulations that were conducted in order to test the feasibility of the DAF solution for estimating the intra-rater reliabilities of a group of raters, and in order to evaluate the accuracy of the method in the presence of sampling errors.

We then move on to show the benefit of using DAF estimates for studying the proximity relations between raters, by analyzing dis-attenuated inter-correlation matrices.

Lastly, we demonstrate the application of the DAF estimation method to empirical data.

The numerical relation between MIC and DAF estimates

Given a set of *n* rater reliabilities $\{r_{11}, r_{22}, \ldots, r_{nn}\}$ and assuming that the true inter-rater correlations are 1.0, we can generate the *nXn* matrix of **observed** inter-rater correlations **C** with entries:

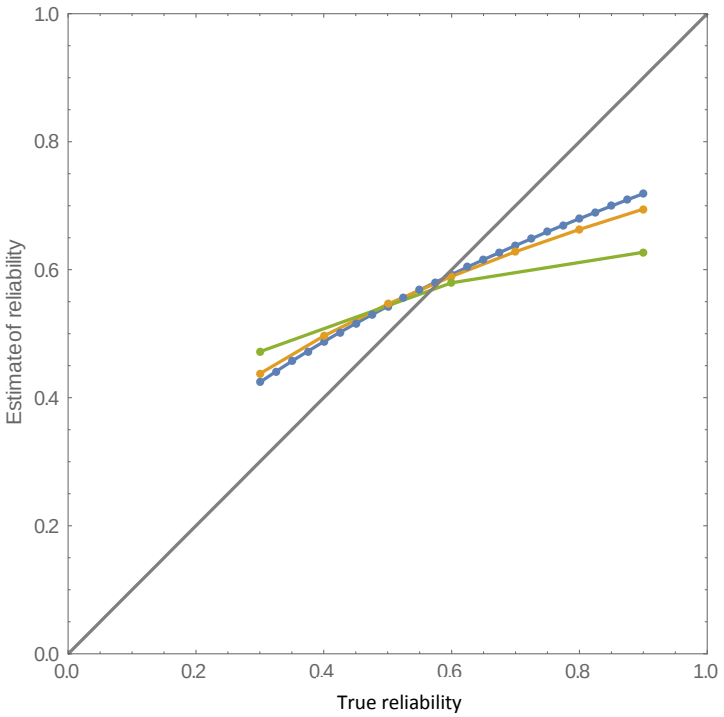
$$c_{ij} = \sqrt{r_{ii} r_{jj}}\,.$$

Note that the entries in the main diagonal of this matrix are the true reliabilities. The mean of the *i'th* row entries – excluding the diagonal entry – is an MIC estimate of the intra-rater reliability of rater *i*.

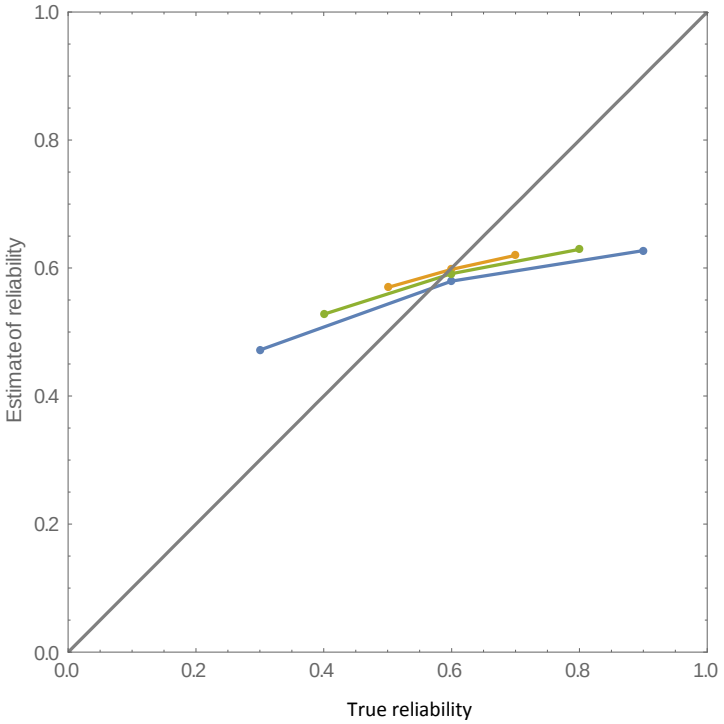
To study the relation between the MIC reliabilities and the true reliabilities, several sets of reliabilities are generated, sets that differ in the number and spread of the reliabilities. The reliabilities in each set are equally spaced, thus representing reliabilities from a uniform distribution. Each set of true reliabilities is a basis for generating a unique inter-rater correlations matrix, which in turn is used for the calculation of the corresponding MIC estimates.

The relation between the true reliabilities and the MIC estimated reliabilities are displayed in Figure 1, panels **a** and **b**. In panel **a** three sets of rater reliabilities that differ by size (number of raters) are presented. As can be clearly seen, more extreme estimates are also more biased. Reliabilities below the mean of the set are overestimated while those above the mean are underestimated. The bias is larger for smaller sets of reliabilities, but increasing the set size cannot eliminate the bias. In Panel **b** three sets of rater reliabilities (each based on three raters) that differ in their range are presented. As is shown, as the range of the true reliabilities becomes narrower, the MIC estimates get higher.

Note that the bias of the estimates is inherent to the definition of the MIC estimates; it is not a result of sampling error since the demonstrated relations are calculated as expectations of the true and the estimated reliabilities.



a. MIC estimates as a function of true reliabilities for three sets of true reliabilities that differ in their size. The true reliabilities are in the range of 0.3 to 0.9, reliability set-sizes are 3, 7 and 25. (The number of dots on each line corresponds to the set-size)



b. MIC estimates as a function of true reliabilities for three sets of true reliabilities that differ in their spread. There are three reliability values per set, with true reliabilities that are in the range of 0.3 to 0.9, 0.4 to 0.8 and 0.5 to 0.7.

Figure 1: MIC estimates as a function of true reliabilities

Estimating DAF reliabilities given MIC estimates

Suppose that there are *n* MIC estimates for a corresponding number of raters (where each MIC estimate is the mean correlation of a specific rater with the other raters) and the researcher is interested in estimating their DAF reliabilities. It is possible to calculate the DAF estimates. In the case of n= 3 raters:

Equation 5

$$d_1 = \frac{(m_1+m_2-m_3)(m_1-m_2+m_3)}{(-m_1+m_2+m_3)}$$

Where d_1 is the DAF estimate of rater number 1, and the *m*'s are the MIC estimates of the three raters. The estimates of d_2 and of d_3 would be identical except for a change of indices.

If n is greater than 3, then the algebra is not that simple, but a corresponding set of equations can be solved numerically. When *n*>3 the following formula can give acceptable estimates of the DAF reliability estimate (*d*) given the set of *n* MIC estimates.

Equation 6

$$d = 2.249\, m - 1.254\, mean + 0.049\, range$$

Where *mean* is the mean of the set of MIC estimates and *range* is the difference between the minimal and maximal MIC estimates in the set. This regression equation was found by simulation of a wide range of MIC sets, differing in location, dispersion and range. Using this prediction formula, the adjusted R^2 between the predicted DAF estimates and the true DAF reliabilities is 0.998. Ninety-six percent of the predicted DAF's fall within ±.01 of the true value.

Simulation studies[1]

The first question to investigate is whether the suggested procedure indeed recovers the intra-rater reliabilities of a group of 15 raters who jointly rated 500 essays. This assessment design deviates markedly from standard rating procedures that seldom employ more than two ratings per essay. The design, however, is not initially meant to simulate reality, but to investigate the soundness of the MIC and DAF estimates in noisy data.

The "true scores" of the 500 essays were generated from a standard (0 , 1) normal distribution. Fifteen raters were simulated, assuming different intra-rater reliabilities ranging from 0.55 to 0.97 in increments of .03. For each rater, a set of 500 ratings were generated by adding 500 "error" components to the set of 500 essay true scores. The error components

[1] All the calculations and simulations were conducted by using the Mathematica system, (Wolfram, 2015) version 10.3.

were generated from a normal distribution with a mean of 0.0 and a standard deviation denoted by se_i of:

Equation 7

$$se_i = \sqrt[2]{1/r_{ii} - 1}$$

Where r_{ii} is the intra-rater reliability of rater *i*.

Thus, we have a set of 500 essay true scores (in response to a single prompt), 15 reliabilities of the raters, and – for each rater – a set of 500 ratings. Note that the expectation of the ratings per rater is zero, and the expectation of the variance of the ratings of rater *i* is $1+ se_i^2$.

The 15X15 matrix of inter-rater correlations among the 15 raters is presented inTable 1, and is used to recover the original intra-rater reliabilities according to the following procedure: from the set of 15 raters we can create 455 triads (15! / (12! 3!)). Each triad is a combination of three different raters, where each rater appears in 91 (14*13/2) of them. So for each rater we can solve the system of equations (Equation 4) 91 times, and then average the results to get a DAF estimate (and the standard deviation) of the reliabilities.

Table 1: Inter-rater correlations. Each correlation is based on 500 ratings.

rater	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	1.0														
2	.55	1.0													
3	.58	.59	1.0												
4	.62	.58	.61	1.0											
5	.60	.63	.63	.70	1.0										
6	.64	.66	.64	.70	.70	1.0									
7	.64	.67	.65	.71	.71	.73	1.0								
8	.66	.65	.66	.71	.74	.73	.76	1.0							
9	.66	.66	.68	.71	.74	.76	.75	.78	1.0						
10	.69	.69	.70	.71	.75	.75	.79	.80	.81	1.0					
11	.69	.72	.68	.74	.78	.77	.77	.80	.83	.85	1.0				
12	.70	.70	.73	.76	.79	.79	.81	.82	.85	.87	.87	1.0			
13	.71	.71	.71	.77	.81	.80	.80	.84	.85	.87	.88	.90	1.0		
14	.71	.71	.74	.77	.82	.81	.84	.85	.86	.88	.89	.92	.92	1.0	
15	.73	.74	.75	.80	.83	.83	.84	.87	.88	.90	.92	.94	.94	.96	1.0

The true reliabilities and the statistics of the DAF estimates are listed in

Table 2 . The relation between true reliabilities and the DAF estimates is displayed inFigure 2. The statistics of the DAF estimates need some explanation. Each DAF estimate is based on 91 triads as was explained above. The statistics of the DAF estimates are based on these 91 triads: the mean of the 91 values, their median, standard deviation, minimum and maximum values, and lastly, the skewness of the distribution of the estimates.

Table 2: True reliabilities and DAF reliability statistics of 15 raters based on 500 ratings per rater

Rater #	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
True r_{ii}	.55	.58	.61	.64	.67	.70	.73	.76	.79	.82	.85	.88	.91	.94	.97
Mean DAF estimate	.56	.57	.58	.65	.70	.72	.74	.77	.79	.83	.85	.90	.91	.93	.97
Median	.56	.56	.58	.65	.70	.71	.73	.77	.79	.83	.85	.90	.91	.93	.98
Sd	.02	.02	.02	.02	.02	.02	.02	.02	.02	.02	.02	.02	.02	.02	.02
Min	.52	.51	.54	.59	.65	.68	.69	.73	.73	.77	.80	.85	.85	.88	.93
Max	.60	.63	.63	.73	.76	.79	.82	.82	.83	.88	.92	.93	.96	.98	1.03
Skewness	.14	.25	.68	.02	.20	.82	.82	-.02	-.65	-.53	0.35	-.37	-.33	-.43	-.29

The correlation between the true reliabilities and the DAF estimates is 0.995. The regression line and the 15 data points are plotted in Figure 2. The exact equation of the regression line is: -0.002 + 1.009 X, showing that the regression line almost coincides with the identity line. As is evident fromTable 2, the standard deviations of the 91 estimates per rater are 0.02 for all raters; the distributions of the 91 estimates are not symmetric; and there is slight skewness – for low-reliability raters the distributions are positively skewed, while for the high reliability raters they tend to be negatively skewed. The amount of skewness, however, is not marked, as can be seen by comparing the mean and median for each rater.

So the answer to the first question that was posed above is that the intra-rater reliabilities can be recovered successfully and accurately.

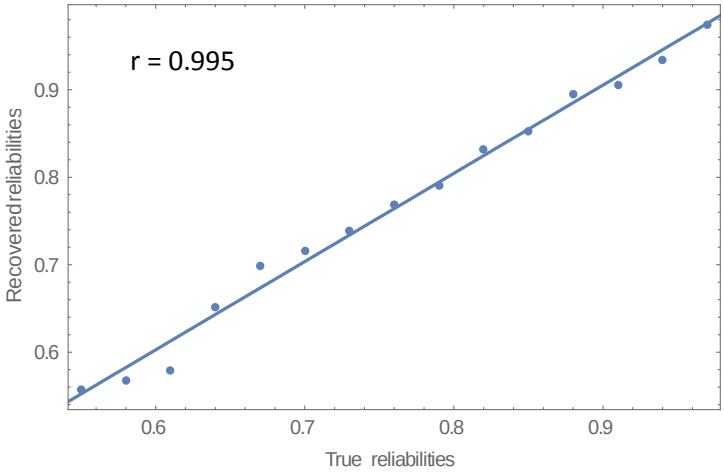


Figure 2: True reliabilities and DAF estimates

A second question is: what is the relation between the true reliabilities and the MIC estimates of reliability? These data are displayed in Table 3; although each MIC estimate is based on 14 values per rater, for purposes of comparison with the DAF estimate statistics, the same 91 triads of raters associated with each rater, are used to calculate local MIC estimates that are the average of the inter-correlations of each specific rater with the two other raters in the triad. These statistics are also presented inTable 3.

Table 3: **True reliabilities and MIC reliability statistics of 15 raters based on 500 ratings per rater**

Rater #	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
True r_{ii}	.55	.58	.61	.64	.67	.70	.73	.76	.79	.82	.85	.88	.91	.94	.97
MIC estimate	.66	.66	.67	.71	.73	.74	.75	.76	.77	.79	.80	.82	.82	.83	.85
Median	.66	.66	.67	.71	.73	.74	.75	.76	.77	.79	.80	.82	.82	.84	.85
Sd	.04	.04	.04	.04	.05	.04	.04	.05	.05	.05	.05	.05	.05	.05	.05
Min	.57	.56	.59	.59	.61	.64	.64	.66	.66	.69	.69	.70	.71	.71	.74
Max	.72	.73	.74	.78	.82	.82	.84	.86	.87	.89	.91	.93	.93	.94	.95
Skewness	-.29	-.33	-.07	-.39	-.23	-.13	-.15	-.14	-.11	-.03	-.02	-.08	-.09	-.13	-.15

The correlation between the MIC estimates and the true reliabilities is also very high (r=0.991), similar to the correlation between the true reliabilities and the DAF estimates of reliabilities, but the slope of the regression line is much shallower (slope=0.473), with overestimation of the low true reliabilities and underestimation of the high true reliabilities. The MIC estimates per rater are negatively skewed at all levels of true reliability, and the standard deviations of the estimates per rater are twice as large as the standard deviations of the DAF estimates. The MIC and DAF estimates of reliabilities as a function of the true reliabilities are displayed in Figure 3.

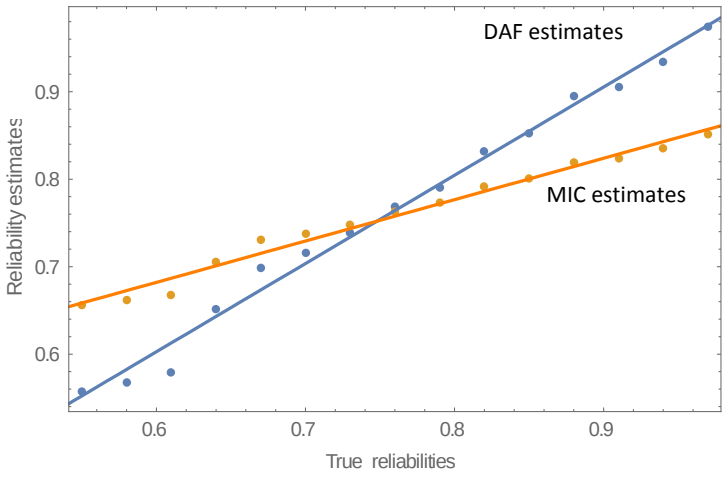


Figure 3: **MIC and DAF estimates as a function of true reliability**

Three comments about the estimates:

1. The average MIC estimates per rater that are calculated on the basis of the 91 triads is mathematically identical with the MIC estimate were it calculated as a simple mean of the correlations of each rater with its fourteen peers.
2. The DAF estimate of intra-rater reliability for rater *i* is *not* statistically independent of the estimate relating to rater *j*. For any triad of raters *i j* and *k*, the estimate of r_{ii} is based on r_{ij}, r_{ik} and on r_{jk}, and the same can be said for the estimates of r_{jj} and of r_{kk}; hence, all three estimates are derived from the same three values. A similar dependency is present in the MIC estimates. There is also a strong dependency

between the estimates given by the two methods that relate to the same rater. This is clearly seen in Figure 3, where the signs of the deviations from the regression line are similar for the two types of estimates.

3. Each of the DAF estimates is based on a summation of terms, where each term is a product and a ratio involving three correlations (e.g. $r_{11} = r_{12}\, r_{13} / r_{23}$). In contrast, the MIC estimate is based on summation of correlations. It is expected, therefore, that although the DAF estimates are more accurate, they involve larger standard errors. This point is investigated in the next section.

Standard errors of the estimates

In order to look at the expected accuracy of the two types of estimates and their sampling errors, the simulation that was described above was repeated 100 times with the same rater parameters. In each simulation a different set of essay true scores and a different array of measurement errors for each essayXrater combination were generated by the procedure that was described above.

The averaged results of the 100 replications are presented in Table 4 and Figure 4, together with their respective standard deviations, which are the estimates of the standard errors of the estimates.

Table 4: Mean reliability estimates and their standard errors over 100 replications (standard errors in parentheses).

Rater #	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
True r_{ii}	.55	.58	.61	.64	.67	.70	.73	.76	.79	.82	.85	.88	.91	.94	.97
DAF estimates	.55 (.03)	.58 (.03)	.61 (.02)	.64 (.03)	.67 (.02)	.70 (.02)	.73 (.02)	.76 (.02)	.79 (.02)	.82 (.01)	.85 (.01)	.88 (.01)	.91 (.01)	.94 (.01)	.97 (.00)
MIC estimates	.65 (.02)	.67 (.02)	.69 (.02)	.70 (.02)	.71 (.01)	.73 (.01)	.74 (.01)	.76 (.01)	.77 (.01)	.78 (.01)	.80 (.01)	.81 (.01)	.82 (.01)	.84 (.01)	.85 (.01)

Evidently, the average DAF estimates are very close to their true values. It is also evident that the reliability estimates of the more reliable raters are more reliable. The MIC estimates deviate from the true reliabilities. They overestimate the lower true intra-rater reliabilities and underestimate those at the high end of the range, as was already demonstrated. The standard errors of the MIC estimates are, as expected, smaller than the standard errors of the DAF estimates. (Note that although Figure 4 depicts symmetric standard errors of the estimates, the actual distribution is negatively skewed).

The results show clearly that the DAF estimates of the reliabilities are very accurate, although their standard errors tend to be slightly larger than those of the MIC estimates.

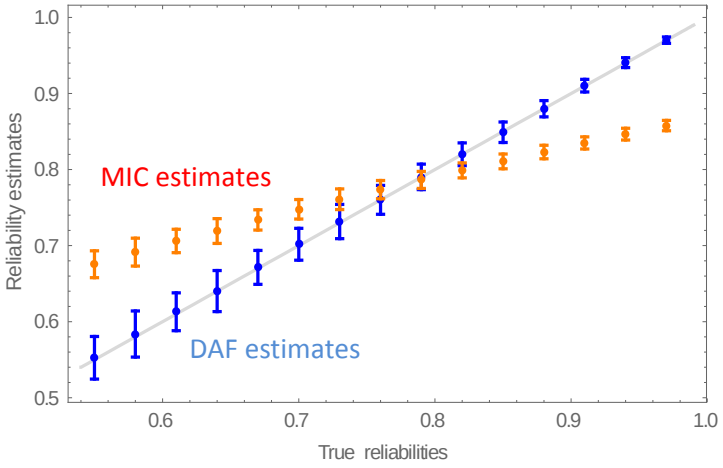


Figure 4: Mean reliability estimates and their standard errors over 100 replications. The straight line is the identity (x=y) line.

Correcting the inter-rater correlations

To check the consistency of the data, the recovered intra-rater reliabilities were used for dis-attenuating the observed inter-rater correlations. The model for generating the data assumed that the true inter-correlations between raters are perfect, i.e. r_{ij} =1.0, for all raters *i* and *j* (this assumption is relaxed in the next section). The dis-attenuated inter-rater correlations are shown in Table 5. There are of course deviations from 1.0 (henceforth: 'dis-attenuation errors'), and some of the dis-attenuated correlations are greater than 1.0, as can happen due to sampling error (Lord & Novick, 1968). The maximal absolute dis-attenuation error is .05, the mean error is 0.0002 and the standard deviation of the dis-attenuation errors (RMSE) is 0.015. Since the standard errors of the estimated intra-rater reliabilities are larger for low-reliability raters, as is evident in the data presented in Table 4, the dis-attenuation errors are, on average, larger for the low-reliability raters.

Table 5: Dis-attenuated inter-rater correlations.

rater	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1		0.98	1.02	1.03	0.96	1.02	0.99	1.01	0.99	1.01	1.01	0.99	1.00	0.99	0.99
2	0.98		1.04	0.95	0.99	1.03	1.04	0.99	0.98	1.01	1.03	0.99	1.00	0.98	1.00
3	1.02	1.04		0.99	0.99	1.00	0.99	1.00	1.00	1.01	0.97	1.01	0.98	1.01	0.99
4	1.03	0.95	0.99		1.04	1.02	1.02	1.00	0.99	0.97	0.99	1.00	1.00	0.99	1.00
5	0.96	0.99	0.99	1.04		0.98	0.99	1.01	0.99	0.99	1.01	1.00	1.02	1.01	1.01
6	1.02	1.03	1.00	1.02	0.98		1.01	0.98	1.01	0.97	0.99	0.99	1.00	0.99	0.99
7	0.99	1.04	0.99	1.02	0.99	1.01		1.01	0.98	1.01	0.97	1.00	0.98	1.01	0.99
8	1.01	0.99	1.00	1.00	1.01	0.98	1.01		1.00	1.00	0.99	0.99	1.00	1.00	1.00
9	0.99	0.98	1.00	0.99	0.99	1.01	0.98	1.00		1.00	1.01	1.01	1.01	1.00	1.00
10	1.01	1.01	1.01	0.97	0.99	0.97	1.01	1.00	1.00		1.01	1.00	1.00	1.00	1.00
11	1.01	1.03	0.97	0.99	1.01	0.99	0.97	0.99	1.01	1.01		1.00	1.00	1.00	1.00
12	0.99	0.99	1.01	1.00	1.00	0.99	1.00	0.99	1.01	1.00	1.00		1.00	1.00	1.00
13	1.00	1.00	0.98	1.00	1.02	1.00	0.98	1.00	1.01	1.00	1.00	1.00		1.00	1.00
14	0.99	0.98	1.01	0.98	1.01	0.99	1.01	1.00	1.00	1.00	1.00	1.00	1.00		1.01
15	0.99	1.00	0.99	1.00	1.01	0.99	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.01	

We showed that the DAF estimation of intra-rater reliability is quite accurate. This result is not dependent on the particular scale and the intra-rater reliability of the rater, as long as the scales are linearly related. But, as has been noted above, this is based on the assumption that all the raters are indeed evaluating, or measuring, the same construct. If some of the raters are using a similar but not identical construct (a different rating rubric, or a different interpretation of the rubric), then the assumption of perfect correlation between the true scores is not valid anymore. This can happen if, for example, in assessing essays by the holistic method, some raters relate more to the grammatical aspects of the essays, while others put more emphasis on the quality of discourse. Testing programs try to minimize this variability between raters by training and monitoring the rating process (cf. standards 6.8 and 6.9, Standards for Educational and Psychological Testing, AERA, APS, NCME, 2014), but – knowing what we know about the fallibility of human judgements – this kind of variability cannot be totally avoided. In such a case, if we assume, erroneously, that there is a perfect correlation between the true ratings when in actuality there is not, then some of the intra-rater reliabilities will be under-estimated, and the reconstructed inter-rater correlations will be over-estimated. This situation is discussed next.

Finding clusters of raters

If indeed there are sub-groups of raters who employ different constructs, then the observed correlations between raters within a subgroup will be on average) higher than the correlations between raters coming from different sub-groups. However, low inter-rater correlation can also be due to lower intra-rater reliabilities. We could assume that high inter-correlation within a subgroup of raters signifies a common rating construct, but it could also be a result of the raters all being very reliable, in spite of their using different rating constructs.

In essence what we have to do is to verify that the inter-rater correlation matrix is essentially unidimensional; that the only factor that contributes to the variability of the inter-rater correlations is the intra-rater reliabilities. There are various methods for identifying multi-dimensionality in correlation matrices (e.g. Budescu, Cohen & Ben-Simon, 1997; Hambleton & Rovinelli, 1986; Hattie, 1985; Nandakumar, Yu, Li & Stout, 1998; Svetina & Levy, 2012), or, put differently – for identifying clusters of variables (in our case – raters). One way is to decompose the inter-rater correlation matrix by applying principal components analysis (PCA). Alternatively, we can use clustering methods on the vectors of correlations of each rater with other raters, and find the main clusters in the data. The vectors/raters that show greater similarity – in other words, that have the same pattern of correlations with other raters – probably come from the same subgroup.

Finding clusters of raters in a homogenous matrix

Applying PCA to the matrix of observed inter-rater correlations (Table 1) produces a scree plot with a strong first principal component (see Figure 5) on which the first 7 raters (low reliability raters) load positively, while the remaining 8 raters load negatively. However, PCA on the dis-attenuated inter-correlations (Table 5 with 1.0 in the main diagonal) shows a very shallow scree plot, with much lower variances of the components (see Figure 5). So, observed correlations, even from a homogenous matrix, i.e. a matrix based on data representing a single rating construct, produce a marked first PC, but this component captures the differences in the reliability of the raters. The first five principal components of the two PCA analyses are shown in panels *a* and *b* of Table 6.

Table 6: First five Principal Components of homogeneous matrices

a. Observed inter-rater correlation matrix

	Principal Component #				
Rater #	**1**	**2**	**3**	**4**	**5**
1	4.799	2.903	1.561	0.548	0.063
2	3.876	0.153	-3.314	0.564	0.946
3	3.895	-2.900	1.505	1.352	0.796
4	2.668	-0.411	0.320	-3.202	0.770
5	1.117	-0.782	-0.501	-0.202	-2.658
6	0.539	0.617	-0.135	0.989	-0.591
7	0.373	-0.253	-0.132	-0.614	-0.851
8	-0.522	-0.129	0.172	-0.096	-0.900
9	-0.369	0.344	0.381	-0.099	0.335
10	-1.620	0.024	0.098	0.288	0.580
11	-1.908	-0.074	-0.081	0.208	0.193
12	-2.557	0.183	0.066	0.042	0.018
13	-3.06	0.151	-0.117	0.119	0.315
14	-3.459	0.133	0.130	0.041	0.519
15	-3.773	0.039	0.048	0.061	0.465

b. Dis-attenuated inter-rater correlation matrix

	Principal Component #				
Rater #	**1**	**2**	**3**	**4**	**5**
1	2.835	3.100	1.566	0.458	1.574
2	1.582	1.272	-0.488	-1.724	0.314
3	-0.704	3.763	-0.069	0.958	-2.324
4	2.470	-1.796	0.160	-0.505	-2.133
5	-5.504	0.628	1.425	-1.475	-0.237
6	0.440	-2.036	4.105	1.207	0.406
7	-0.940	0.317	-0.565	1.915	0.568
8	-1.224	-0.595	-1.863	2.285	1.438
9	-0.356	-0.198	-0.084	-2.485	1.616
10	1.809	-0.316	-0.764	-1.109	0.036
11	-0.297	-1.259	0.053	0.267	-1.757
12	-0.686	-1.078	-0.944	0.052	1.110
13	0.748	-0.385	-0.478	0.380	0.393
14	0.567	-0.894	-1.018	0.068	-0.615
15	-0.741	-0.523	-1.037	-0.292	-0.391

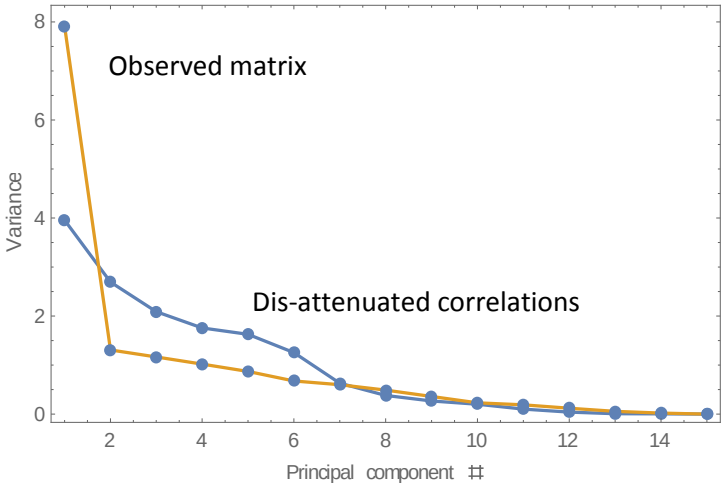
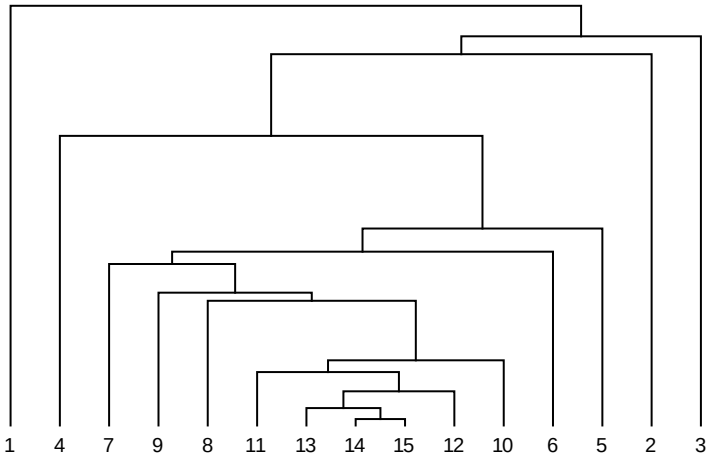


Figure 5: Scree plots of observed and dis-attenuated correlation matrices.

As for clustering, no significant clustering is found when applying hierarchical clustering on the same data, as can be seen in the dendrograms displayed in the two panels of Figure 6.

a. The observed inter-rater correlation matrix.



b. The dis-attenuated inter-rater correlation matrix.

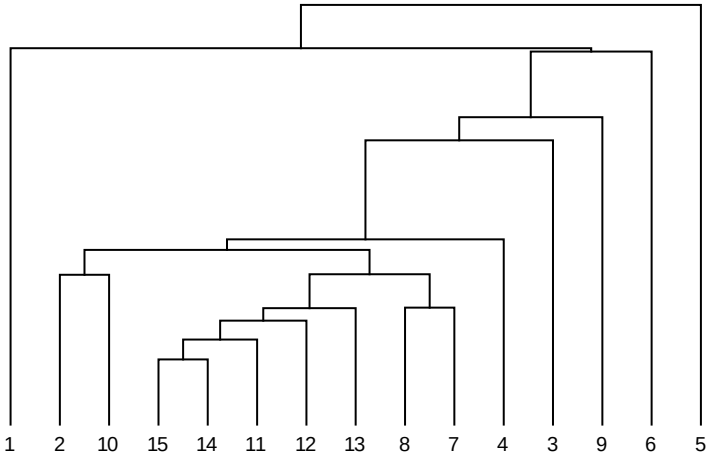


Figure 6: Dendograms of clustering

Finding clusters of raters in a heterogeneous matrix

To create a heterogeneous group of raters, i.e. a group that can be divided into subgroups, each one with a slightly different construct of rating, a dataset of 500 essays rated by 30 raters was generated. For each of the 500 essays, two true-scores – T_a and T_b – were simulated. T_a and T_b can be thought of as representing two different rating rubrics. The two true-score vectors are correlated; they were generated from a bi-normal standard distribution [{0,0},{1,1}], with a correlation of $\boldsymbol{\rho}$. The 30 raters were divided evenly between two groups. Those in group I had the same distribution of intra-rater reliabilities shown in

Table 2 and so did the raters in group II, but the ratings were based on T_a for group I and on T_b for group II. Thus, the expected value of the correlation between the true scores of

group I and the true scores of group II is ρ. The simulation was replicated twice: once with ρ=0.8 and then with ρ=0.9. These values of ρ (correlation between true scores!) are high, yet not perfect. The resulting data matrix represents ratings given by two groups of raters – half of whom rated according to their construct of what is a good essay, and half according to a different construct, and the expected correlation between the two constructs (which can be thought of as true concurrent validity), is 0.8 (or 0.9).

Results for ρ=0.8. First, let us examine the estimates of the reliabilities in the heterogeneous matrix, presented in Table 7. All the estimates fall short of the true values of the reliabilities, more so for higher true reliabilities and less so for lower ones. While the mean of the true reliabilities is 0.76, that of the DAF estimates is 0.67. This is the result of violating the assumption that the true inter-rater correlation is 1.0. This can be also verified by studying the matrix (not shown here) of dis-attenuated inter-rater correlations; while in a homogenous matrix the mean dis-attenuated correlation hovers closely around 1.0, here it is 0.98 (with a standard deviation of 0.13). The estimated reliabilities were used to create a matrix of dis-attenuated inter-correlations of the raters.

Table 7: The true and DAF values of 30 intra-rater reliabilities in a heterogeneous matrix. Raters 1-15 belong to group I and raters 16-30 belong to group II.

	Rater #														
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
True reliabilities	.55	.58	.61	.64	.67	.70	.73	.76	.79	.82	.85	.88	.91	.94	.97
DAF estimates	.53	.53	.52	.56	.54	.62	.63	.70	.72	.71	.74	.75	.83	.84	.86

	Rater #														
	16	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**
True reliabilities	.55	.58	.61	.64	.67	.70	.73	.76	.79	.82	.85	.88	.91	.94	.97
DAF estimates	.52	.46	.52	.57	.60	.63	.65	.65	.70	.74	.79	.80	.81	.84	.88

A PCA was applied to both the observed inter-correlation matrix and the dis-attenuated matrix. The first two principal components of the original matrix account together for about 80% of the variance, while the dis-attenuated correlation matrix shows a single principal component that accounts for more than 95% of the total variance. The first five PCs of the two matrices are listed in the two panels of Table 8. The full PC matrix is represented graphically in Figure 7.

The scree plots for the two analyses are displayed in Figure 8. In both analyses, there is one PC that distinguishes well between the two subgroups of raters. But, while it is the first and dominant PC in the dis-attenuated matrix, it is the second and less dominant PC of the observed matrix. The first PC of the observed matrix distinguishes between raters of high and low reliability irrespective of the group they come from. It follows that in the dis-attenuated matrix the differentiation between raters due to their differing reliabilities is much less pronounced, and this matrix is more suitable for differentiating between sub-groups of raters based on their conceptualization of what a 'good' essay is.

Table 8: First five Principal Components of correlation matrices, $\boldsymbol{\rho}$ =0.8.

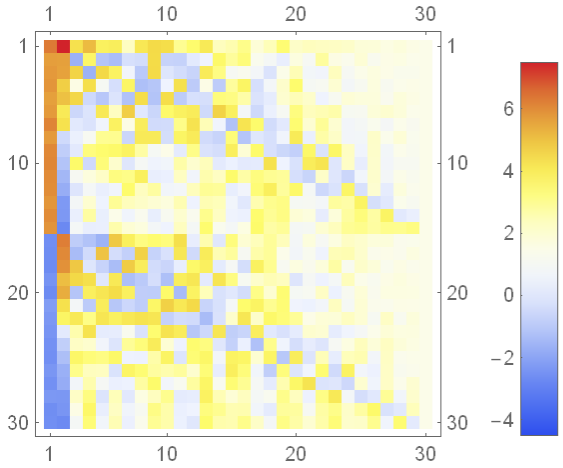
a. Observed inter-rater correlation matrix

	Principal Component #				
Rater #	**1**	**2**	**3**	**4**	**5**
1	9.010	4.089	0.871	2.671	0.581
2	4.512	2.397	-2.603	0.206	-1.664
3	4.263	2.419	1.640	-2.658	1.375
4	1.959	2.573	-0.051	1.091	-0.039
5	2.798	2.209	1.435	-0.693	-1.382
6	0.805	2.250	-0.785	-0.402	0.677
7	0.700	2.991	-0.279	-0.462	-0.114
8	-0.605	2.261	-0.137	-0.094	0.164
9	-1.322	2.896	-0.273	0.200	0.314
10	-1.938	2.799	0.357	-0.762	0.260
11	-2.597	2.674	-0.045	-0.032	0.048
12	-3.251	2.536	-0.318	-0.084	-0.170
13	-4.518	2.184	-0.055	0.012	-0.111
14	-4.727	2.487	-0.151	0.115	-0.158
15	-5.624	2.163	-0.043	-0.156	-0.141
16	5.877	-3.019	-1.008	-1.457	-2.472
17	5.330	-3.050	-1.903	-0.696	2.222
18	5.002	-3.153	2.155	-0.126	-0.894
19	2.769	-2.575	1.327	1.119	-0.171
20	3.512	-2.978	-1.583	0.576	0.874
21	1.216	-2.208	0.015	-1.080	0.501
22	0.513	-2.293	0.646	0.209	0.779
23	-0.199	-2.373	0.104	0.957	-0.158
24	-1.096	-2.619	0.084	0.052	-0.094
25	-1.877	-2.645	0.384	0.331	0.286
26	-2.815	-2.283	0.227	0.099	0.047
27	-3.082	-2.533	-0.048	0.124	-0.042
28	-4.501	-2.123	0.063	0.262	-0.319
29	-4.689	-2.572	0.0100	0.378	-0.133
30	-5.426	-2.502	-0.036	0.304	-0.064

b. Dis-attenuated inter-rater correlation matrix.

	Principal Component #				
Rater #	**1**	**2**	**3**	**4**	**5**
1	7.483	1.521	0.479	0.186	0.508
2	4.573	-0.854	-0.193	0.424	0.336
3	4.865	-0.558	-0.193	-0.470	-0.285
4	4.962	0.347	-0.169	-0.125	-1.128
5	4.378	-0.125	-0.471	0.296	-0.292
6	4.644	-0.173	-0.494	0.228	0.307
7	5.864	-0.402	0.396	0.135	-0.187
8	4.605	-0.280	0.502	-0.283	-0.448
9	5.614	-0.035	-0.352	-0.724	0.354
10	5.705	0.027	0.425	0.323	0.488
11	5.369	-0.259	0.265	0.031	-0.063
12	5.263	0.116	-0.194	-0.161	-0.05
13	4.748	0.059	0.000	-0.049	0.153
14	5.233	0.121	-0.057	-0.100	0.015
15	4.833	0.106	-0.056	0.075	0.119
16	-5.647	1.003	-0.184	-0.159	-0.288
17	-5.868	0.427	-0.645	-0.897	0.062
18	-6.022	-0.184	0.972	-0.914	0.372
19	-5.097	-0.210	-0.664	0.159	0.882
20	-5.864	-0.302	1.036	0.333	-0.194
21	-4.422	0.606	0.307	0.440	0.086
22	-4.773	0.261	-0.249	0.789	-0.312
23	-4.902	-0.698	0.153	-0.155	0.057
24	-5.238	0.048	-0.505	-0.009	-0.177
25	-5.286	-0.449	-0.201	0.082	-0.179
26	-4.774	-0.184	-0.109	0.543	0.045
27	-5.159	-0.019	0.038	0.147	-0.236
28	-4.608	-0.012	-0.017	0.093	0.024
29	-5.275	0.123	0.058	-0.174	-0.085
30	-5.205	-0.023	0.122	-0.065	0.114

a. Observed matrix PC's.



b. Dis-attenuated matrix PC's.

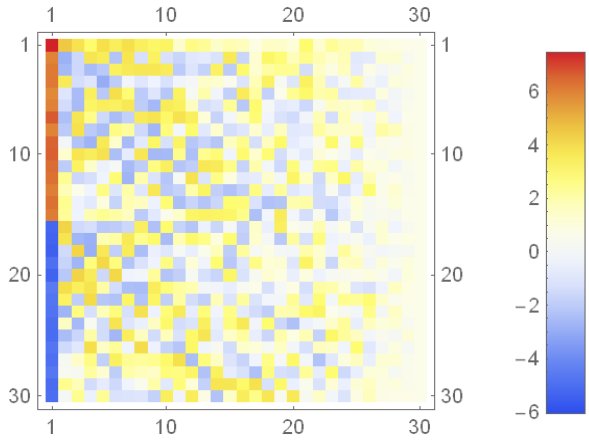


Figure 7: Graphical representation of the loadings on the principal components. $\boldsymbol{\rho}$ =0.8. Warm and cold colors designate positive and negative loadings respectively. The legend to the right of each matrix gives the color scale.

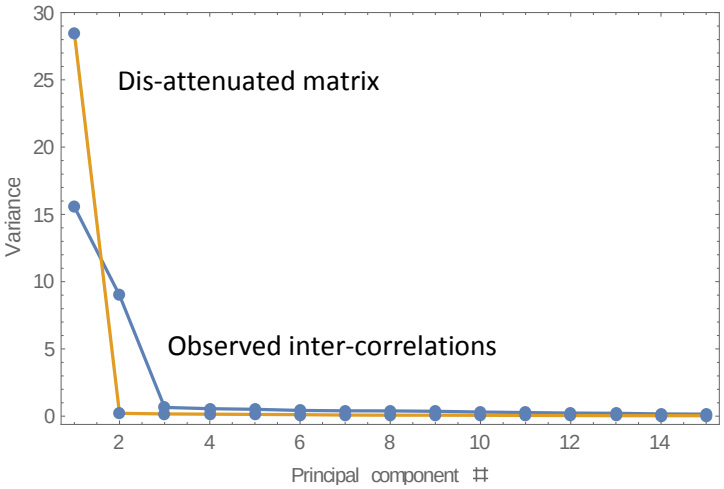
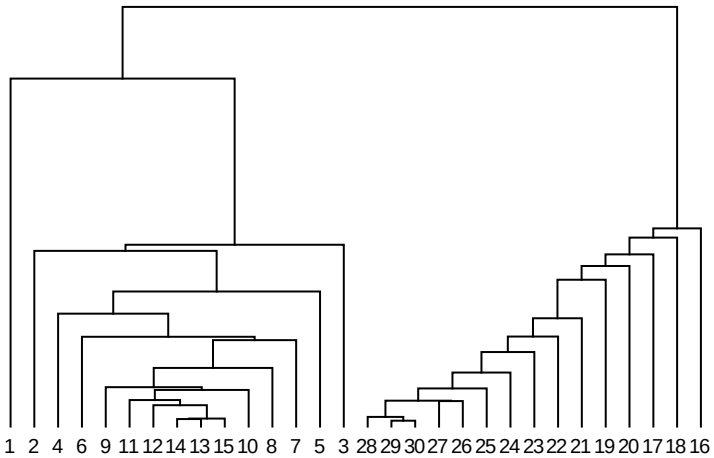
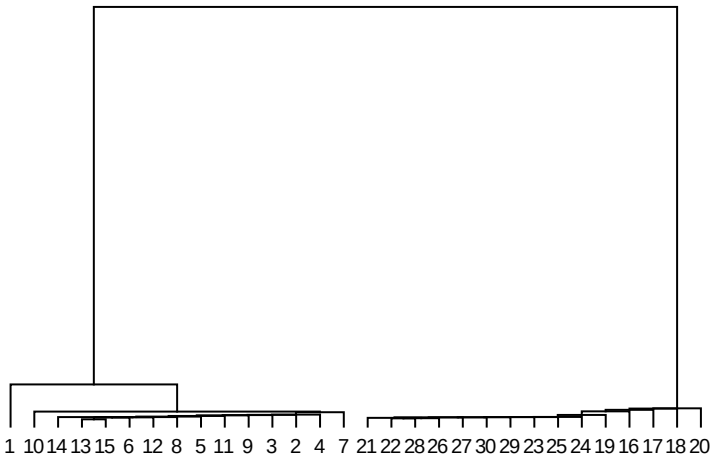


Figure 8: Scree plot of PCA on simulated data, the first 15 principal components, $\boldsymbol{\rho}$ =0.8.

Hierarchical clustering of the two matrices shows distinct differences between them. Clustering the rows of the observed inter-correlations matrix produces 2-4 clusters, with good a distinction between the two groups of raters but with the additional effect of the intra-rater reliabilities. In contrast, applying hierarchical clustering on the rows of the dis-attenuated correlation matrix produces two clearly distinct clusters. The dendrograms of the clustering of the two matrices are presented in Figure 9, which brings out graphically the sharp distinction between the two matrices.



a. Dendrogram of clustering the observed inter-rater correlation matrix.



a. Dendrogram of clustering the dis-attenuated inter-rater correlation matrix.

Figure 9: Dendrograms of two heterogeneous matrices, ρ =0.8.

Results for ρ=0.9. The same analysis was performed for 30 raters X 500 essays where the true scores of the two rater sub-groups correlated at a level of 0.9. Looking at the estimates of the reliabilities (Table 9), the estimates are systematically lower than the true values, but the bias is smaller than the case of a heterogeneous matrix with ρ =0.8. The mean DAF-estimated reliability is 0.71 (with a standard deviation of 0.13), compared with a mean true-reliability of 0.76.

Table 9: The true and DAF values of 30 intra-rater reliabilities, ρ =0.9.

	Rater #														
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
True r	.55	.58	.61	.64	.67	.70	.73	.76	.79	.82	.85	.88	.91	.94	.97
Estimated	.54	.55	.55	.55	.60	.58	.67	.73	.75	.76	.78	.80	.87	.88	.91
	Rater #														
	16	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**
True	.55	.58	.61	.64	.67	.70	.73	.76	.79	.82	.85	.88	.91	.94	.97
Estimated	.54	.48	.54	.59	.62	.66	.68	.68	.73	.77	.82	.83	.85	.88	.93

Looking at the observed inter-rater correlations, it is harder to identify the two groups of raters. This is because when ρ is high relative to the true intra-rater reliabilities, high observed inter-correlation occurs either because the two raters have high intra-rater

reliabilities or because they belong to the same group, and it is difficult to discriminate between these two situations.

The scree plots, the graphical representation of the PC loadings matrix and the dendrograms are presented in Figure 10. The first five PCs are listed in Table 10.

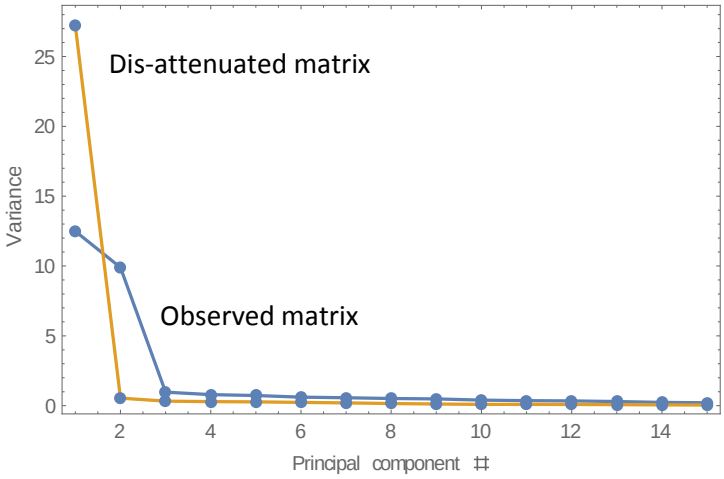
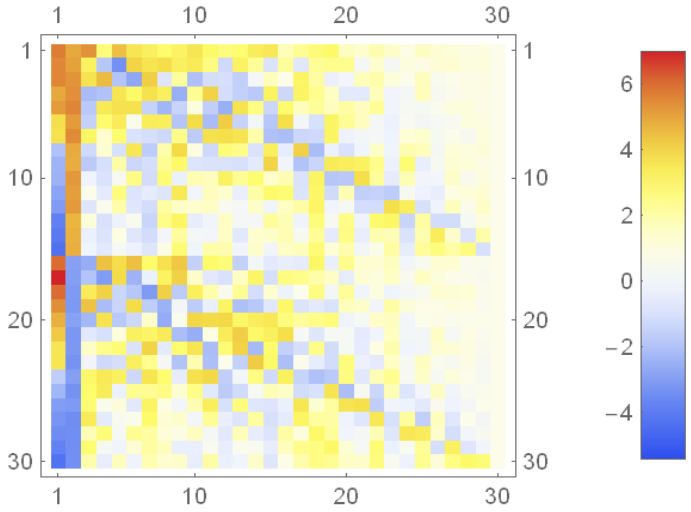
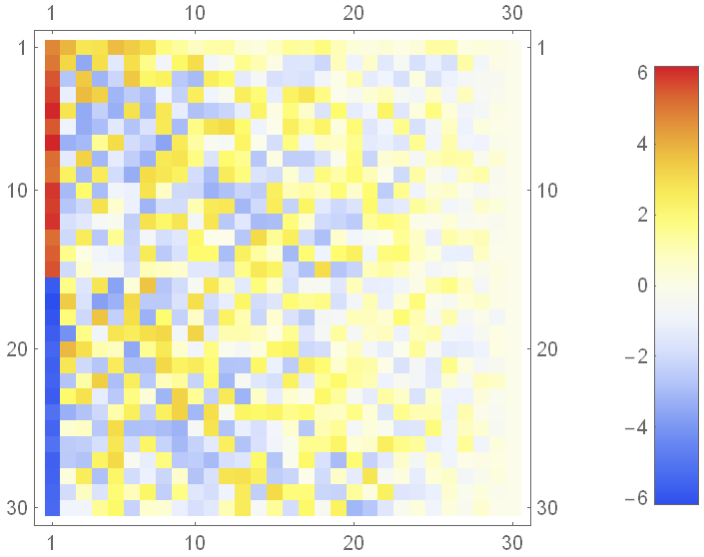


Figure 10: Scree plot of PCA on simulated data, the first 15 principal components, $\boldsymbol{\rho}$ =0.9

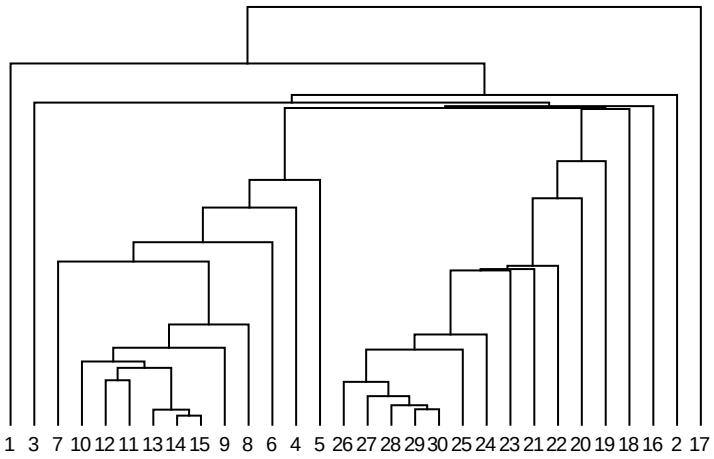


a. Observed matrix PC's.

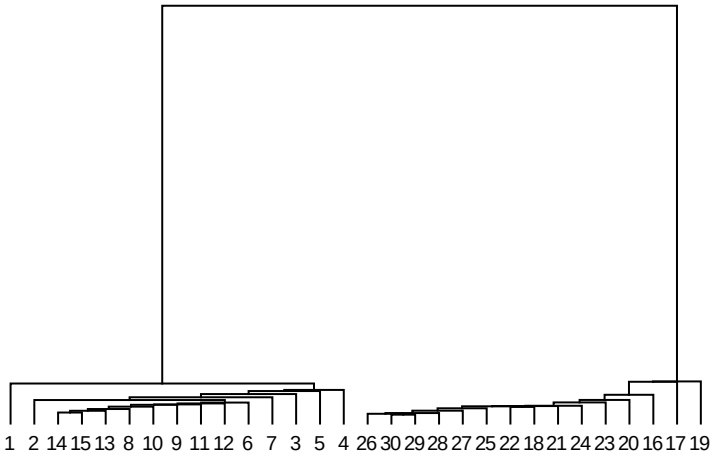


b. Dis-attenuated matrix PC's.

Figure 11: Graphical representation of the loadings on the principal components. $\boldsymbol{\rho}$ =0.9. Warm and cold colors designate positive and negative loadings respectively. The legend to the right of each matrix gives the color scale.



a. Dendrogram of clustering the observed inter-rater correlation matrix



b. Dendrogram of clustering the dis-attenuated inter-rater correlation matrix.

Figure 12: Dendrograms of two heterogeneous matrices, $\boldsymbol{\rho}$ =0.9

Re-estimation of reliability

Separation of the raters into two sub-groups paves the way to re-estimate the rater intra-reliability in each one separately. Applying the calculation to the two heterogeneous matrices (one with a $\boldsymbol{\rho}$ of 0.8 and the other with 0.9) gives the correct results as shown below. Each of the heterogeneous matrices was divided into two groups according to the clusters found, and then the corresponding inter-rater correlation matrix of each subgroup was used to re-estimate the reliabilities. The results are presented in Table 10, together with the true reliabilities. The means of the re-estimated reliabilities are 0.77 and 0.76 for the two sub-groups (with standard deviations of 0.14 and 0.13) for the $\boldsymbol{\rho}$ =0.8 matrix, and 0.76

and 0.75 (standard deviations of 0.14) for the $\boldsymbol{\rho}$ =0.9 matrix. This compares favorably with the true means of 0.76 (with a standard deviation of 0.13). These reliabilities can, therefore, be used to reconstruct the true inter-correlation matrix of all 30 raters.

Table 10: DAF-estimated reliabilities in each subgroup. Also shown are the true reliabilities.

a. $\boldsymbol{\rho}$ =0.8

	Group I - Raters 1-15														
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
True r	.55	.58	.61	.64	.67	.70	.73	.76	.79	.82	.85	.88	.91	.94	.97
DAF estimates	.52	.61	.61	.69	.65	.72	.75	.77	.82	.84	.86	.88	.92	.94	.97

	Group II – Raters 16-30														
	16	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**
True r	.55	.58	.61	.64	.67	.70	.73	.76	.79	.82	.85	.88	.91	.94	.97
DAF Estimates	.58	.60	.61	.66	.65	.69	.72	.75	.79	.83	.85	.87	.91	.94	.97

b. $\boldsymbol{\rho}$ =0.9

	Group I - Raters 1-15														
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
True r	.55	.58	.61	.64	.67	.70	.73	.76	.79	.82	.85	.88	.91	.94	.97
DAF Estimates	.56	.57	.59	.64	.63	.70	.73	.76	.79	.82	.85	.87	.91	.93	.98

	Group II – Raters 16-30														
	16	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**
True r	.55	.58	.61	.64	.67	.70	.73	.76	.79	.82	.85	.88	.91	.94	.97
DAF Estimates	.58	.53	.59	.63	.65	.70	.72	.73	.76	.81	.85	.89	.91	.93	.98

Estimation of reliabilities with empirical data

The method for estimating true reliabilities was applied to an empirical dataset[2]. The data consisted of ratings of 250 essays conducted by 13 well-trained raters. The raters were instructed to rate the essays on two scales of 1-6. The ratings given on the two scales by each rater to each essay were added up, thus creating ratings on a scale from 2 to 12. The mean essay rating of all 250 essays is 6.87, with a standard deviation of 1.42.

The mean rating for each rater and the corresponding standard deviation are presented in Table 11. Also listed for each rater is the MIC estimate – the mean inter-rater correlation with the other 12 raters. The mean of the mean ratings is 6.87, quite close to the middle of the rating scale (7), and the standard deviation of the means is 0.55. The most severe rater is rater #6 and the most lenient is rater #12. The raters differ also in the spread of ratings. While rater #12 uses a wide range (his/her ratings have a standard deviation of 2.4), rater # 4 has the narrowest spread of ratings (a standard deviation of 1.29, which is almost half of the standard deviation associated with rater # 12). In many treatments of rating data, the authors assume that raters differ only on the severity/leniency dimension (e.g. Brennan, 2001; Wright & Masters, 1982; and Longford, 1994). The present data challenge this assumption.

Table 11: Descriptive statistics and MIC estimates of the raters

	Rater #												
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**`10**	**11**	**12**	13
Mean	6.24	6.99	7.50	7.09	6.30	5.80	6.99	6.74	6.43	7.38	7.42	7.54	6.94
SD	2.02	1.72	1.76	1.29	2.10	2.27	1.81	1.91	1.87	2.31	1.98	2.40	1.85
MIC estimate	.58	.57	.48	.52	.51	.59	.55	.45	.52	54	.58	.51	.61

The inter-rater correlations of the empirical dataset are presented inTable 12. The correlations are quite modest, ranging from 0.33 to 0.72, with a mean of 0.54 and a standard deviation of 0.08.

[2] Cohen Y. & Allalouf A. *(2016) Scoring of essays by multiple raters; procedure and descriptive statistics.* Technical report TR-16-02. Jerusalem: NITE.

Table 12: Raters' intercorrelations

Rater #	1	2	3	4	5	6	7	8	9	10	11	12	13
1	1.0												
2	.66	1.0											
3	.56	.46	1.0										
4	.54	.56	.51	1.0									
5	.54	.56	.43	.54	1.0								
6	.63	.62	.47	.56	.56	1.0							
7	.59	.59	.47	.51	.53	.67	1.0						
8	.41	.50	.47	.38	.33	.46	.47	1.0					
9	.52	.56	.47	.53	.50	.53	.51	.48	1.0				
10	.56	.60	.48	.47	.48	.60	.56	.57	.49	1.0			
11	.64	.60	.45	.57	.56	.67	.61	.48	.56	.58	1.0		
12	.60	.47	.48	.56	.48	.58	.50	.36	.49	.43	.61	1.0	
13	.65	.67	.50	.55	.60	.72	.63	.48	.58	.64	.69	.56	1.0

Is the correlation matrix unidimensional? Cluster analysis does not reveal any distinct clusters of raters. On the other hand, PCA reveals at least two PC's before the knee-point in the scree plot, suggesting that the assumption of unidimensionality (true inter-correlation of 1.0 among all raters) cannot be held. Nevertheless, the DAF estimates of the reliabilities were estimated, and are presented in Table 13 together with the MIC reliabilities.

The fact that each essay was rated by 13 raters can be used to estimate the reliabilities in yet another way. We can get an estimate of the true score of each essay by averaging all the ratings given to that essay. This is not really a true score, because it involves only a finite number of raters, but it is a close enough estimate. The correlations of the actual ratings with this estimate of true scores are the basis of estimating the intra-rater reliabilities. A detailed description of this estimate is given in appendix A. The intra-rater reliabilities which are based on this method are very close to the DAF estimated reliabilities, and thus give strong support to the DAF estimates. These estimates are shown in Table 13 and in Figure 13 as "r_{it} estimates".

Table 13: Reliability estimates for 13 raters

	Rater #												
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**`10**	**11**	**12**	**13**
DAF Estimated reliabilities	.63	.61	.42	.51	.47	.66	.57	.37	.50	.54	.65	.48	.70
MIC Estimated reliabilities	.58	.57	.48	.52	.51	.59	.55	.45	.52	.54	.58	.51	.61
r_{it} estimates	.62	.60	.41	.50	.46	.65	.57	.36	.49	.53	.64	.47	.70

The correlation between the DAF reliabilities and the MIC reliabilities is practically 1.0 (r=0.998), and, as was demonstrated in the simulation studies, the DAF estimates of reliabilities are much more spread out – their range is from .37 to .70 compared with a range of .45-.61 for the MIC estimates of reliability. The data of Table 13 are presented graphically in Figure 13 , where the raters are ordered by magnitude of the MIC estimates. The estimation of intra-rater reliabilities can be further improved by estimating the reliabilities within each cluster, but this requires research which is beyond the scope of this report.

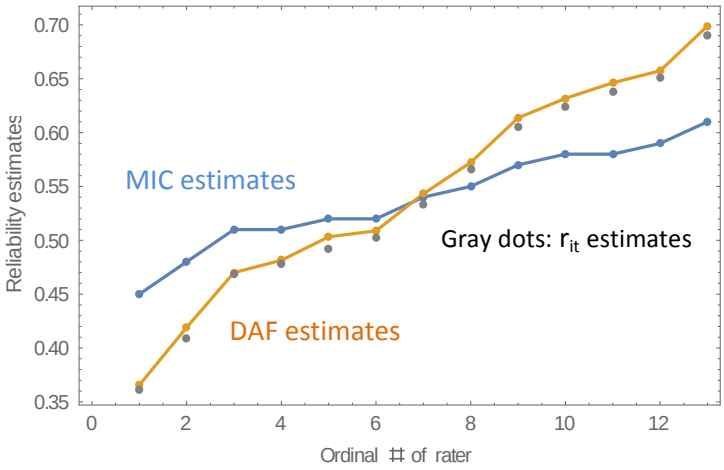


Figure 13: Reliability estimates for 13 raters. Note that the rater number is the ordinal number of the rater when sorting the raters by the magnitude of the reliability estimate.

Hierarchical clustering analysis of the observed inter-rater correlation matrix does not reveal any significant clustering of the raters. When the (DAF) dis-attenuated correlations are analyzed one rater stands out – rater number 8 – who apparently adopted a different scoring criteria or just did not do the task as required. Note that this rater would also be singled out by traditional means; he/she is the rater with lowest mean inter-rater correlation. In addition, the analysis reveals two distinct clusters, as can be seen in Figure 14. One cluster includes raters 1, 4, 5 and 12, and the other includes all the others except rater no. 8. However, there is still a large variation (large distances) within each cluster.

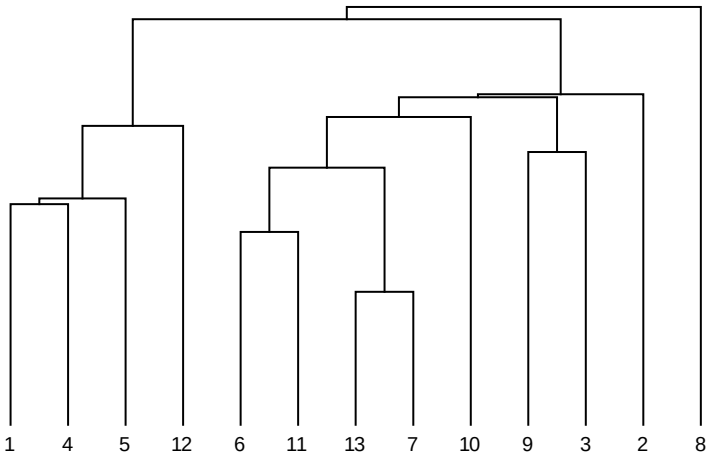


Figure 14: Dendrogram of the DAF dis-attenuated correlation matrix.

To sum up the analysis of the empirical data:

1. The DAF estimates have a greater spread than the MIC estimates, as was demonstrated in the simulations.
2. The DAF estimates are validated by using the correlation with an estimate of the vector of true scores.
3. Dis-attenuating the inter-rater correlations allows separation of the raters into clusters of raters, where the raters in each cluster are relatively homogenous in terms of the criteria that they adopt for marking the essays.
4. Since there is a high ordinal correlation between the MIC and DAF estimates, for the purpose of identifying low performing raters, the two kinds of estimates are interchangeable.

Discussion

The suggested method for estimating intra-rater reliability in the framework of classical test theory can be of use whenever there is interest in the reliability of a specific rater and not in the reliability of the raters as a group. One application is in setting quality standards for the performance of raters. We have shown that there is perfect ordinal correlation between the DAF and the MIC estimates. So, for identifying the least consistent rater or raters it does not matter whether we use one estimate or the other. But if we want to set numerical criteria for rater performance, an accurate estimate of the intra-rater reliability is required.

Accurate estimates of intra-rater reliability are also required in the context of calibrating raters. It is a known (and usually disregarded) fact that raters differ in the range of values that they use. While the leniency/severity of raters can be easily corrected by adjusting the means, when it comes to the variability in the spread of ratings, one has to decide what part

of the intra-rater variability is due to the spread of the (assumed) true scores, and what can be accounted for by the error component. Since the variance of the ratings within a rater is a sum of the error variance and the variance of the true ratings, simply equating the rating variance across raters may preserve the error component and even magnify it when applied to raters whose ratings have small variance. Adjustment of the variance of the raters has to take into account the true score variance and not the combined error + true score variance. The intra-rater reliability can be utilized in order to find the variance of true ratings per rater.

Having a good estimate of the true variance opens the way to differential weighting of raters. Differential weighting may not be acceptable in operational programs, but at least it can be used for research purposes. This point requires further research, as do the other points made here.

In the simulations detailed above, and in the demonstration of applying the procedure to real data, we based the analysis on full matrices of raters X essays. The reader may ask whether the same methods can be applied to sparse matrices, where essays are allocated randomly to pairs of raters. It should be noted that the method of DAF estimation is applicable whenever there are triads of raters who share pair-wise sets of essays. However, as we have demonstrated, multidimensionality of the ratings may pose a problem for the DAF estimates.

Lastly, it should be noted that the standard errors of DAF estimates are larger than those of the MIC estimates. This is because a MIC estimate involves the sum of correlations while the DAF estimate involves a product and a ratio of correlations. Therefore, the sample size in each situation has to be taken into account when deciding which method to use. This point also needs, of course, further study.

Summary

It is suggested that a novel way to estimate the inter-rater reliability be incorporated in studies of raters' behavior. The validity of the method was demonstrated via simulations and by investigation of an empirical dataset.

We have briefly pointed out certain areas in which the method can be of use, such as the calibration of raters and the differential weighting of raters. Some of the limitations of the method, namely, its dependence on the dimensionality of the data and on sample sizes, were noted.

As happens many a time, the solution to one problem – in this case the estimation of intra-rater reliability – opens a set of new questions. Further research will probably highlight the ways and the contexts in which the suggested method is most useful and applicable.

References

AERA, APS & NCME, 2014. Standards for Educational and Psychological Testing, AERA: Washington, DC.

Brennan, R. L. (2001). *Generalizability theory*. New York: Springer.

Budescu D.V., Cohen Y., Ben-Simon A. (1997) A revised Modified Parallel Analysis for the Construction of Unidimensional Item Pools. *Applied Psychological Measurement,* 21(3) 233-252.

Guilford, J. P. (1954). *Psychometric methods: By J.P. Guilford.2d Ed*. New York: McGraw-Hill.

Haertel, E. H. (2006) Reliability. Chapter 3 in: Brennan, R. L., (2006). *Educational measurement*. (4th edition) National Council on Measurement in Education & American Council on Education. Westport, CT: Praeger Publishers.

Hambleton K.H., Rovinelli R.J. (1986) Assessing the Dimensionality of a set of Test Items. *Applied Psychological Measurement,* 10 (3) 287-302.

Hattie, J. (1985) Methodology Review: Assessing Unidimensionality of Tests and Items. *Applied Psychological Measurement*, 9(2) 139-164.

Joreskog, K.G. (1971) Statistical analysis of sets of congeneric tests. *Psychometrika, 36*, 109-133.

Longford, N.T. (1994) Reliability of Essay Rating and Score Adjustment. *Journal of Educational and Behavioral Statistics*. 19 (3) 171-200.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. London: Addison-Wesley.

Nandakumar R., Yu F., Li H., Stout W. (1998) Assessing Dimensionality of Polytomous Data. *Applied Psychological Measurement,* 22(2) 99-115.

Svetina D., Levy R. (2012) An Overview of Software for Conducting Dimensionality Assessment in Multidimensional Models. *Applied Psychological Measurement,* 36(8) 659-669.

Wolfram Research, Inc., (2015) *Mathematica, Version 10.3*, Champaign, IL.

Wright B.D., Masters G.N. (1982) Rating Scale Analysis. Chicago: Mesa Press.

Appendix A

Estimating reliabilities using the correlation with true scores.

Given data of ratings by multiple raters of multiple essays, the squared correlation of the ratings given by a rater with the **true** scores of the essays is an estimate of the intra-rater reliability (cf. Lord & Novick, 1968, Eq. 9.1.1).

In practical situations we do not have information about the true scores, but a good estimate of the true score of an essay is the mean of the multiple ratings of the essay. The vector of these estimates is not perfectly reliable, but it approaches unity as the number of raters grows.

For the case of *n* raters with an average reliability of r_{ii}, the reliability of the true scores that are based on the average or sum of *n* ratings can be approximated by the generalized version of the Spearman-Brown formula with n and r_{ii}. Let us call it r_{tt}.

Let r_{it} be the correlation of the ratings given by rater *i* with the estimated true scores. When dis-attenuating this correlation by the reliability of rater *i* and the reliability of the estimated true scores, in the case of a unidimensional inter-rater correlation matrix, we should get a perfect correlation. Hence:

$$\frac{r_{it}}{\sqrt{r_{ii} r_{tt}}} = 1.0$$

It follows that:

$$r_{ii} = \frac{r_{it}^2}{r_{tt}}$$

In the application of this formula as described in this paper, each of the 13 raters was examined separately. The approximation of true scores for the estimation of the intra-rater reliability was based on the ratings given by the other 12 raters. The values 0.54 and 12 were used for estimating r_{tt} by the Spearman-Brown formula.