The Technical Whitepaper Series

9 REV.1

Test reliability and validity

The inappropriate use of the Pearson and other variance ratio coefficients for indexing reliability and validity

pbarrett.net
Advanced Projects R&D

Revised 15th July, 2010

Executive Summary

1. For test-retest reliability and validity estimation, psychologists generally use Pearson correlations to express the magnitude of relationships between attributes. For rater reliability where ratings are usually acquired using Likert ordered-class-as-numbered- magnitudes scales, they generally use intraclass (ICC) coefficients and rwg statistics.

2. I initially explore the three main questions asked by anyone working with tests, whether researchers, I/O test publisher psychologists, or clients/consumers of test products who are relying upon statements made by the sellers of tests.

3. I then show why the use of the Pearson or ICC coefficients alone are inappropriate in the context of the three common questions, using logical argument, graphics, and data analysis.

4. A solution to the dilemma posed by #3 is then constructed, introducing the Gower and Double-Scaled Euclidean indices of agreement as obvious choices for use in assessing test and rating reliability, test validity, and predictive accuracy. The final recommendation made is for the Gower coefficient, because of its more direct and obvious interpretation relative to the observation metrics. The Gower is interpreted as the average % of maximum agreement (identity) between the two sets of observations.

5. I compute both agreement and monotonicity, using the Gower and Pearson correlation coefficient respectively; the Pearson being the optimal measure of symmetric scale-free monotonicity.

6. I also develop the bootstrap procedure for assessing the statistical significance of the agreement index.

7. Example dataset analyses are provided to show how the agreement coefficient compares to conventional indices; these include the use of random samples of observations taken from bivariate-normal and uniform distributions.

8. Finally, the results from analyzing three real-world datasets are presented (two validity estimation applications and the examination of test sub-scale score relationship). In the case of the validity estimation applications, conventional validity r-squares of 19% (r = 0.44) and 5% (r = 0.23) can be compared to 90% and 87% agreement respectively using the Gower index. The reason for the somewhat spectacular increase in validity is provided in detailed sub-analyses associated with each example.

9. Three important theoretical developments and thinking have driven this work: Joel Michell's (1997, 2008) explanations of psychometrics as a pathology of science, Leo Breiman's (2001) arguments and results in favor of algorithmic statistics, and most recently, James Grice's development of Observation Oriented Modeling (book submitted for publication).

10. For test publishers, the opportunity now exists to cease producing the usual tables of mostly indifferent and "*not quite certain what they really mean*" validity indices, and instead take another look at their existing datasets which might harbor the kind of validities which need no creative spin nor ad-hoc "*in a perfect world*" statistical corrections.

1. Three Questions asked by Practitioners

When an assessment or observation is made of an individual which results in a quantitative value, a summed scale score, a rating, or an ordered or unordered category/class location, one or more of the questions below will likely be asked by the assessor:

[Reliability of the Assessment]: **If I obtain an assessment or rating of an individual on one occasion, and repeat the process on another, will the individual receive the same score on each occasion?**

There may be many reasons as to why the same score may not be achieved. But, the bottom-line for any user of any assessment is calculating the amount of error expected between one or more assessments made over time on the same individual.

Within psychometrics, this would be called test-retest reliability, differentiated from internal consistency reliability, and the standard error of measurement. These latter two methods are essentially "single-shot" methods of estimating reliability, which are redundant if reliability is estimated over time.

For example, let's assume we want to calculate the reliability of our new prototype automobile engine starter motor. We can approach the problem from a "single" observation in time viewpoint (as psychometricians do when using alpha or other "single-shot" estimates), or as a "time-to-failure" longitudinal exercise, where we repeatedly engage the starter motor to start the main engine (akin to test-retest in psychometrics).

The single-shot method requires that we use a reasonable number of starter motors with their main engines (having determined that all main engines are in working order). Now we simply start all the starter motors and observe how many fail to engage the main engine. That gives us a direct measure of the likely reliability of our starter motors on a single occasion - taking into account the number of starter motors we observed. But, it tells us nothing about what will happen over time, because we never observed what happens second time around. For all we know, they could all have burnt out their contacts as part of their initial use. Yet, this is the exact analog of how psychologists approach reliability. A one-shot exercise, not even using the same "stimulus" (the same-model starter motor), but "items" which might be "similar" to one another or ordered according to some assumed "latent trait". And from this one-shot assessment, they go on to make statements about how "reliable" a test, or a person, might be.

A test may have hopeless internal consistency (interpreted as poor reliability), but an individual might obtain the same score on several occasions by answering the same subset of items (excellent reliability by any other standard).

And that's my point. Reliability seems to involve the notion of elapsed time. It's how all other applied sciences and frankly the rest of the real-world uses the term reliability. E.g. "will it last?", "if I do this again, will the same thing happen?", "will this device do what it should do next time I switch it on?", "will my hard-drive maintain my stored data over time"?

However, test-theory reliability estimates avoid this "elapsed time" issue by treating items or people as "samples from some population or universe" - and attempting to infer the reliability of a test (or person) with reference to sampling distributions, hypothetical true scores, or data-model features. Estimates from such models are thus predicated upon assumptions about data rather than relying upon the actual data at hand.

Yet, what concerns practitioners above all others is not what "should happen in a perfect world" but what is likely to happen in the real world.

This is what the new procedures presented below are designed to provide. They work directly from the observations. No true scores, no restriction of range corrections, no hypothetical sampling distributions, no abstract variance ratios, and no data transformations (no standardization). What we observe is what we work with. The result? A clear statement of the amount of error likely to be incurred for an individual over a specified amount of time on a test, translated into the actual metric of the scores, ratings, ordered classes, or categories. But, data must be acquired from at least two occasions across the same individuals using the same instrument.

[Validity of the Assessment]: **Even if the assessment is reliable, does it assess what I think (or have been told) it assesses?**

An assessment may be reliable, but is it of any use? This is not an exercise in the academic semantics of the word "Validity", as might be found in scientific journal publications or books on "Validity". Practitioners use tests and/or make assessments for a purpose; to help them come to a decision about the likely future occurrence of some important outcome. They will have been told that the test or assessment "measures" one or more attributes, whose magnitudes or ordered classes are predictive of some particular kinds of outcome or event. As Bob Hogan has consistently stated, repeated in a recent book chapter on Personality Assessment (Hogan and Kaiser, in press) …

> "The goal of assessment is not to measure entities, but to predict outcomes; the former only matters if it enhances the latter".

That's what really matters to practitioners and users of tests - predictive accuracy. When an employee or union representative questions the interpretation of test scores from an assessment in an employment court, what the court wants to see is evidence that the assessment does indeed predict that which is considered important for job performance or some other outcome for which the assessment has been used as a decision-making tool.

Invariably, this evidence is given in the form of Pearson correlations between test scores and criterion outcomes. Test manuals are usually packed with such correlational evidence. Sometimes (very rarely), the evidence may take the form of actuarial or classification tables, or ROC curves, with a direct estimate of misclassification and error-rates. But, the overriding index for indicating validity is the Pearson correlation coefficient, which may subsequently be corrected "upwards" to correct for restriction of range or the reliability of the variables in question. In many cases, meta-analyses with "corrections" are used to aggregate results from several small studies (with the accompanying problems and confusions

that can cause if non-identical criteria are aggregated as though they were identical (see Barrett and Rolland, 2009).

However, as will be demonstrated below, using real-world data and a new class of coefficient which uses the actual data at hand rather than a transformed version of it, the Pearson correlation can be seen to severely misrepresent the actual validity of data - in terms of how well test scores/assessments predict important criterion outcomes. Baguley (2009, 2010) had already published some warnings about the use of the Pearson coefficient, but still persisted with trying to devise ways of reporting "agreement" in terms of conventional effect sizes.

Reporting effect sizes is clearly better than reporting p-levels of significance, but effect sizes remain needlessly abstract and frankly confusing to many users of tests who need to know how accurate (or inaccurate) is the assessment in terms of its prediction of important outcomes*, in the metric of those outcomes*.

Using the new methods presented below, a much clearer and more accurate picture of the validity of an assessment can be given to any 3rd party, in a form that is easily understood. The methods use the actual data at hand, make no assumptions about hypothetical sampling distributions, hypothetical true scores, are immune to restriction of range or reliability attenuation, and focus on direct observation agreement and directionality of relationship (and determining monotonicity separate from agreement).

3 [Rater Reliability]: **If I ask raters to rate objects, people, etc., how similarly do they rate them? More specifically, how well do the raters' ratings agree with one another?**

A straightforward and essential question, asked by any practitioner who makes use of an assessment center with observers who rate behaviors, or where ratings are acquired about an individual from multiple raters within a 360-degree assessment, or where two or more supervisors are rating subordinate staff on performance measures, or where nurses are rating patient behavior on a ward, or forensic clinical psychologists or corrections staff are rating offenders based upon judgments made about their previous behaviors in case-records (for actuarial risk assessment procedures). The conventional approaches to assessing rater reliability using ordered-class as interval rating scales are intraclass correlations (ICCs) , Pearson correlations (rarely), Kendall's Concordance, some IRT-based methods (as in Rasch facet analysis), and the multiple-rater coefficients of the *rwg* variety (James, Demaree, & Wolf, 1993).

There are two major problems with current approaches:

1. Pearson ICCs, and rwg coefficients rely upon ratios of variances, and data which are distributed according to the normal distribution. However, organizational rating data tend to be highly skewed, truncated in range, and possesses little variance (because of halo effects or literally because there is not much variation in the behaviors being rated). That means all these coefficients are going to produce indices which are attenuated simply because of a lack of variance in the data. It's pointless trying to correct for this as the data are never normally distributed, and never real-valued continuous numbers (upon which many of the calculations depend for their accuracy).

2. Quantitative psychologists and psychometricians have forgotten what practitioners want to know; which is a simple answer to the simple question: "how well do the raters' ratings agree with one another?". Not how "monotonic" they are, how the observed to true-score ratios can be indexed, how the rater variance ratios may be re-expressed in terms similarity etc.

This logic treats ratings as "what you see is all you have got". Raters either agree with their ratings or they don't. Because two raters might use two different areas of a rating scale, but agree monotonically is irrelevant (e.g. on a 5-point rating scale, on three attributes, one rater rates an individual 3, 4, and 5, while another rates the same individual as 2, 3, and 4). Those ratings do not agree - period. There are no "faceted raters", or some latent variable etc. on which raters can be placed (as in IRT facet analysis). This is just psychometric tomfoolery. Ratings either agree with one another, or they don't. The magnitude levels/descriptions on a rating scale are meaningful as "absolutes". That is, "excellent means excellent", not "Average". Because one rater may rate using different but monotonically equivalent rating levels as another rater is of no interest except as an indication that although the ratings are unreliable (do not agree very well at all), the pattern of observations between raters are monotonically related. As to why, that is a matter for exploration, training, or whatever. The point being that assessing rating reliability is simple, straightforward, and uncomplicated.

And the new agreement procedures recommended below do just that; they answer the basic question by using the actual rating data, untransformed, and providing estimates of similarity which are in the metric of the ratings themselves. The methods use the actual data at hand, make no assumptions about hypothetical sampling distributions, hypothetical true scores, are immune to restriction of range or reliability attenuation, and focus on direct observation agreement and directionality of relationship (separating monotonicity from agreement). Technical Whitepaper #10 in this series presents the entire computational solution/algorithms and computer program for dealing with multiple raters and interrater reliability, but, the indices used are those presented here, along with another which adds a "certainty" component to aid critical judgments (the Kernel Smoothed Distance coefficient).

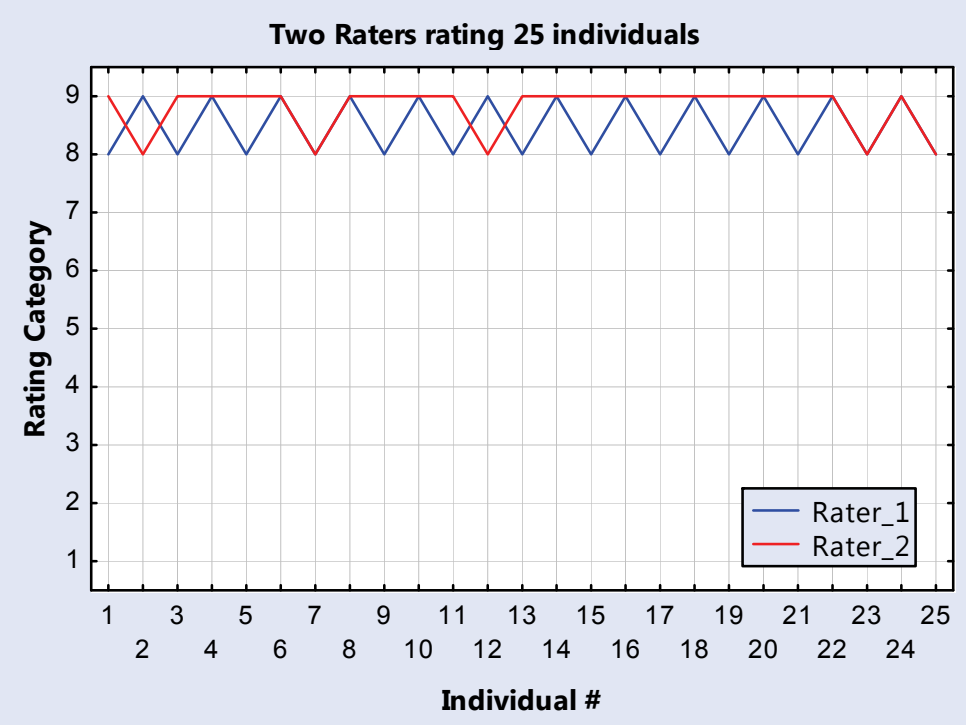
2. The Two Problems affecting Pearson and other Variance-Ratio Statistics: Monotonocity and Restricted Range

In order to show very clearly why conventional IRR and even test-retest coefficients are problematic, I generated data for two raters, using a typical 9-category rating scale as shown below. I've created category descriptors which provide very clear anchors.

It's useful to see such clearly defined category descriptors – as these enable the contrast to be seen quite clearly between what the categories mean, and what the indexes will tell us.

Rater Code	Attribute: Communication Style
9	Truly Excellent; able to convey complex data via useful metaphors, always willing to explain in simple terms, produces truly outstanding presentations.
8	Pretty near Excellent
7	Pretty good really – not outstanding, but efficient
6	Not bad – better than average but not too much
5	Average - ok sometimes, not others – about 50/50
4	Borderline acceptable – room for much improvement
3	Not really acceptable – inefficient and sloppy
2	Pretty near hopeless – many complaints
1	Absolutely hopeless; virtually incoherent, unintelligible, never able to explain complex concepts, presentations are just dreadful.

Now let me generate two very simple sets of data for our two raters who rate 25 individuals using this rating scale.



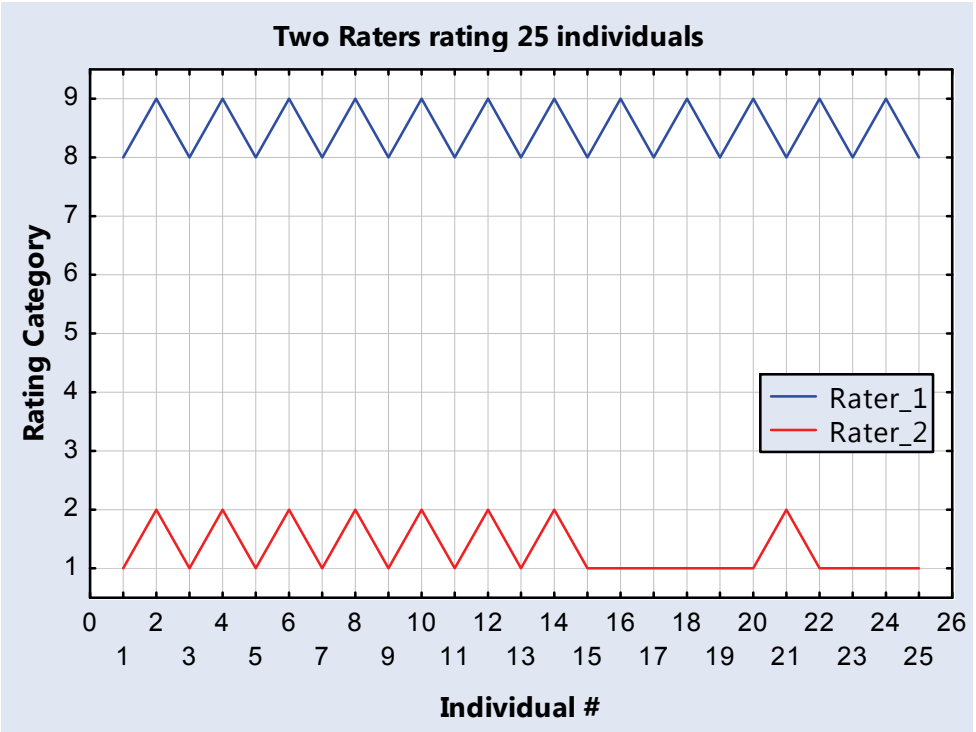
Here we see two raters, both assigning ratings between 8 and 9 on our scale, for 25 individuals. Clearly, these ratings are very similar to one another in terms of what the rating categories 8 & 9 mean.

But the four coefficients (Pearson and ICC models 1, 2, and 3) all tell us that rater reliability is virtually zero. The reality is, rater agreement is very high, while rater correlation is near zero. We computed the $r_{wg(j)}$ statistic, treating rated cases as objects/items. That tells us the ratings are identical; they are not.

You cannot correct these data - because there is no correction for the Pearson's reliance upon using transformed data instead of actual observations (it standardizes the actual observations, then computes the agreement of these transformed observations, not the ones you actually observed). And, there is no correction for "attenuation" for restriction of range as there is no known population variation for these data. What you see "is it".

In many cases, I have seen 360-degree rating data on a 5-point scale which mostly varies between 4 and 5, with the odd "3" thrown in occasionally.

Is it any wonder supervisor ratings of job performance using ICCs are always so low? This is due to a methodological flaw, not an empirical fact; as can be seen in my ISSID 2009 presentation "**Interrater Reliability: measuring agreement and nothing else**" (http://www.pbarrett.net/issid/issid2009.html) and Technical Whitepaper #10, where real-world rating data was used to show the difference between the typical rating coefficients and their counterparts using the new coefficients presented below.



Here we see two raters, both assigning ratings at each end of our rating scale, for 25 individuals. For the first 14 individuals being rated, the monotonicity between assigned ratings is perfect.

But, these ratings across the 25 individuals indicate almost total disagreement, **when you take the meaning of those ratings into account**. Here, only icc2 (the model 2 icc) produces a value which seems to get near to what we are seeing by eye with these data.

But, what if this pattern of data was produced by 50 raters providing a single rating on an attribute for each individual (a model 1 icc)? It's value is -0.97 – an indication of almost perfect negative correlation, but which says nothing about the actual disparity between ratings. The Pearson r is, as expected, quite misleading.

I again computed the $r_{wg(j)}$ statistic, treating rated cases as objects/items. That shows us the ratings are identical. Which they are not.

3. The Solution

Two agreement coefficients seem to do exactly what is required here: the **Gower** (1971) agreement index and a **double-scaled euclidean agreement index** (see Technical Whitepaper #6: Euclidean distance: raw, normalized, and double-scaled coefficients (2005), in which it was first derived); downloaded from:
http://www.pbarrett.net/techpapers/euclid.pdf

These coefficients have been used in marine biology research as well as in some profiling applications (Barrett, 2005; Wellenreuther, Barrett, and Clements, 2009, and within the current Hogan Development Survey and Hogan Cognitive Inventory test manuals).

The formula for the **Gower** index is:

$$Gower_{similarity} = 1 - \left[\frac{\sum_{i=1}^{n} \left(\frac{|case_{1i} - case_{2i}|}{range_i} \right)}{n} \right]$$

$n =$ the number of cases

$range_i =$ the range of the rating/score attribute for case i $(maximum - minimum\ value)$

$case_{1i} =$ the observed value for case i of n in the first set of observations

$case_{2i} =$ the observed value for case i of n in the second set of observations

Where all cases are rated or provide scores on a single attribute, then the range will be that defined by the minimum and maximum values for that attribute. Where cases are in fact rating variables, and two individuals' vectors of ratings are being compared to one another, then each "case" might have a unique range. However, using observations which possess different minimum and maximum possible ranges is not optimal for interpretation purposes, and not recommended; as I explain below in section 3.1.

The Gower coefficient computes a scaled similarity coefficient, utilizing scaled discrepancies. It varies between 0 and +1, where +1 is equal to identity between the two vectors being compared.

It indexes the average absolute discrepancy between observations, where the average discrepancy is expressed as a proportion of the total discrepancy which might be observed for the data. Subtracting this average discrepancy from 1 renders it an index of agreement/similarity.

However, directionality of agreement (and strength of monotonicity) between observation pairs can be allocated to this coefficient (positive or negative relationship between observations); the computational procedure for this is presented in Section 3.2.

The formula for the **double-scaled euclidean agreement index** is (**DSE-s**):

$$dse_s = 1 - \left[\frac{\sqrt{\sum_{i=1}^{n} \left(\frac{(case_{1i} - case_{2i})^2}{range_i^2} \right)}}{\sqrt{n}} \right] = 1 - \sqrt{\left[\frac{\sum_{i=1}^{n} \left(\frac{(case_{1i} - case_{2i})^2}{range_i^2} \right)}{n} \right]}$$

where

$n =$ the number of cases

$range_i^2 =$ the squared range of the rating/score attribute for case i $\left(maximum - minimum\ value\right)$

$case_{1i} =$ the observed value for case i of n in the first set of observations

$case_{2i} =$ the observed value for case i of n in the second set of observations

This index computes the squared discrepancy between two cases'/variable's values, then divides this squared discrepancy by the maximum possible squared discrepancy for that case/variable. Summing and taking the square root of these "scaled" discrepancies across cases/variables yields a scaled 'one-dimensional' Euclidean distance. But, the metric of this scaled and cumulatively summed variable discrepancy distance varies between 0 and some value greater than 1.0. In order to scale this coefficient into a unit (0 to 1) metric, a further scaling operation takes place. That is, the initially scaled Euclidean distance is divided by the square root of the number of variables comprising the distance computation. This second scaling now produces a coefficient which always varies between 0 (no distance between variables) to 1 (maximum possible distance between variables given the designated maximum and minimum values for each variable).

This dual scaling ensures that the coefficient is comparable between studies and samples where different variable magnitudes might otherwise distort a conventional Euclidean distance. Further, because the initial scaling of distance is linear (rather than the non-linear operation used within the more common solution of converting data to normalized z-scores prior to any distance calculation), the linear distance relations between magnitudes on the variables remains unchanged. Finally, in order to complete the process, the double-scaled distance is expressed as a similarity index by subtracting it from 1, thus yielding the double-scaled Euclidean similarity (DSE-S) measure, where 0 for this coefficient indicates maximum possible disagreement, and 1 indicates that all cases/variables possess identical rating magnitudes.

The similarity to the Gower coefficient is obvious, but these will produce different coefficient sizes and distribution densities as the Gower is based upon absolute value discrepancy while the double-scaled Euclidean is based upon squared discrepancies.

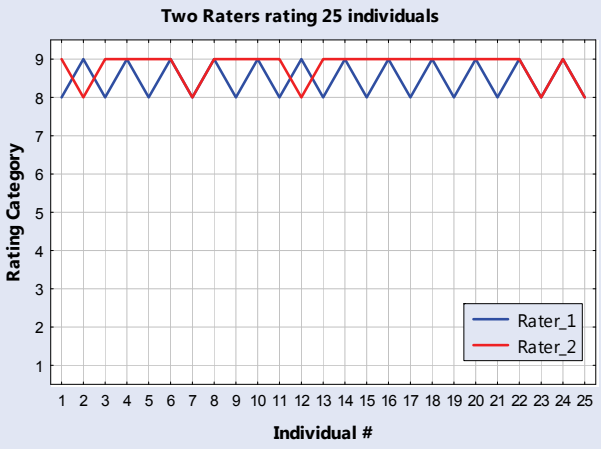
The DSE-S coefficient computes a scaled similarity coefficient, utilizing scaled discrepancies. It varies between 0 and +1, where +1 is equal to identity between the two vectors being compared.

Where cases are in fact rating variables, and two individuals' vectors of ratings are being compared to one another, then each "case" might have a unique range. However, as with the Gower, using observations which possess different minimum and maximum possible ranges is not optimal for interpretation purposes, and not recommended; as I explain below in section 3.1.

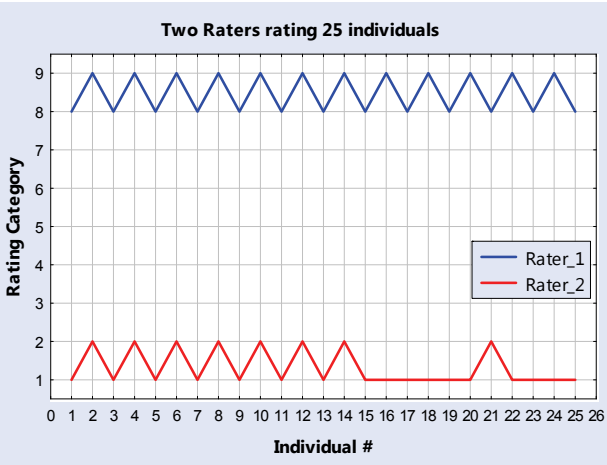
As with the Gower, directionality of agreement (and strength of monotonicity) between observation pairs can be allocated to this coefficient (positive or negative relationship between observations); the computational procedure for this is presented in Section 3.2.

So, if we use these two coefficients on the two sets of data, what happens?

r =.08
icc1 = -.02
icc2 = .07
icc3 = .08
$R_{wg}(j)$= 1.0
Gow = .94
DSEs = .91



r =.54
icc1 = -.97
icc2 = .00
icc3 = .54
$R_{wg}(j)$= 1.0
Gow = .10
DSEs = .10



Both the Gower (Gow) and DSEs provide the kind of similarity estimates which match what the observer sees by eye - when looking at the actual observations rather than transformed versions or variance ratios computed from them.

3.1 How to interpret the Magnitude of a Gower or DSE-S Coefficient.

So far so good, except that the indices need the kind of interpretation which is linked directly to the magnitude of discrepancy between observations. It's pointless producing indices which are no more interpretable than effect sizes, correlations, or ICCs; none of which convey information about precisely how much observations differ from, or agree with one another.

In this regard, the Gower coefficient looks the obvious choice. Why, because it's similarity can be expressed as a percentage of maximum possible similarity (the arithmetic identity between two sets of observations where those observations are numbers). Indeed, even if we don't use numbers as arithmetic quantities, but as symbols indicating ordered classes, it is possible to rework the operations in the index to express class transitions in a similar way.

For example, look at the following scores on two tests, with the minimum and maximum possible scores for each test between 0 and 50 ..

	Var1	Var2
1	30	5
2	30	5
3	30	5
4	30	5
5	30	5
6	30	5
7	30	5
8	30	5
9	30	5
10	30	5

The discrepancy between each pair of observations is 25, which is exactly half the maximum possible discrepancy range of 50. So each paired scaled discrepancy is exactly 0.5, with a resulting Gower and DSE-s of **0.5**.

That really is the logic of a discrepancy-based coefficient in a nutshell. A value of 0.5 tells you that the similarity between observations is 50% of maximum (which is absolute identity).

Another way of expressing this is that the average discrepancy between observations is one half of the maximum possible discrepancy.

However, we have to be careful here. For the Gower, " the maximum possible discrepancy" is the range of the data (the difference between the maximum possible and minimum possible values). The Gower is the average of the *absolute* discrepancies, divided through by the range, subtracted from 1 to provide the measure of similarity.

For the DSE-s, we take the squared difference between the maximum possible and minimum possible values. The average of these squared discrepancies is divided through by the range squared, then the square root taken of the result, which is subtracted from 1 to provide the measure of similarity.

In essence, we have lost the direct interpretability of similarity with the DSE-s, because it uses squared rather than actual discrepancies. The cost of maintaining comparability with euclidean distance has resulted in a coefficient which Is not as straightforward to interpret as the Gower.

Look at the following data, where the minimum and maximum possible scores for each test is now between 1 and 10 ..

	Var1	Var2
1	1	1
2	4	5
3	5	4
4	2	1
5	1	8
6	4	3
7	3	2
8	2	1
9	7	7
10	9	8

The discrepancy between each pair of observations is no longer equal (as in our previous example). The Gower and DSE-s are respectively **0.84** and **0.74** - which can be expressed as similarities of 84% and 74% respectively.

For comparative purposes, the Pearson correlation for these data is: 0.59.

The Gower can be interpreted as:
the average *absolute* discrepancy between observations, expressed as a % of absolute identity.

If we look more closely at what the Gower is using to arrive at the 84%...

	Var1	Var2	Absolute discrepancy between Var 1 and Var 2
1	1	1	0
2	4	5	1
3	5	4	1
4	2	1	1
5	1	8	7
6	4	3	1
7	3	2	1
8	2	1	1
9	7	7	0
10	9	8	1

The sum of these discrepancies is 14, divided through by the number of pairs of observations (10) yields a value of 1.4.

It is this average discrepancy which is scaled by the range of maximum possible discrepancy (10-1 = 9), which yields the result 0.15556, indicating a 16% average discrepancy between Var 1 and Var 2 observation magnitudes. If we subtract this from 1, we achieve our Gower similarity/agreement index of 84%. The DSE-s possesses no such simple logic, because of the squaring of observations and the subsequent square root function of the scaled average discrepancy.

Can we compare ratings or attributes for pairs of attribute observations where the range for each attribute differs by pairs?
Yes, the formulae for both the Gower and DSE-s permit this. BUT, interpreting the coefficients becomes more awkward, as the maximum possible discrepancy varies per attribute. There is no simple way to interpret the coefficient in these circumstances, except insofar as one might interpret the magnitude of a Pearson or ICC coefficient. That is, the minimum possible value for a Gower of DSE-s coefficient is 0, indicating maximum possible disagreement; with 1 indicating absolute identity between pairs of observations. So, 0.8 or 80% could be interpreted a coefficient magnitude indicating good agreement - although we are unable to specify the actual magnitude of expected discrepancy across each attribute (without further calculations being made which focus on each attribute range and observed discrepancy for that attribute).

Instead, I now always linearly convert any data to be compared into the same metric. That is, if pairs of observations are acquired from attributes whose magnitude ranges are not equivalent, prior to Gower (or DSEs) calculations, I linearly convert them all to a common unit metric. I do not standardize or use any other kind of transformation which changes the nature of the data itself. For example, assume we wish to compare the profile similarity of person 1 to person 2, using their scores on 10 personality scales, where some scales possesses a different possible score range.

	Person 1	Person 2	Minimum possible attribute value	Maximum possible attribute value
1	1	2	1	5
2	6	7	1	10
3	12	18	1	20
4	3	5	1	5
5	7	10	1	12
6	40	31	1	40
7	12	6	1	20
8	14	8	1	20
9	5	10	1	10
10	20	10	1	20

One approach might be to use a Pearson correlation, which will standardize each column of data and re-express them as though coming from a distribution with mean 0 and standard deviation of 1.0. The ICCs will take into account the mean and variance of each person's data, as well as the row means and SDs. The results of analyzing these data "as is" are:

Pearson r = .87
ICC Model 1 = .83
ICC Model 2 = .82
ICC Model 3 = .83
Gower = .66 (66% of maximum possible similarity)

The problem for the Gower is answering that question, what is the 66% a percentage of, in terms of the actual expected discrepancy between the two sets of observations?

If we now transform the data into a common unit metric, using the smallest range attribute as the common range (1-5), and re-expressing every other attribute range in that metric:

	Person 1	Person 2	Minimum possible attribute value	Maximum possible attribute value	Linearly rescaled Person 1	Linearly rescaled Person 2
1	1	2	1	5	1	2
2	6	7	1	10	3.222	3.667
3	12	18	1	20	3.316	4.579
4	3	5	1	5	3	5
5	7	10	1	12	3.182	4.273
6	40	31	1	40	5	4.077
7	12	6	1	20	3.316	2.053
8	14	8	1	20	3.737	2.474
9	5	10	1	10	2.778	5
10	20	10	1	20	5	2.895

and now make the same computations on these data using all coefficients, the results are:

Pearson r = .13
ICC Model 1 = .17
ICC Model 2 = .14
ICC Model 3 = .13
Gower = .66 (66% of maximum possible similarity)

The Pearson and ICCS are completely different - all reduced from above 0.87 to below 0.20, simply because we re-expressed the different attribute observations in a common metric.

The Gower remains the same value (66%) because the formula takes into account varying attribute ranges. I simply re-expressed the attribute magnitudes into a common metric, using a linear transformation which does not affect magnitude ratios.

But, because of this common metric, we can now interpret the 66% in terms of the expected average similarity between any pair of observation as 66%, or each pair of observations will be expected to differ by 34% from each other, on average.

Taking into account the common measurement range of the attributes (=4), that represents an expected *absolute* discrepancy between observation pairs of 1.36.

Ultimately, we could express the 66% in terms of each attribute measurement range (untransformed) - by computing what 34% error looks like for each attribute.

	Person 1	Person 2	Minimum possible attribute value	Maximum possible attribute value	Rounded Expected 34% Absolute Discrepancy
1	1	2	1	5	1
2	6	7	1	10	3
3	12	18	1	20	6
4	3	5	1	5	1
5	7	10	1	12	4
6	40	31	1	40	13
7	12	6	1	20	6
8	14	8	1	20	6
9	5	10	1	10	3
10	20	10	1	20	6

And remember, the indices computed directly from these data were:

Pearson r = .87
ICC Model 1 = .83
ICC Model 2 = .82
ICC Model 3 = .83
Gower = .66 (66% of maximum possible similarity)

Yet if we transform them into a common-unit metric prior to the same calculations, we observe:

Pearson r = .13
ICC Model 1 = .17
ICC Model 2 = .14
ICC Model 3 = .13
Gower = .66 (66% of maximum possible similarity)

Another reason never to use the Pearson correlation nor ICC for any kind of multi-attribute profiling, unless the data really do conform to the assumptions inherent in both the Pearson and ICCs, and all observations are expressed in a common metric.

But even when we transformed all data into a common unit metric, the results indicated by these coefficients seem at odds with the actual observed discrepancies observed between observations (in the table on the previous page). The picture doesn't look anywhere as bleak as a correlation or ICC of 0.13.

3. 2 How to detect directionality and magnitude of monotonicity

Sometimes, it is important to know whether observations are not only in agreement to one another, but whether they share a directional relation, and how strong such a relation is. That is, observations on one attribute tend to increase as observations on another increase, or whether as observations on one attribute increase, they decrease on the other (the monotonicity principle). The Pearson correlation and ICC coefficients do provide such information. Coefficients of agreements based upon absolute or squared discrepancies do not.

So, in order to provide these coefficients with a sign indicating directionality of observation relationship, along with the magnitude of agreement between observations, but only using the simple properties of the observations themselves (not transformed or using summary distribution-related parameters), I needed to develop an algorithm which would impart relationship sign to an agreement index. For binary observations (Yes-No, Agree-Disagree, Present-Absent), I could use a variety of signed agreement indices. However, where the number of response options/magnitude steps increase in ordered class, or semi-interval data, we need a method which will provide an indication of relationship direction.

The solution as conceived in my original exposition of the issues worked only as long as there were no "tied" values within a vector of observations (i.e. all observations were unique values). That was clearly not acceptable.

So, another approach was undertaken by visualizing the problem "geometrically"; which involves calculating the angle of a least-squares best-fit line through the cluster of points defined by two vectors of data, where the raw-data vectors are initially mean-centered and normalized into a unit-metric, Cartesian coordinate space of {1-.0 to +1.0}. Yep, it can be done graphically/computationally, but why bother when all I have actually re-created is the Pearson correlation!

Right now, I see no other optimal way of determining monotonicity which is asymmetric (i.e. it doesn't matter which variable is used as X or Y). The Pearson works well in this regard as long as you remember it indexes monotonicity "quantitatively"; that is, it assumes the order relations between vectors are additive. But, the Gower is also functioning on the basis that the discrepancy between vector values is quantitative. In the end, what matters is that you are aware that manipulating the numbers this way does not necessarily mean that the attribute magnitudes themselves vary in the same way. That is, rating judgments or personality attributes, for example, may not vary in the same way as numbers do. We can take advantage of the convenience of working with numbers, but we don't have to assume that the attributes themselves "behave" in the same way. Which is why it's always important to assess the theory-relevant predictive accuracy of whatever you claim to be a "measurement" of something.

I'm using the Pearson coefficient to index monotonicity and the Gower to index the agreement between magnitudes. The Pearson cannot be used for this latter task as it removes precisely that information required to assess agreement between the observations. The examples above make this very clear.

I've retained the examples below to show why computing monotonicity and agreement as separate indices helps better understand how attribute magnitudes relate to one another.

Example 1: minimum possible value = 1, maximum = 5

	Var1	Var2
1	1	5
2	2	4
3	3	3
4	4	4
5	5	5

Pearson Monotonicity = .0
ICC Model 1 = –.11
ICC Model 2 = .0
ICC Model 3 = .0
Gower = .70 (= 70 % of maximum possible similarity)

Example 2: minimum possible value = 1, maximum = 10

	Var1	Var2
1	1	10
2	2	9
3	3	8
4	4	7
5	5	6
6	6	5
7	7	7
8	8	7
9	9	7
10	10	5

Pearson Monotonicity = –.75
ICC Model 1 = –.63
ICC Model 2 = –.58
ICC Model 3 = –.62
Gower = .62 (= 62 % of maximum possible similarity)

Example 3: minimum possible value = 0, maximum = 1

	Var1	Var2
1	0.5	0.3
2	0.3	0.2
3	0.2	0.3
4	0.1	0.1
5	0.4	0.4
6	0.5	0.5
7	0.7	0.2
8	0.4	0.4
9	0.2	0.7
10	0.3	0.5

Pearson Monotonicity = –.06
ICC Model 1 = –.00
ICC Model 2 = –.06
ICC Model 3 = –.06
Gower = .84 (= 84 % of maximum possible similarity)

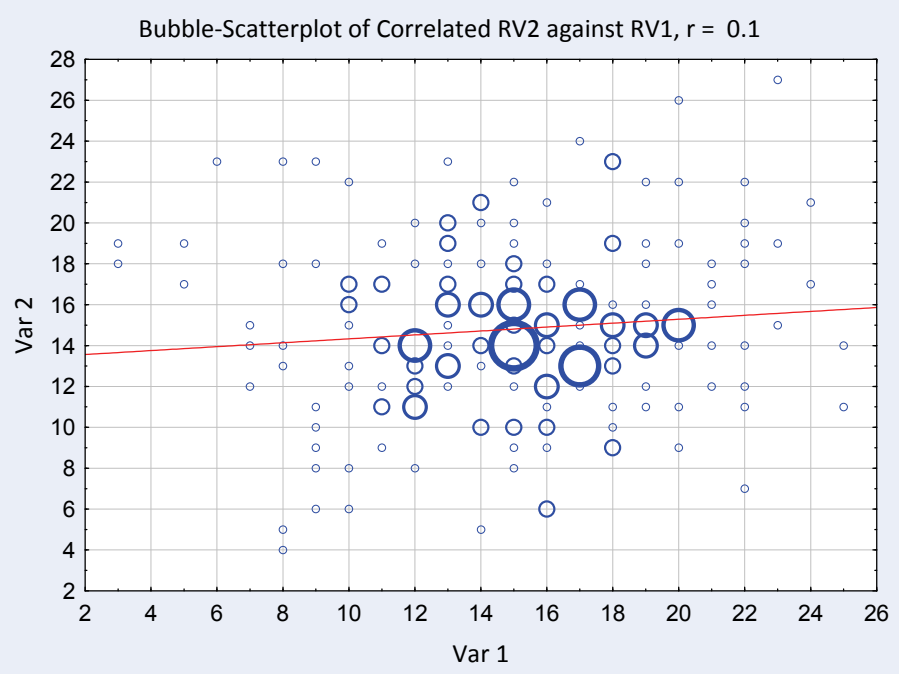
if the **minimum possible value was -1,with a maximum = 5,** then:
Gower = .92 (= 92 % of maximum possible similarity)

because the average discrepancy (0.16) is scaled by the maximum possible discrepancy (2) instead of (1). Relative to the extended maximum possible absolute discrepancy range, the observed average absolute discrepancy of 0.16 is trivial.

Example 4: minimum possible value = 1, maximum = 30. 200 pairs of observations, generated as *bivariate-normal* distributed integers with a Pearson correlation of 0.1.

Pearson Monotonicity = .10
ICC Model 1 = .10
ICC Model 2 = .10
ICC Model 3 = .10
Gower = .85 (= 85 % of maximum possible similarity)

I've plotted the data as a bubble-frequency -scatterplot on the next page, with the larger circles reflecting higher frequency of observations. Clearly there isn't much monotonicity here …



The reason for the high Gower index is the volume of pairs of observations near the middle of each Variable range (as expected from a perfect normal bivariate distribution with such a low Pearson correlation between values. If we look at the frequency distribution of absolute score discrepancies, we see:

Category	Frequency table: Discrepancy: =abs(v3) (test-gow3.sta)			
	Count	Cumulative Count	Percent	Cumulative Percent
0	15	15	7.5	7.50
1	34	49	17.0	24.50
2	23	72	11.5	36.00
3	20	92	10.0	46.00
4	26	118	13.0	59.00
5	22	140	11.0	70.00
6	14	154	7.0	77.00
7	15	169	7.5	84.50
8	7	176	3.5	88.00
9	6	182	3.0	91.00
10	5	187	2.5	93.50
11	3	190	1.5	95.00
12	2	192	1.0	96.00
14	3	195	1.5	97.50
15	3	198	1.5	99.00
16	1	199	0.5	99.50
17	1	200	0.5	100.00
Missing	0	200	0.0	100.00

The average absolute discrepancy is 4.4, with a median value of 4, over a total possible magnitude discrepancy range of 29.

The interquartile range (middle 50%) of signed discrepancies lie between -3 and +4 (over a total possible magnitude discrepancy range of ± 29)

This example highlights the difference between a coefficient sensitive to monotonicity, and one which is sensitive to the actual magnitude discrepancies of observations.

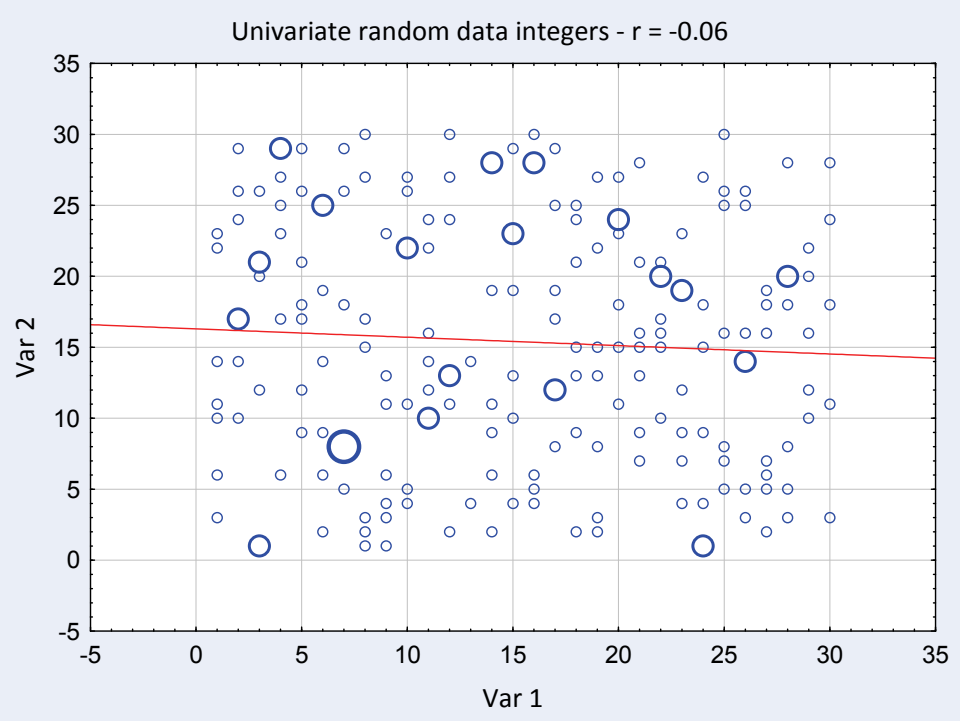
Random Data Examples

These are important - as the values of the Gower (and DSE-s) are not 0 when observations are entirely random, as examples 5 and 6 show below.

Example 5: minimum possible value = 1, maximum = 30. 200 pairs of observations, generated as *uniform* distributed random integers.

Pearson Monotonicity = –.06
ICC Model 1 = –.06
ICC Model 2 = –.06
ICC Model 3 = –.06
Gower = .65 (= 65 % of maximum possible similarity)

I've plotted the data as a bubble-frequency -scatterplot, with the larger circles reflecting higher frequency of observations.



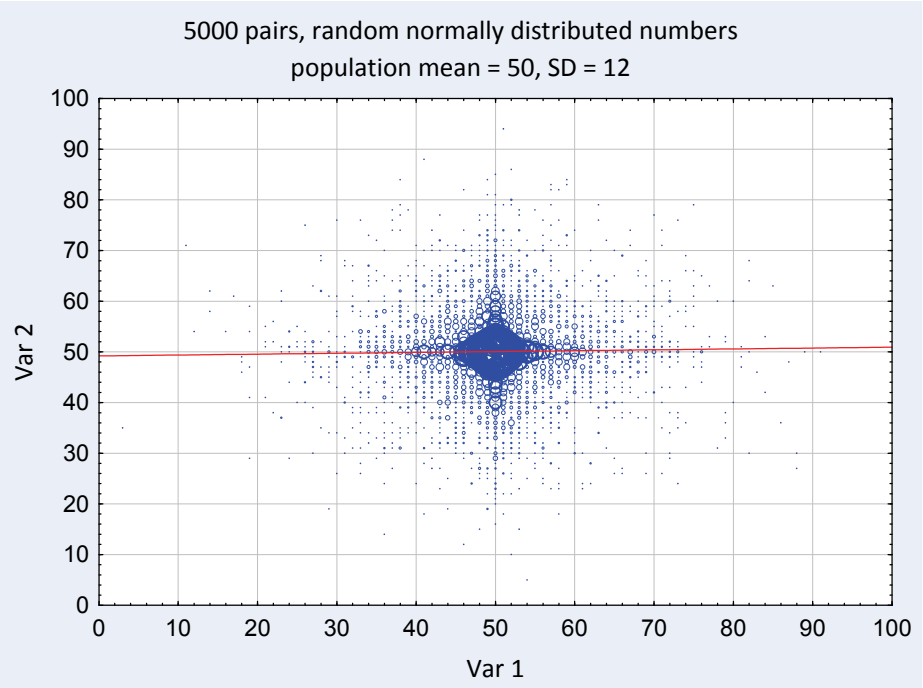
As can be seen, there are fewer pairs of very similar observations (reflected in the decreased Gower index). The middle 50% of signed discrepancies now lie between -10 and 9, which given a ± 29-point maximum possible discrepancy range, looks pretty too high for comfort.

What this brings home is that the expected value of the Gower for random data (depending upon the distribution of those random numbers) will never be near zero, because that represents a situation where pairs of observations are at the minimum and maximum value of the total magnitude range, which is not random.

Example 6: minimum possible value = 0, maximum = 100. 5000 pairs of observations, generated as *normally* distributed random integers, sampled from a perfect normal distribution with mean of 50 and SD of 12.

Pearson Monotonicity = .02
ICC Model 1 = .02
ICC Model 2 = .02
ICC Model 3 = .02
Gower = .91 (= 91 % of maximum possible similarity)

It's worth dissecting these data to see why the Gower is so high. First, a bubble-frequency scatterplot.



As can be seen from the dense blue area, the majority of observations are clustered around the joint mean of 50.

Let's look at the frequency distribution of the absolute score discrepancies themselves (between pairs observations):

Category	Frequency table: Discrepancy: =abs(v1-v2) (Test Data, 5000 normal random cases 0-100.sta)			
	Count	Cumulative Count	Percent	Cumulative Percent
0	236	236	4.720	4.72
1	421	657	8.420	13.14
2	398	1055	7.960	21.10
3	404	1459	8.080	29.18
4	334	1793	6.680	35.86
5	300	2093	6.000	41.86
6	287	2380	5.740	47.60
7	263	2643	5.260	52.86
8	251	2894	5.020	57.88
9	219	3113	4.380	62.26
10	223	3336	4.460	66.72
11	203	3539	4.060	70.78
12	174	3713	3.480	74.26
13	117	3830	2.340	76.60
14	141	3971	2.820	79.42
15	125	4096	2.500	81.92
16	104	4200	2.080	84.00
17	97	4297	1.940	85.94
18	79	4376	1.580	87.52
19	83	4459	1.660	89.18
20	74	4533	1.480	90.66

in fact the actual average *absolute* discrepancy for these data is 9,with a median discrepancy of 7, relative to a total possible discrepancy range of 100 units. That's why the Gower is so high. The pairs of observations are actually quite similar to one another, given the total possible magnitude range of each variable, and the size of absolute discrepancy we have observed.

But, the monotonicity for these data is 0.02, which is negligible; indicating no consistent relationship between score magnitudes.

What these random data examples highlight is that:

1. Both agreement and monotonicity need to be interpreted, separately.

2. The magnitude of a Gower coefficient reflects absolute discrepancy/agreement relative to the maximum possible discrepancy/agreement given the magnitude range of each variable.

3. The Gower magnitude for random data is affected by the choice of sampling distribution for "random" variables. A bivariate uniform random distribution will ALWAYS produce lower similarity estimates than sampling from a bivariate normal distribution, because of the clustering of observations around each respective mean in the bivariate normal distribution sampling.

The ICCs and Pearson especially are mostly sensitive to monotonicity, and not agreement between observations; the proof of that is contained in the graph and results on pages 8 & 9 of this document.

The key to using the Gower index is to utilize its feature of expressing similarity in % terms, which can be converted into actual expected discrepancies between pairs of observations, where the vector of discrepancies can also be further analyzed into a frequency distribution of magnitude discrepancies.

Remember, we are indexing direct observational agreement, independent of the distribution of those observations, their monotonicity, their variances, and their standardized values.

3.3 Statistical Significance?

The obvious approach is by creating an empirical bootstrap sampling distribution for Gower coefficients, parameterized by the magnitude range of the particular set of observations from which a Gower has been created, and whether the observations within each attribute are thought to be equally possible (a uniform distribution function) or more likely to occur near some specified mean value for an attribute, with a certain variance around that mean (the normal distribution).

Right now, I advocate a uniform distribution as without any prior evidence of what the population mean and SD should be for a random number distribution for two attributes, choice of distribution parameters remains ad-hoc. And, by compressing/expanding those distributions around the mean (by varying the SD), you can change the distribution of bootstrap sample Gower coefficients, and thus the expected magnitude at the 95th percentile.

In the past, I have generated sampling distributions containing between 10,000 to 32,000 coefficients; 20,000+ coefficients seems to produce sufficiently precise percentile intervals.

4. Three Real-world applications.

Three applications are presented here, which demonstrate the utility of this new approach to indexing observation agreement and directional relationships. All are taken from a combined New Zealand and US dataset of university students, who took part in the initial round of trialing of the new Hogan Cognitive Inventory (HCI). Not all students from each country provided data for the three applications, so sample compositions change per application.

Without going into details, the test is designed to assess reasoning among higher-ability business executives. The items are scored correct/incorrect, with a prototype total test score-range between 0 and 22. While data was acquired on the test items themselves, some relevant outcome criteria were also collected from the US students, specifically their ACT scores. In addition, the US students were also asked to provide self-report ratings of behaviors indicative of aspects of their cognitive functioning e.g. "I use information from one solution to solve other problems."

The three applications consist of a typical criterion validity coefficient estimation, a concurrent validity exercise, and the determination of the correlation between two subscales of the test, "Rule Applying", and "Rule Finding".

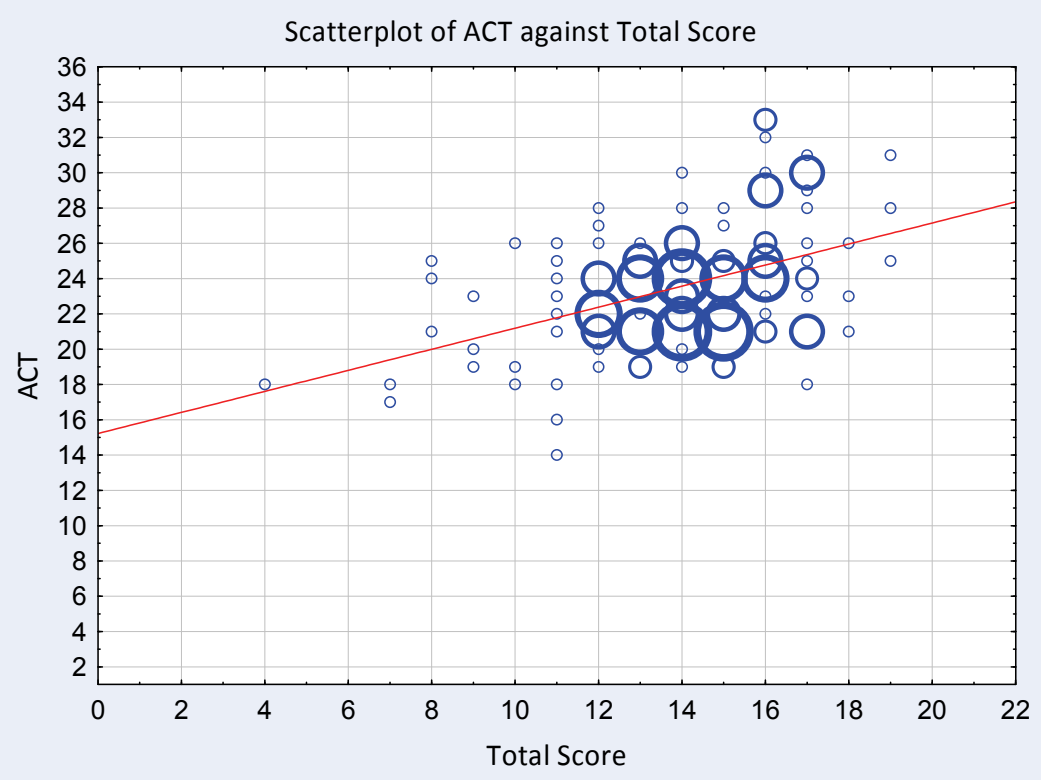
4.1 Estimating the Validity of the HCI Total Test Score (predicting ACT scores)

The Pearson correlation between the ACT score and HCI total scale score was **0.44** (n=135, US students only).

Of importance here is the paper by Koenig, Frey, & Detterman (2008) analyzing ACT and standard IQ test scores. On pages 158-159 of the paper the authors state:

> "As discussed in the opening of this article, ACT, Inc. claims that the ACT is not an IQ test, but rather measures the preparedness of the test-taker for advanced education. Given the results of the current study, this statement is misleading. Colleges that use scores on the ACT and SAT for admission decisions are basing admissions partially on intelligence test results. Whether this is acceptable or efficient practice is beyond the scope of this article, but we argue that the testing companies have a responsibility to the public to accurately describe what these widely-used tests measure. "

So, the ACT is in essence, a measure of general psychometric intelligence (g). The correlation is clearly important in this context. But, as the correlation/agreement discussions above have indicated, there are grounds for thinking that this "validity correlation" might be substantially underestimating the actual relationship between the ACT score and HCI test score. In this context, it is useful to examine the HCI total scale score bubble-frequency scatterplot showing the full possible measurement range of each variable.



What's clear is that the both HCI Total Scores and ACT scores are located more in the upper quadrant of each score range. Basically, higher ACT scores are associated with higher HCI scores.

The Pearson correlation of 0.44 is an indication of this monotonicity, but with an r^2 of just 20%, the accuracy of prediction looks rather poor if we only had access to that coefficient, and had not seen the scatterplot. Some might try and correct the coefficient for supposed "restriction of range", but why? This is simply a recognition that the Pearson coefficient is largely inaccurate unless the assumptions upon which the coefficient is dependent are completely met. It's a fact that students at university will have higher ACT scores then those who failed high-school. What's the point of correcting a correlation on this basis. The data acquired are exactly what a practitioner or decision-maker will have to work with, in the real world, on a daily basis; not some artificial, statistically-sanitized version.

Instead, I used a coefficient which is only sensitive to the absolute agreement between scores, **not monotonicity**, taking into account the potential measurement range of each variable. The logic is that if the bulk of ACT scores are located higher within their possible range, as well as the bulk of HCI scores being located higher within their possible range, then doesn't this tell us something about how well these scores "go together", irrespective of whether monotonicity is retained within the paired observations?

Prior to computing the Gower coefficient, all HCI scores were linearly translated into a 1-36 integer range to match the magnitude range of the ACT scores. The Gower coefficient was then computed on

these common-metric data, with a value of **0.90** (90% similarity), with an average absolute discrepancy between ACT and HCI total scores of 3 (3.3) , over a 35-point score range.

The frequency distribution of absolute score discrepancies is:

	Frequency table: Absolute Discrepancy: =abs(v1-v2) (Spreadsheet1)			
Category	Count	Cumulative Count	Percent	Cumulative Percent
0	15	15	11.11	11.11
1	21	36	15.56	26.67
2	24	60	17.78	44.44
3	22	82	16.30	60.74
4	19	101	14.07	74.81
5	7	108	5.19	80.00
6	8	116	5.93	85.93
7	9	125	6.67	92.59
8	5	130	3.70	96.30
10	2	132	1.48	97.78
11	2	134	1.48	99.26
12	1	135	0.74	100.00
Missing	0	135	0.00	100.00

indicating that 75% of cases possessed an absolute score discrepancy of 4 or lower. That's why the Gower is so high. Relative to the total score range, the discrepancy between observed scores is pretty low.

The Pearson simply indexes the strength of the positive monotonic trend: for higher ACT scores to be associated with higher HCI scores.

Statistical Significance?

In terms of statistical significance of this Gower coefficient, I created a distribution of 20,100 Gower coefficients by constructing a pairwise bootstrap of 201 sets of random observations, each set consisting of 135 integer magnitudes sampled from a uniform distribution whose bound-values were between 0 and 12.

The minimum Gower coefficient observed in this distribution was 0.597. The maximum was 0.764. Our observed Gower coefficient of 0.90 is clearly not a "chance" phenomenon.

4.2 Estimating the Validity of the HCI Total Test Score (predicting self-report behavioral/ cognitive-style scores)

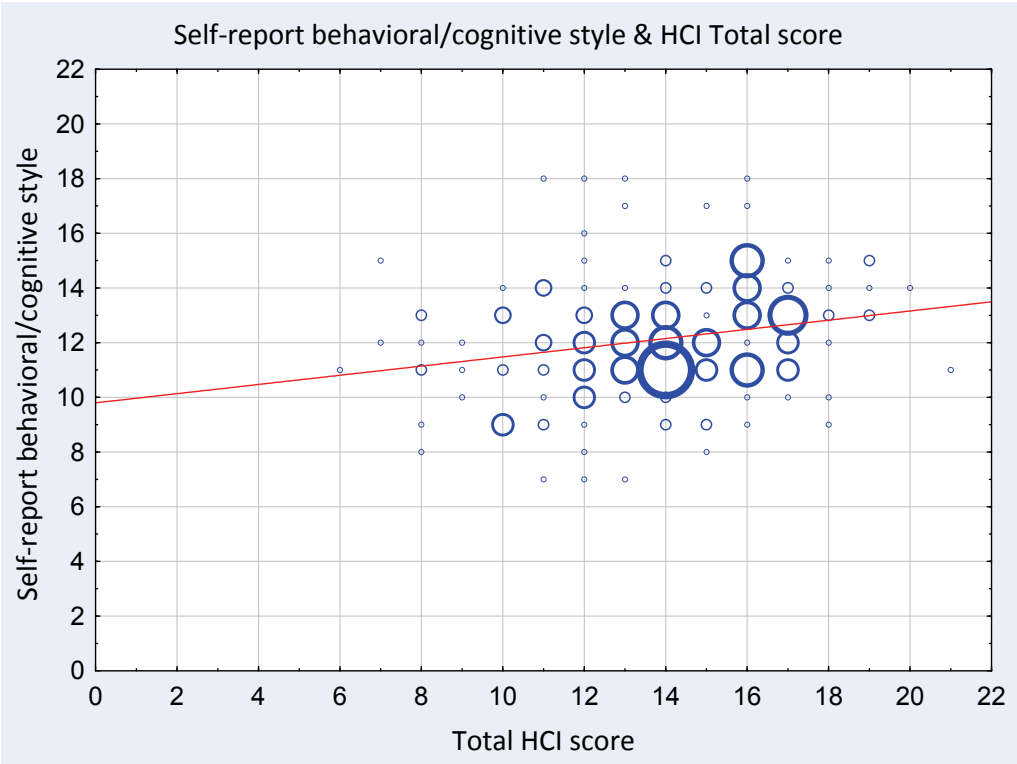
The responses to 16 items assessing behavioral/cognitive style features were shown to form a reasonably homogenous composite. The item responses were acquired using a 5-point Likert scale, yielding a total possible score range of between 16 and 80. The 16-80 score range of the composite scale was linearly compressed into the respective score range of the total HCI scale {0-22} using integer rounding. 184 complete cases of US student data were available for analysis.

The Pearson monotonicity coefficient between the behavioral/cognitive style composite scale and the HCI total scale score was **0.23**. The Gower index of agreement was **0.87** (87% of maximum possible agreement).

The discrepancy frequency distribution between the 16-item behavioral/style composite and the HCI total score (unified common metric between 0 and 22) is shown below, indicating just why the Gower agreement is so high. The median *absolute* discrepancy between the two sets of scores is 3 (within a 22-point measurement range). The interquartile range is between 0 and 4, and the middle 80% of signed discrepancy observations lie between -3 and 5.

	Frequency table: Discrepancy: =v1-v2 (Spreadsheet126)			
Category	Count	Cumulative Count	Percent	Cumulative Percent
-8	1	1	0.42	0.42
-7	1	2	0.42	0.85
-6	1	3	0.42	1.27
-5	5	8	2.12	3.39
-4	4	12	1.69	5.08
-3	10	22	4.24	9.32
-2	4	26	1.69	11.02
-1	14	40	5.93	16.95
0	14	54	5.93	22.88
1	27	81	11.44	34.32
2	24	105	10.17	44.49
3	26	131	11.02	55.51
4	19	150	8.05	63.56
5	16	166	6.78	70.34
6	12	178	5.08	75.42
7	3	181	1.27	76.69
8	1	182	0.42	77.12
9	1	183	0.42	77.54
10	1	184	0.42	77.97
Missing	52	236	22.03	100.00

A bubble-frequency scatterplot of the two sets of scores is shown below.



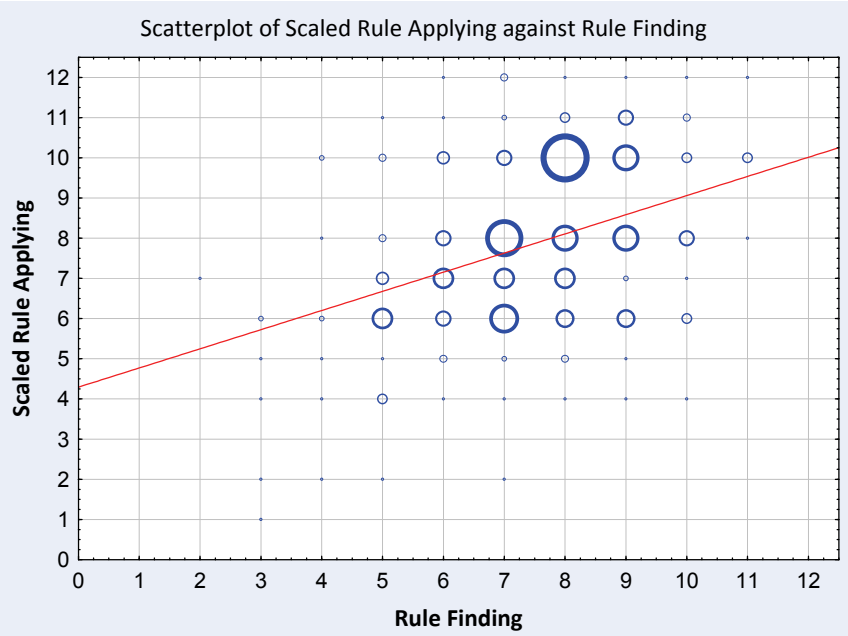
As shown by the size of those frequency bubbles, clustered around 10-14 on both scales, these scores are much more similar to one another than indicated by ICC or Pearson indices, whose values are typical for these kinds of comparisons (ICC models 1, 2, and 3 produce coefficients respectively of .09, .18. and .22).

Once again, the use of an appropriate validity coefficient coupled with an independent assessment of monotonicity has demonstrated results which are very different from relying on a conventional Pearson /variance-ratio approach to estimating validity.

4.3 Estimating the agreement/relationship between the two subscales of the HCI test.

Using a combined US and NZ student sample of 236 cases, the Pearson correlation between the two subscale scores of the HCI was computed as **0.39**. That would normally be taken as evidence of some degree of independence between the two scores, important for interpretation/feedback processes. However, the Gower index is **0.85**, with a monotonicity index of **0.013**.

The bubble-frequency scatterplot for these data is:



which shows that same clustering effect. The frequency distribution of signed score discrepancies is:

	Frequency table: Discrepancy: =v2-v3 (Spreadsheet3)			
Category	Count	Cumulative Count	Percent	Cumulative Percent
-6	4	4	1.7	1.7
-5	8	12	3.4	5.1
-4	9	21	3.8	8.9
-3	16	37	6.8	15.7
-2	39	76	16.5	32.2
-1	46	122	19.5	51.7
0	30	152	12.7	64.4
1	41	193	17.4	81.8
2	20	213	8.5	90.3
3	14	227	5.9	96.2
4	6	233	2.5	98.7
5	2	235	0.8	99.6
6	1	236	0.4	100.0
Missing	0	236	0.0	100.0

The interquartile range (middle 50%) of signed discrepancies is between -2 to +1, and the middle 80% of these discrepancies between -3 and +2.

The median *absolute* discrepancy is 2.

These numbers have to be put into the context of a 13-point measurement range (0-12).

The Pearson coefficient is low is largely due to the lack of monotonicity within regions of the data. So, even when scores are very similar to one another (say just one-score point apart), if there is little systematic variation of directionality of these discrepancies, the Pearson will ignore the obvious agreement because of that lack of monotonicity.

5. Conclusions

By separating out agreement from monotonicity, the new approach to estimating reliability and validity is suggestive of the proposition that psychological test and rater reliability and validity has been substantively underestimated in what might be the vast majority of investigations. Those seeking predictive accuracy and profile agreement using Pearson and other indices which depend upon variance ratios might now reconsider this methodology.

Some datasets may not show the expected gain in predictive accuracy, validity, or reliability. But, given the logic. rationale, and the separation of monotonicity from agreement in the Gower and DSEs, I suspect most if not all datasets will actually benefit from using this new approach. By "benefit", I mean a really dramatic increase in validity and reliability, in a way that is easily interpretable and understood with direct reference to the observations themselves.

In common with James Grice at Oklahoma State University, and the philosophy, logic, and principles of his Observation Oriented Modeling, this new approach to estimating test and rater reliability and validity gives primacy to observations. Not hypothetical sampling distributions, not hypothetical true scores, not hypothetical latent variables, not transformed observations, and not derived parameters or variance ratios constructed from the observations.

What's really difficult in this kind of work is not the computations themselves, but the necessary cognitive disconnection from the statistical data model assumptions which currently influences the thinking of many psychologists who use data-model-predicated statistical analyses. It is as Leo Breiman (2001) has stated in the abstract to his paper on Statistical Modeling

> " There are two cultures in the use of statistical modeling to reach conclusions from data. One assumes that the data are generated by a given stochastic data model. The other uses algorithmic models and treats the data mechanism as unknown. The statistical community has been committed to the almost exclusive use of data models. This commitment has led to irrelevant theory, questionable conclusions, and has kept statisticians from working on a large range of interesting current problems. Algorithmic modeling, both in theory and practice, has developed rapidly in fields outside statistics. It can be used both on large complex data sets and as a more accurate and informative alternative to data modeling on smaller data sets. If our goal as a field is to use data to solve problems, then we need to move away from exclusive dependence on data models and adopt a more diverse set of tools. "

You have now seen what awaits the test publisher/investigator who is prepared to set robust explanatory/ predictive accuracy as the primary goal of any analysis. It is an approach characterized by the need to answer some rather obvious questions about the data (observations) themselves, and not transformed or idealized versions of them.

The forthcoming whitepaper #10: Assessing interrater reliability: directly, simply, and with no data model/test theory assumptions, extends this same logic into the rater agreement assessment, while enhancing the 'certainty' of agreement using another novel coefficient: Kernel Smoothed Distance.

References

Baguley, T. (2009) Standardized or simple effect size: What should be reported? *British Journal of Psychology*, 100, 3, 603-617.

Baguley, T. (2010) When correlations go bad. *The Psychologist*, 23, 2, 122-123.

Barrett, P.T. (2005) Person-Target Profiling. In André Beauducel, Bernhard Biehl, Michael Bosnjak, Wolfgang Conrad, Gisela Schönberger, and Dietrich Wagener (Eds.) *Multivariate Research Strategies: a Festschrift for Werner Wittman*. Chapter 4, pp 63-118. Aachen: Shaker-Verlag.

Barrett, P.T. and Rolland, J.P. (2009) *The Meta-Analytic Correlation between the Big Five Personality Constructs of Emotional Stability and Conscientiousness: Something is not quite right in the woodshed*. Advanced Projects R&D Ltd., Strategic whitepaper series #3. Downloaded from: http://www.pbarrett.net/stratpapers/metacorr.pdf

Breiman, L. (2001) Statistical Modeling: the two cultures. *Statistical Science*, 16, 3, 199-231.

Gower, J. C., 1971. A general coefficient of similarity and some of its properties. *Biometrics,* 23,623-637.

Grice, J. (submitted) Observation Oriented Modeling: An Introduction. Book manuscript under review.

Hogan, R. and Kaiser, R. (in press) Personality. In J.C. Scott and D.H. Reynolds (eds.*) Handbook of Workplace Assessment*. Chapter 4.

James, L.R., Demaree, R.G., &Wolf, G. (1993) rwg: an assessment of within-group interrater agreement. *Journal of applied Psychology*, 78, 3, 306-309.

Koenig, K.A., Frey, M.C., & Detterman, D.K. (2008) ACT and general cognitive ability. *Intelligence*, 36, 2, 153-160.

Michell, J. (1997) Quantitative science and the definition of measurement in Psychology. *British Journal of Psychology*, 88, 3, 355-383.

Michell, J. (2008) Is Psychometrics Pathological Science?. *Measurement: Interdisciplinary Research & Perspective*, 6, 1, 7-24.

Wellenreuther, M., Barrett, P.T., & Clements, K.D. (2009) The evolution of habitat specialisation in a group of marine triplefin fishes. *Evolutionary Ecology,* 23, 4, pp. 557-568.

Acknowledgments

I would like to thank Bob and Joyce Hogan for permission to report some of the analysis results associated with the trialing of the prototype Hogan Cognitive Inventory, and James Grice at OSU for his perceptive insights on features of this work.