Polskie Forum Psychologiczne, 2013, tom 18, numer 1, s. 5-16

Artykuły
[*Articles***]**

A NOTE ON THE CORRELATION OF GAIN SCORES AND ARCHIEVEMENT LEVEL

Jürgen Rost
Christian-Albrechts-Universität zu Kile

Summary. The negative correlation between gain score and initial status is one of the classical dilemmas in the measurement of change. A simple but efficient method is proposed to get valid information about the relationship between change and the level of achievement. After finishing and submitting this paper, the author became aware of the fact that the proposed rotation of the 2-dimensional space defined by pre- and post-test has already been presented and discussed by P.D. Oltham 50 years ago. However, there is no reason to withdraw the paper, since the majority of empirical researchers still try to derive correct results on the relationship of level and growth without the simple but efficient method of rotating the data space by 45 degrees. Adressed to these researchers, I would say 'it's time to make a change.'
Key words: archievement level, rejeated measurement.

The book on *Problems in measuring change* by Harris (1963) was a milestone in the development of methods for measuring change. It addresses methodological problems and statistical artefacts on the common way of analysing change data that also were considered to be dilemmas, i.e. unsolvable problems (Bereiter, 1963). In the meantime, many solutions for these problems have been discussed in the literature, but not all of them have been accepted as a solution by applied researchers. One of them is the negative correlation between initial status and change (Rogosa and Willett, 1985; Wainer and Brown, 2004?), another one the regression towards the mean (Campbell and Kenny, 1999) and a third one the lack of reliability of difference scores (Collins, 1996; Mellenbergh, 1999; Fischer, 2003).

In this paper, there will be proposed only one new argument, and this may even not be new. However, if one proceeds in accordance with this, the three men-

Adres do korespondencji: Jürgen Rost, e-mail: an@j-rost.de

tioned problems loose their property of being a problem. There is no widely accepted taxonomy of problems of measuring change, but a source of many problems is connected to the question, if the amount of change in a two-time-points measurement situation depends on the individual starting level. The easiest way of finding an answer to this question is to correlate pretest and gain (difference) score. However, this is exactly what should not be done.

The negative correlation between pretest and gain score

Whenever two stochastically independent variables X and Y are measured and their difference $D = Y - X$ is correlated with either the subtrahend or the subtractor, the correlation is negative in the case of X and positive in the case of Y. Moreover, if both distributions are uniform, it can be calculated that $r(X,D) = -\sqrt{0.5} = -0.707$ and $r(Y,D) = \sqrt{0.5} = 0.707$. This fact is not surprising, because one of the two variables correlated here, D, is a function of the others, X and Y.

In the context of measuring change, this becomes to be a problem, as the correlation of X and D often is calculated for empirical data in order to get information about the dependence of change on the initial status. The size of these correlation coefficients has no empirical meaning, because it is a mixture of information from the data and logical necessity. There is much less interest in the correlation of D and Y, which is positive for independent variables. From an applied perspective, the positive correlation between D and Y makes sense, because high Y-values can easier be reached by persons with a high learning gain. The backside of the coin, i.e. the negative correlation between D and X, contradicts the educational expectation, that bright students (high X-values) have higher learning effects, because they are better prepared to understand and encode new instructional material or whatever the treatment provides. Students who perform worse are expected to have smaller learning gain.

Of course, those considerations are obsolete, since the reported correlations have no interpretations at all, they are statistical phenomena without any empirical meaning. The statistical relations make it impossible to interpret an empirically assessed correlation between pretest and gain score or posttest and gain score. The empirical relations, e.g. whether poorly performing students have a smaller learning effect, cannot be answered on the basis of this correlation. It seems to be one of the unsolved problems of change measurement to disentangle the empirical and the statistical proportion of the calculated correlation between pretest and learning gain.

Many attempts to circumvent this problem have been discussed in the literature. Not successful are those proposals, that focus on the error of measurement. The measurement error may strengthen the negative correlation between pretest and difference, because the error that is part of X is correlated with the negative value of itself as part of the difference. The impact of the error of measurement can be controlled by appropriate attenuation formulas (Lord, Novick, 1968, p. 73). But the artificial negative correlation remains after removing the error of measurement.

A very elegant way of eliminating the effect of measurement error would be not to use the same data for calculating the pretest and the difference score. If, e.g., half of the items of the test instrument were used for computing the pretest score and the other half for computing the difference score, the correlation of -0.73 between pretest and difference drops to -0.56 in a Rasch homogeneous test with 20 items.

A small artificial data example shows that the negative correlation can be observed even for data that are free of any error of measurement. Figure 1 shows a scatter-plot of 16 persons with uniformly distributed and uncorrelated measures *X* and *Y* (figure 1a).

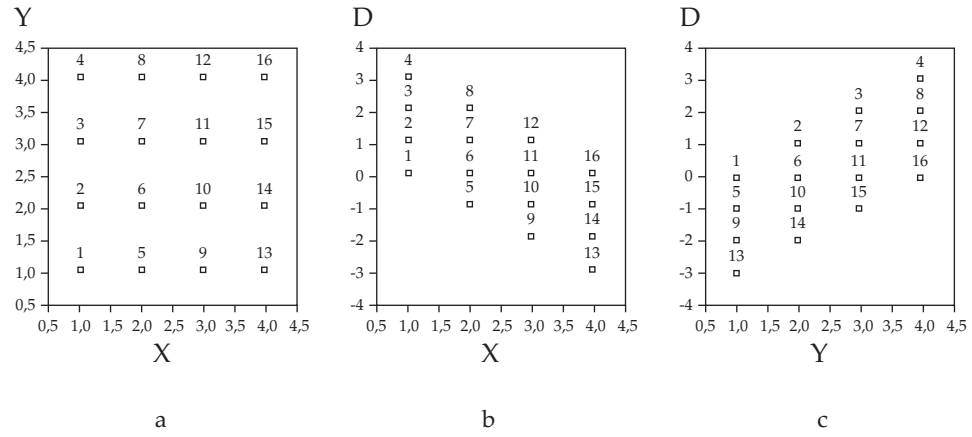

Figure 1. 16 artificial persons with uniform and uncorrelated pre- and posttest measures (a) and their correlations with the difference score *Y-X*; (b) and (c)

Figures 1b and 1c show the negative correlation between *X* and $D = Y - X$, and the positive between *Y* and *D*, respectively. Correlating *D* with *X* produces the same value of a negative correlation as the correlation of *D* and *Y* shows in the positive direction (-0.707 and + 0.707, which is cosine (45º) or $\sqrt{0.5}$). One conclusion of the inspection of these figures would be that the artificial distortions between the difference scores and achievement level should vanish, if both measures, *X* and *Y*, were used to estimate the achievement level.

The zero correlation between difference and sum score

The most straightforward way of taking into account both measures, *X* and *Y*, for the measurement of the achievement level[1], would be to take the sum of both, $S = X + Y$, as an estimate of the individual level. Figure 2 shows that *D* and *S* are uncorrelated as *X* and *Y* are.

[1] P.D. Oldham published in 1963 this transformation and provided some more good arguments for this method.

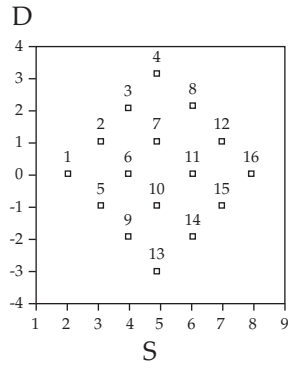

Figure 2: The artificial data in a rotated space

It can also be drawn from figure 2 that the two-dimensional space defined by *S* and *D* simply is a 45º rotation of the original space defined by *X* and *Y*. Obviously, the calculation of the sum or difference of two variables is not two arbitrary algebraic operations, but they define a rotation of the two-dimensional space. It can be proven that the angle of rotation, caused by addition and subtraction of the coordinates exactly is 45º.

Figure 3 shows a person located in the *X*-*Y*-system at point (6, 3) so that this person has coordinates 6 + 3 = 9 and 3 – 6 = -3 in the *S*-*D*-system. Both points define a triangle where the tangens of the angle at the origin equals the ratio of the coordinates of this point. The tangens of alpha (see figure 3) equals the ratio of the original coordinates, 6 and 3, which is 0.5, and the tangens of beta is the ratio of 9 and -3, which is -0.33. Since tangens (26,56º) = 0.5 and tangens (18.44º) = 0.33, the sum of alpha and beta exactly is 45º.

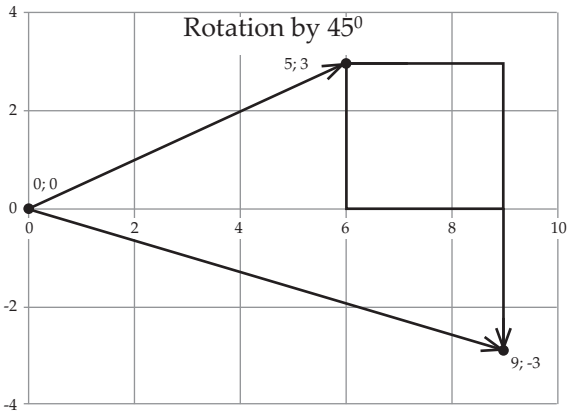

Figure 3: The rotation by 45°

This at the first glance somewhat surprising result can be proven by means of vector algebra[2]. Let A and B the vectors in the 2-dimensional space of figure 3:

$$A = \begin{pmatrix} x \\ y \end{pmatrix} \text{ and } B = \begin{pmatrix} x+y \\ y-x \end{pmatrix},$$

then the fundamental theorem about the angle between two vectors in a two--dimensional space,

$$\cos(y) = \frac{AB}{\sqrt{AA}\sqrt{BB}} \tag{1}$$

gives the following result:

$$\cos(y) = \frac{x(x+y) + y(y-x)}{\sqrt{x^2+y^2} + \sqrt{(x+y)^2 + (x-y)^2}} \tag{1}$$

$$= \frac{x^2+y^2}{\sqrt{x^2+y^2}\sqrt{x^2+y^2}\sqrt{2}} \tag{2}$$

which, again, is the cosine of 45º.

From this perspective, the calculation of the correlation between X and D relates two variables, that were taken from different rotations of the same two-dimensional space. Correlating X and D generates the problem of a negative bias that can be avoided by deciding for one of the two representations of data, the X-Y-representation or the S-D-representation. Of course, both representations can be used to describe the data, but not in the same statistical analysis (e.g. X as a predictor and D as the criterion in a regression analysis).

The observed measures X and Y define a bivariate distribution, which usually can be described by three parameters, i.e. the variances of pre- and posttest and the correlation of X and Y: var(X), var(Y), and corr(X,Y). In the case of measuring change, one is interested in the difference scores D, that are obtained by rotation of the X and Y axis of the bivariate distribution by 45 degrees. The second axis of the rotated space is defined by the sum of X and Y. The three parameters of the rotated representation, var(S), var(D), and corr(S,D), can be obtained from the original parameters by the following transformations.

$$\text{var}(S) = \text{var}(X) + \text{var}(Y) + 2\text{cov}(X,Y)$$

and $$\text{var}(D) = \text{var}(X) + \text{var}(Y) - 2\text{cov}(X,Y), \tag{3}$$

where cov(X,Y) is the covariance of X and Y. Replacing the covariances by the corresponding correlation terms,

$$\text{cov}(X,Y) = \text{corr}(X,Y)\sqrt{\text{var}(X)}\sqrt{\text{var}(Y)}, \text{ gives the results}$$

$$\text{var}(S) = \text{var}(X) + \text{var}(Y) + 2\,\text{corr}(X,Y)\sqrt{\text{var}(X)}\sqrt{\text{var}(Y)}$$

[2] My thanks to Gunnar Friege, who suggested the proof by vector algebra.

and $$\text{var}(D) = \text{var}(X) + \text{var}(Y) - 2\,\text{corr}(X,Y)\,\sqrt{\text{var}(X)}\,\sqrt{\text{var}(Y)}. \quad (4)$$

The covariance and correlation of S and D then is

$$\begin{aligned}\text{cov}(S,D) &= \text{cov}(X + Y,\ Y\text{-}X)\\ &= \text{cov}(X,Y) - \text{var}(X) + \text{var}(Y) - \text{cov}(X,Y)\\ &= \text{var}(Y) - \text{var}(X), \quad (5)\end{aligned}$$

and $$\text{corr}(S,D) = (\text{var}(Y)\text{-var}(X)) \,/\, \sqrt{\text{var}(S)}\,\sqrt{\text{var}(D)}.$$

The important conclusion of these equations is that the covariance of S and D simply is the difference of post- and pretest variance. The correlation of S and D, however, depends on the correlation of pre- and posttest which is part of the variances of S and D. The correlation of S and D, then, is zero if pre- and posttest have the same variance. The correlation can only be negative, if the pretest variance is higher than the posttest variance, and it is positive, if the posttest variance is higher.

Another conclusion is that the difference scores need not be calculated in order to compute their variance or the correlation of D and S. As a numerical example, data from the PISA 2003 study will be used.

Data example: The German PISA 2003 longitudinal science test

The following data example is taken from the German longitudinal extension of the PISA 2003 assessment (PISA-Konsortium Deutschland, 2006). In this national extension 4353 ninth grade students from the main sample were tested a second time one year later. In the science domain, the German science test (Rost et al., 2005) has been used to investigate the achievement gain of students in this population. In PISA studies the variances and correlations of the latent distributions are computed and reported, which can be taken as estimates of the true-score variances and the correlations of true-scores (Mislevy et al., 2002).

It turned out that the latent variances of pre- and posttest have the same variance, i.e. $\text{var}(X) = \text{var}(Y) = 86^2$ (Walter et al. 2006, p.112). According to the equations above, the learning gain in this population is not related to the achievement level of the students ($\text{corr}(S,D) = 0$). The correlation of pre and posttest was $\text{corr}(X,Y) = 0.78$ (p.113), which gives a variance of the gain scores $\text{var}(D) = 57^2$ (p.112) and a variance of the sum scores $\text{var}(S) = 162^2$.

A deeper understanding of these results emerges, if the formulas above are rewritten for the situation of equal variances of pre- and posttest (which is given in the present data example).

$$\text{var}(D) = 2\text{var(T)} - \text{corr}(X,Y)\ 2\text{var(T)}$$

and $$\text{var}(S) = 2\text{var(T)} + \text{corr}(X,Y)\ 2\text{var(T)}. \quad (6)$$

In the variance of difference scores, twice the test variance (7396) is reduced by the percentage of itself defined by the correlation of X and Y (0.78)

$$\text{Var}(D) = 14792\text{-}0.78*14792 = 57^2.$$

In case of the variance of the sum score, twice the test variance is increased by the same amount

$$\text{Var}(S) = 14792 + 0.78*14792 = 162^2.$$

The correlation of pre- and posttest defines the proportion of the reduction or increment of the test variances that produces the variances of the difference or sum

scores. However, the standard deviation of the sum score (162) is rather high, when the measures of pre- and posttest have a standard deviation of 86. The reason for this large value is given by the fact, that the variable S is the sum of two variables and therefore has double sized values (see Oldham, 1962, p. 973). In order to make them comparable with the original test scores, they should be divided by 2, i.e.

$$\text{var}(S/2) = ¼\text{var}(S) = 6582 = 81^2.$$

The same can be done with the difference score, because also both test scores, X and Y, contribute to the differences, i.e. each score is responsible for half the difference,

$$\text{var}(D/2) = ¼\text{var}(D) = 813 = 28{,}5^2.$$

The variances of $S/2$ and $D/2$ add to the test variance,

$$\text{Var}(S/2) + \text{var}(D/2) = \text{var}(X) \qquad (7)$$
$$6582 + 813 = 7396$$

About 89 percent (6582/7396) of the test variance contributes to the measurement of the achievement level and only 11 percent (813/7396) to the measurement of change. This relation in size makes sense, because the trait measured by this test (science literacy) has to be assumed stable, even for more than one year. The result hat 11 percent of the test variance is due to temporal fluctuations, has to be considered as a big amount (section 6 below). However, the result that this amount of individual variance of learning is not related to the level of achievement has not expected apriorily. If the (fallacious) correlation between X and D would have been calculated,

$$\text{cov}(X,D) = \text{cov}(X,\ Y\text{-}X) = \text{cov}(X,Y) - \text{var}(X)$$

and

$$\text{corr}(X,D) = (\text{cov}(X,Y)\text{-var}(X)) \,/\, \sqrt{\text{var}(X)}\,\sqrt{\text{var}(D)}, \qquad (8)$$

the result of $\text{corr}(X,D) = -0.33$ would have (mistakenly) contradicted our expectation of a positive correlation between achievement level and learning gain.

The use of X-Y- and S-D-correlations for predicting change

One reason for the persisting use of the X-D-correlation in the analysis of change data is the prediction of the efficiency of some treatment on the level of individuals or subpopulations. Such treatments may be a therapy, teaching, programs for changing attitudes or motivation, skill training or whatever. Two kinds of predictions have to be distinguished in this context.

First, a study has been performed in order to better understand the relationships between the level of achievement, motivation etc. and the amount of change of this variable. In this case, the proposed correlation of S and D certainly is an elegant solution of the problem of artificial negative correlation between pretest and change. The sum of two measures of a variable is a better estimate of the achievement level than the pretest alone.

The second case is given, if a prediction of the learning outcome has to be done on the basis of only the pretest. Such a situation is given, e.g., when persons have to be selected or assigned to some therapeutic program or training courses. Since the prognosis refers to the increment D and the sum score S is not available yet, a regression of D on the pretest X seems to be unavoidable.

In that case, the regression equation for predicting the gain score D has to be taken from an earlier study, where both measures, S and D, had been available. This correlation of S and D can be taken for predicting D on the basis of the current pretest X. The rationale is, that X is in the case of no other information the best estimate of the level measure S. Therefore, the change can be predicted by the pretest, but using coefficients that stem from an analysis of S and D. This will be illustrated by a small data example.

The artificial data example of uncorrelated X and Y (see figure 1a) has been modified, so that there is now a X-Y correlation of r = 0.707 . In fact, the scatterplot of this modified data set is exactly that of figure 1c, except the axes now are X and Y. A reasonable good prediction of Y can be made on the basis of X. However, it can be seen in figure 1c, that the variance of the posttest (which is now the vertical axis) is much higher than of the pretest (sd(X) = 1.15 and sd(Y) = 1.63, sd: standard deviation). This indicates a positive correlation of level and change, which is corr(S,D) = 0.45 in the example.

A prediction of D, therefore, would be possible on the basis of S. The regression equation is in the given example

$\widehat{D} = 0.2\ S + 0.2$,

where $\widehat{D}$ is the predicted difference. The four pretest values 1, 2, 3, and 4 get the following predictions of D: 0.4, 0.6, 0.8, and 1.0. Depending on the size of the regression weight (b = 0.2) the range of the predicted D is smaller or larger. This prediction would not be possible on the basis of the correlation of pretest and gain score, corr(X,D), which is zero in this example.

This example illustrates, that there is no need to mix variables from the two representational systems. If the posttest measure Y is not available yet for the individuals under consideration, a prognosis of D on the basis of the pretest measure X is possible. The regression equation, i.e. the regression of D on S, has to be taken from an earlier study.

Regression towards the mean

The regression towards the mean is a general principle in statistical analysis, that is not specific for the measurement of change. It simply describes the fact, that the predicted values are closer to the mean of the distribution than the observed criterion values are. This phenomenon can easily be seen from the most simple regression equation $\widehat{Z}_y = \text{corr}(x,y)Z_x$, where Z_y and Z_x are the z-transformations of predictor X and criterion Y. In some sense, the predictions underestimate high scores and overestimate low scores of the criterion.

This phenomenon has been considered to be relevant in the measurement of change, because the prediction of the post test on the basis of a pre test, produces difference scores that are negatively correlated with X (Campbell and Kenny 1999, Rogosa 1988). Low level performers get a higher (predicted) score in the post test (closer to the mean) and high level performers get a lower score in the post test, due to the regression effect. When the differences of the predicted post test and the

observed pre test were correlated with the pretest, the negative correlation between *X* and *D* can be seen again.

This regression phenomenon suggests that there is a second source that contributes to the (well known) negative correlation of *X* and *D*, and this second source is active, when predictions were made by means of regression analysis. However, this effect of making the correlation negative vanishes on the same way, as the first mechanism discussed above. If the regression is applied to the *X*-*Y*-system, then the range (and variance) of the predicted *Y*-values is smaller than of the original *Y*. This may lead to an underestimation of the correlation between level and growth, because the correlation of *S* and *D* is a simple function of the difference of post- and pretest variance (see equation (5)).

But why should the predicted y-values be used instead of the observed? It makes no sense to work with predicted *Y*-values as long as the original values are available. But if this is necessary, one should stay to the system that has been chosen, *X*-*Y* or *S*-*D*. Calculating a linear regression of *D* on *S*, the regression to the mean effect occurs, but not by lifting the poor performers higher and dropping the high performers. When *D* is the criterion variable, the variance of the estimated *D* is smaller than of the observed *D*. But the correlation of $\widehat{D}$ and *S* is not distorted by the regression to the mean.

The regression to the mean only is a problem, if the criterion *Y* is predicted by *X* and the predicted Difference score, $\widehat{D} = \widehat{Y} - X$, is taken as an estimate of the learning gain. In that case, the correlation of ^*D* and *X* is even more negative then the correlation of *D* and *X* already is. When the *S*-*D* representation is applied, none of these statistical artefacts is given.

The low reliability of difference scores

Difference scores are known to have a low reliability. This fact has led to a long debate about the question if change should be measured at all, beginning with Cronbach and Furby (1970), continued by Collins (1996), Mellenbergh (1999) and with a preliminary end by Fischer (2003). The crucial point is the distinction between reliability and precision, i.e. gain scores can be highly precise and nevertheless very unreliable, depending on the variances and covariances of the measures in the population.

Only one argument that is often used in this context, shall be addressed here. The low reliability of difference scores is attributed to the fact, that in difference scores two variables are involved, *X* and *Y*, and both variables contribute their error of measurement to the variable *D*. This is certainly true, but the sum score *S* also reflects two errors of measurement, but usually is very reliable.

Let *X'* be the true score of *X* and *Y'* the true score of *Y* and Ex and Ey their error variables, then

$$\mathrm{Var}(Y\text{-}X) = \mathrm{Var}(Y' + \mathrm{Ey} - X'\text{-}\mathrm{Ex}) = \mathrm{Var}(Y'\text{-}X') + \mathrm{Var}(\mathrm{Ey}) + \mathrm{Var}(\mathrm{Ex}), \quad (9)$$

because all covariance terms with an error variable are zero. So it is true, that the variance of difference scores covers two errors of measurement, whereas the

pre- and posttest scores cover only one. But it is also true, that the sum of two measures, e.g. $S = X + Y$, have the same amount of error variance as their difference:

$$\mathrm{Var}(X + Y) = \mathrm{Var}(X' + \mathrm{Ex} + Y' + \mathrm{Ey}) = \mathrm{Var}(X' + Y') + \mathrm{Var}(\mathrm{Ex}) + \mathrm{Var}(\mathrm{Ey}). \quad (10)$$

Therefore, within the *S*-*D*-system, the difference between the reliabilities of gain measures (*D*) and level measures (*S*) cannot be due to error of measurement. Not the error variance produces the difference in reliability, but the true score variance. The true score variance of a sum usually is higher than the true score variance of a difference. To be precise, the variance of a sum of two variables is higher than the variance of a difference, when the covariance of both variables is positive:

$$\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X,Y)$$
$$\mathrm{Var}(X\text{-}Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)\text{-}2\mathrm{Cov}(X,Y). \quad (11)$$

The variance of a difference is higher, if the covariance is negative. In the case that *X* and *Y* are measures of the same variable at two different time points, their correlation usually is high positive. But if the goal is to measure change, then pre- and posttest should not correlate very high.

As a result, *S* and *D* have the same amount of measurement error, but due to their different true score variance, their reliabilities are very different. The conclusion is rather trivial, i.e. change can only be measured reliably, if there is substantial change. But change can be measured with high precision even if the gain scores have no strong variance (Fischer 2003).

Conclusions

Some of the persisting dilemmas in the measurement of change are due to a fallacious separation of information about *gain* and information about *level* in the data. If the difference of posttest *Y* and pretest *X*, $D = Y\text{-}X$, is taken as a measure of change, then it is problematic to base the level measure only on the pretest *X*. The well known (artificial) negative correlation between difference score and initial status can be avoided when the level measure is not defined by *X*, but symmetrically, by *X* and *Y*, e.g. by $S = X + Y$. *S* operationalises the 'middle' status, instead of the initial status (as *X* does).

The two-dimensional space of *S* and *D* is a simple rotation of the *X*-*Y* space by 45° , preserving the empirical correlations between level and gain. The correlation of level (*S*) and gain (*D*) depends on the variances of *X* and *Y*, as well as on the correlation of *X* and *Y*, and can be calculated by this information without calculating the *D*-scores or *S*-scores. Both representations of change data, the *X*-*Y*-system and the *S*-*D*-system work well, as long as their variables are not mixed, as happens in the correlation of *D* and *X*. The transformation of the *X*-*Y*-system into the *S*-*D*-system is a special kind of extracting the first two, unrotated principal components in the *X*-*Y*-system.

From a state-trait perspective, both variables *X* and *Y* reflect a trait component and a state component. The trait component is stable over the two time points of measurement, the state component reflects the difference of the two measures. The mean of the two variables is a measure of what is common to both variables, the *trait*. The difference of the two variables is a measure of what is different, the *state.*

The prediction of change can be done by the mean level *S*, in case that posttest data are available. In the case of only *X*-measures have been assessed, the prediction of *D*-scores is possible on the basis of *X* only. However, the regression coefficients have to come from another data set including the posttest, where a regression of *D* on *S* has been conducted.

As a consequence of staying with one representation, *X*-*Y* or *S*-*D*, the statistical phenomenon of *regression towards the mean* has lost its distorting effect on the prediction of change. Because of the symmetry in the *S*-*D*-system, it cannot longer be argued that difference scores suffer from their double-sized error of measurement, whereas *X* or *Y* have a single error of measurement only: the sum scores *S* have the same two error variables as *D*. However, the sum scores have a higher (true score) variance than the difference scores and, hence show a higher reliability than difference scores. The variances of *S* and *D* are an empirical result and provide information about stability of the measured trait and the effects of the treatment.

Some problems of measuring change have not been addressed here, e.g. floor and ceiling effects in the pre- and posttest, or changes of the psychometric structure of the measurement instrument between pre- and posttest. But measurement of change has lost some irritating phenomena, simply by switching from *X* to *S*.

Tu and Gilthorpe (2006) presented a very detailed analysis and evaluation of the Oldham method, which is essentially the method presented here. They conculde „that Oldham's method has been misunderstood for many years" (p. 456), a statement that still can be regarded as true.

Literatura cytowana

Bereiter, C. (1963). *Some persisting dilemmas in the measurement of change*. W: C.W. Harris (ed.), *Problems in measuring change*, Madison, Univ. of Wisconsin Press.

Campbell, D.T., Kenny, D.A. (1999). *A primer on regression artifacts*. New York Guilford.

Collins, L.M. (1996). Is reliability obsolete? A commentary on 'Are simple gain scores obsolete?', *Applied Psychological Measurement*, 20, 289-292.

Cronbach, L.J. and Furby, L. (1970): How should we measure change – or should we? Psychological Bulletin 74, 1, 68-80.

Fischer, G.H. (2003). The Precision of Gain Scores Under an Item Response Theory Perspective: A Comparison of Asymptotic and Exact Conditional Inference About Change. *Applied Psychological Measurement*, *vol. 27, 1, pp. 3-26*.

Harris, C.W. (Ed.). *Problems in measuring change*. Madison, WI: The University of Wisconsin Press.

Mellenbergh, G.J. (1999). A note on simple gain score precision. *Applied Psychological Measurement*, 23, 87 – 89.

Mislevy, R.J., Beaton, A.E., Kaplan, B. and Sheehan, K.M. (1992). Estimating population characteristics from sparse matrix samples of items. *Journal of Educational Measurement*, 29, 133-164.

Rogosa, D. (1988). Myths about longitudinal research. W: K.W. Schaie, R.T. Campbell, W. Meredith, S.C. Rawlings (ed.): *Methodological issues in aging research*, New York: Springer, 171-209.

Rogosa, D.R., & Willett, J.B. (1983). Demonstrating the reliability of the difference score in the measurement of change. *Journal of Educational Measurement*, 20, 335-343.

Rost, J., Walter, O., Carstensen, C.H., Senkbeil, W., Prenzel, M. (2004). Naturwissenschaftliche Kompetenz. W: PISA-Konsortium Deitschland (eds.): PISA 2003 – Der Bildungsstand der Jugendlichen in Deutschland – Ergebnisse des zweiten internationalen Vergleichs. Münster: Waxmann, 111-146.

Wainer, H., Brown, L.M. (2004). Two statistical paradoxes in the interpretation of group differences: Illustrated with medical school admission and licensing data. *The American Statistician*. 58, 117-123.