Measuring the Reliability of Picture Story Exercises like the TAT

Nicole Gruber and Ludwig Kreuzpointner

Universität Regensburg

Author Note

Nicole Gruber, Department of Psychology, Universität Regensburg; Ludwig Kreuzpointner, Department of Psychology, Universität Regensburg.

Correspondence concerning this article should be addressed to Ludwig Kreuzpointner, Chair of Psychology VI, Department of Psychology, Universität Regensburg, Universitätsstraße 31, D-93053 Regensburg, Germany. E-mail: ludwig.kreuzpointner@ur.de

Abstract

As frequently reported, psychometric assessments on Picture Story Exercises, especially variations of the Thematic Apperception Test, mostly reveal inadequate scores for internal consistency. We demonstrate that the reason for this apparent shortcoming is not caused by the coding system itself but from the incorrect use of internal consistency coefficients, especially Cronbach's α. This problem could be eliminated by using the category-scores as items instead of the picture-scores. In addition to a theoretical explanation we prove mathematically why the use of category-scores produces an adequate internal consistency estimation and examine our idea empirically with the origin data set of the Thematic Apperception Test by Heckhausen (1963) and two additional data sets. We found generally higher values when using the category-scores as items instead of picture-scores. From an empirical and theoretical point of view, the estimated reliability is also superior to each category within a picture as item measuring. When comparing our suggestion with a multifaceted Rasch-model (Blankenship et al., 2006; Tuerlinckx et al., 2002) we provide evidence that our procedure better fits the underlying principles of PSE.

Keywords: reliability, projective test, Thematic Apperception Test, Picture Story Exercise, internal consistency

Measuring the Reliability of Picture Story Exercises like the TAT

Many psychological constructs cannot be measured directly. In the classical test theory (e.g. Lord & Novick, 1968) each observed score is decomposed into a true and error score. To examine the reliability of a test —one of the central criterions of its goodness— many methods were developed. If a test measures a time-stable construct, the score achieved in the first session should not differ from the score in the second session (retest reliability). If a test contains items all measuring the same construct, these items should be highly statistically related (esp. method of split-half or internal consistency).

When these general methods of calculating reliability are used for projective tests, they mostly yield unacceptable scores. One well known projective measure is the Thematic Apperception Test (TAT), by McClelland since 1989 mostly called Picture Story Exercise (PSE). Participants view some pictures, each for about half a minute and are then instructed to write a short story about it by answering some leading questions. Within about five minutes they have to respond. The central assumption of a PSE is that participants identify themself with the protagonist of the picture when writing the story and thus project their own needs into their story. Those stories are coded for implicit motives using a special coding system. This coding technique of PSE has been widely used on different versions. Most common is the measure of the need for achievement (Heckhausen, 1963; McClelland, Atkinson, Clark & Lowell, 1958, Winter, 1994). Heckhausen's PSE (1963; English language translation by Schultheiss, 2001) assessed two components of need for achievement separately: (1) hope of success (HS) and (2) fear of failure (FF). Heckhausen stated this two components as interrelated with each other. He calculated a "net hope score" (NH) as HS – FF and the "resultant achievement motivation" as HS + FF. But recent studies and theories (e.g. the quadripolar model of Covington & Roberts, 1994) imply the distinction of the two components.

Even though the achievement TAT is a well-researched and empirically validated assessment, its reliability has often been criticized. For example, Entwisle (1972) stated in a review about PSE (or like she called: fantasy-based measures of achievement motivation) that the internal consistency “rarely exceeds .30 to .40” (p. 377), but listed a few results with obvious higher as well as obvious lower values. For retest reliability she found studies with values about .30 and lower. When equivalent forms were used the values were mostly higher (>.50). Schultheiss and Pang (2007) observed a loglinear degression of retest reliability for the time between two measurements and concluded for the one day a mean stability coefficient of .71, .60 for a week until .25 for 10 years. They stated that the retest reliability of TAT mostly is in an acceptable range. Lundy (1985) also assessed the reliability of two PSE, found retest reliabilities “in the same range as those of these [MMPI, CPI and 16PF] three popular and representative objective personality tests” (p. 143) of .48 and .56, but alphas of .32 and .31 for the first and -.18 and .22 for the second measurement a year later. He conceded (p. 144): “The inevitable conclusion is that the assumptions of classical psychometrics are not met with TAT, and that alpha is therefore an inappropriate measure for this test.” Current researchers such as Tuerlinckx, De Boeck and Lens (2002) have “accepted the unreliability of TAT” (Blankenship et al., 2006, p. 100). But claiming PSE as “test-theory free” because of low reliability scores is no solution, indeed it shows that reliability calculations for projective tests have always been a big problem.

McGrath and Caroll (2012) reported in their critical review about PSE low internal consistency and retest stability but an adequate inter-rater reliability. But inter-rater agreement is not a measure of reliability in the context of the classical test theory, it is a prerequisite of reliability because the measure indicates the independence of the results from the persons who scored the results (i.e. objectivity). We focus in this article on the internal consistency of the PSE. Therefore we first review the Coefficient α by Cronbach (1951) and the six lambdas of Guttman (1946). Then we introduce a new reliability calculation using the categories instead of the picture-scores. We contrast this measure with calculations on dichotomous item-level data. We also examine whether Rasch-scaling is appropriate for PSE. Finally we demonstrate empirically on three data sets which internal consistency method best fits the Heckhausen PSE.

Internal consistency

Cronbach (1951) emphasised that by only demonstrating whether two halves of a test are consistent with each other, not all possible variations are examined. To assess the consistency of all items, he constructed the α Coefficient, which is one of the most frequently used measures for internal consistency. One possible reason for its wide use could be that historically it was easy to compute and the measure is a perfect fit to self-rating questionnaires with a high number of similar items. However, α could inflate the reliability of a test, especially self-rating scales, because people like to reflect a consistent self-concept (Brunstein & Schmidt, 2004). But if the items are not equivalent or even if they are heterogeneous, α can produce misleading reliability scores and therefore should not be used. Rae (2007) discussed the problem of α and stated that the assumption for its use "implies that every person's true score on any given component differs from his or her respective true score on any other component by only an additive constant" (p. 177). Borsboom (2005) queried the correct usages of the true score concept in most measurement research at all. When he termed the concept of lower bounds as "probably the most viable defence that could be given for the standard practice in test analysis" (p. 30), he questions a procedure which is in use more than 70 years. Six years before Cronbach (1951), Guttman (1945) proposed six coefficients for internal coefficients and established the concept of lower bounds. Guttmanns λ_3 is the exact equivalent of Cronbach's α. Guttman started with λ_1, which is very similar to λ_3, but the calculation did not include the number of items. As an improvement, he included the numbers of items as well as the covariances in the λ_2 coefficient. Additionally he developed as a short version of λ_2 λ_3, because it "is easier to compute than λ_2" (p. 274) by ignoring the covariances. However therefore, two prerequisites for its use are strict homogeneity and positive covariances. In all other cases, Guttman (1945) suggested to use λ_2, despite the increased computational requirement. Three further coefficients were developed: λ_4 is a measure for split-half reliability for which covariance is not calculated, λ_5 is another measure developed for the case that one item has "large absolute covariances with the other items compared with the covariances among those items" (p 277). λ_6 is a measure when data fit to regression model like McClelland (1985) assumed for PSE by using the multiple regression error variance instead of the item variances. Although Fleming (1982) by referencing Lundy (1985; Fleming cited an unpublished version of 1980) suggested the assessment of the reliability of PSE in using linear regression, a calculation of λ_6 for a PSE was not findable.

Revelle and Zinbarg (2009) subsume the discussion about the use of several coefficients as lower bound of internal consistency. They recommend to use ω_t (McDonald, 1999), especially in "contexts, such as applied prediction, in which we are concerned with the upper bound of the extent to which a test's total score can correlate with some other measure and we are not concerned with theoretical understanding regarding which constructs are responsible for that correlation" (Revelle & Zinbarg, 2009, p. 152). In case of a unidimensional construct $\omega_t = \alpha$. When the goal is to assess "the degree to which the total scores generalize to latent variable common to all test items" (p. 152), the use of ω_h is more appropriate. For PSE all of the above named methods can be applied. But this does not mean that they are all appropriate.

According to the Dynamics of Action Theory (DoA), Atkinson and Birch (1970) described the problem when using a PSE that a trait (the motive) can only be measured by the situational state (the motivation) which has been known to fluctuate for several reasons. An inherent response behaviour to the pictures is the lowering of the need for achievement activation force through writing an achievement-thematic story. The need to write about an achievement topic in the next picture (the next item) decreases. Consequently, the progress will be up and down, especially for highly motivated people. McClelland (1980) referred to Atkinsons doctoral thesis since then this “so-called sawtooth effect in the achievement content of successive stories has been known” (p. 31), when people do not write the same or similar story just because of the instruction to be creative. Consequently this effect leads to low values of statistical indices of internal consistency. So Atkinson, Bongort and Price (1977) tested their theoretical assumption with computer simulations and found according to their hypothesis that the high criterion validity of TAT was consistent with very low and even negative reliability scores. Reumann (1982) hypothesized that ipsative variability, which is associated with low internal consistency, will increase the criterion validity (assessed with an arithmetic task) of the motivational imaginary story. The results revealed an outlandish internal consistency of -1.23 (assessed with Coefficient α) referred to a good criterion validity of .62. Reumann suggested that calculating internal consistency using α would not be effective, because he expected this measure in a well-constructed PSE to become infinitely negative.

Tuerlinckx et al. (2002) also tested the theoretical assumptions of Atkinson and Birch (1970), but could not validate them. They found that some pictures stimulate a high achievement motive and some do not but no evidence was found to explain why. Thus, the result best fits a model of spontaneous-drop-out, which was later theoretical explained by Schultheiss, Liening and Schad (2008) using the Cognitive Affective System Theory (CAST; Mischel & Shoda, 1995). This theory offers an explanation for finding which disagree with the drive reduction suggested by the DoA theory and is very similar to the explanation provided by McClelland (1980): People learn to satisfy their needs in different situations throughout their life. So some people think of an instructor and worker when they see two men standing on a workbench, others think of father and son or two friends drinking beer. This suggests that each score results from an interaction between the picture-cue and the personal background of a person which cannot be controlled (another reason for fluctuation).

This unpredictable change of item difficulty is an immense problem for the calculation of the reliability, because all unpredictable changes serve as measurement error. Another problem is that not all pictures correlate highly and positively with each other. Every picture can stimulate the motive in a different way, which leads to completely different stories to the extent that they correlate negatively. Moreover, α increases with the number of items, but PSE comprises few items, because people get tired after more than six pictures (McClelland, 1985).

In sum, we state that α is not an appropriate reliability coefficient when using the sum of occurring categories of each picture as items of a PSE, because the items (picture-scores) are inhomogeneous. We provide another approach of calculation to eliminating the inhomogeneity.

Category vs. picture reliability

Therefore, we introduce an idea that eliminates the inhomogeneity by taking a closer look at the internal consistency of the coding system. The scores of pictures are always related to the underlying coding system, but after thorough review of the literature the reliability of a coding system has not been assessed in any study according to the reliability of the projective test. The only exceptions are Kuhl (1978), who assessed reliability in the context of Rasch-scaling methods, and Fleming (1982) by mention the possibility to use the scores of categories for regression equations. Our idea is to use the categories instead of the pictures as items. For example, when calculating the reliability of the *hope of success* scale, the scores for each of the six pictures are not used but the scores of the six categories. Each item consists of the number of pictures which fits the criterions of the category. The overall participant score remains the same. Generally, we assume that calculating reliability using categories instead of pictures as corresponding items would be a much more adequate measure for internal consistency of PSE. The categories of the coding system are constructed to correlate positively and to be homogeneous. Participants with a high need for achievement are expected to write more elements which fit the criteria of the categories. Though the influence of the length of the stories of a subject, which affected the motivescore - e.g. Pang and Schultheiss (2005) found a correlation of .23 -, has less impact for the estimation of the reliability. Thus, the relevance of the saw-tooth-effect according to the DoA or the picture cue effects as specified in the CAST will be minimized. To shortly explain this with an example data-matrix (see table 1).

%% table 1 about here %%

This is a very constructed and shorten data-matrix of a PSE data set. We just use three categories (Cat 1, Cat 2 and Cat 3) and three pictures (A, B, C). This way the data of PSE can be seen as a two-level matrix consisting of 0 and 1, whereby categories are nested within the pictures. As for the first three subjects the sum of pictures and the sum of categories are equal, for the second three subjects the equal scores within the three categories leads to different scores for the pictures. So there are high intercorrelations for the categories and low intercorrelations for the picture-scores (see table 2).

%% table 2 about here %%

This statement is also checked mathematically by reviewing the formulas provided below. Strongly simplified, the internal consistency measured with α can be seen as a relationship of test variance (V_t) and item variance (V_i) (Cronbach, 1951, p. 304 (13)):

$$\alpha = \frac{n}{n-1}\left(1 - \frac{\sum V_i}{V_t}\right) \tag{1}$$

Note: i is counter of *n* items.

An obvious feature of equation 1 is that regardless of using categories or picture-scores, the denominator will be the same because the test variance is the same in both cases. Hence, there are only two reasons why reliability calculated over categories would be higher than reliability calculated over pictures. First, the item variance for categories is lower than for pictures, which makes sense if we follow the assumptions of Atkinson and Birch (1970) or of Schultheiss et al. (2008) that each picture stimulates the motive to varying extents. But all categories are always related to the same criteria, thus the variance of categories should be lower. Second, higher covariances are expected when using category-scores instead of picture-scores. If the coding system is valid, all categories should positively correlate. We can neither hypothesize it for pictures (e.g. DoA) nor observe it (e.g., Reumann, 1982).

Given the denominator, the difference between these two types of reliability measures depends on the numerators. We found a direct connection of internal consistency calculated from pictures and from categories. We first decomposed the variances of the internal consistency numerator calculated over pictures. The variance of a sum is the sum of the summand variances and each summand pair covariance (e.g. Kenney & Keeping, 1951, p. 72, (4.32)). To calculate variance and covariance, we can use equation 22 and 24 by Guttman (1946; p. 269): $(\sigma_{sj})^2 = E(x_{ijk}-\mu)^2$ and $\gamma_{xgxj} = E(x_{igk}-\mu_g)(x_{igk}-\mu_j)$ respective, where E is the expected value that can be estimated as sum of all components divided by *n*.

To simplify and because the actual measure of variability matters little for our suggestions, we preferred the sum of squares (SS) instead of the variance of the picture-scores (Var_p):

$$SS_p = n \cdot Var_p = \sum_{p=1}^{n} (x_p - \bar{x}_p)^2 = \sum_{j=1}^{N} \sum_{p=1}^{P} \sum_{c=1}^{C} \left(x_{jpc} - \bar{x}_{pc}\right)^2 + 2\sum_{j=1}^{N} \sum_{p=1}^{P} \sum_{c=1}^{C-1} \sum_{b=c-1}^{C} \left[\left(x_{jpc} - \bar{x}_{pc}\right)\left(x_{jpb} - \bar{x}_{pb}\right)\right] \quad (2)$$

Note: j is counter of N subjects, *p* is counter of P pictures, *c* is counter of C categories.

The picture SS is the sum of all category SS by picture-scores and the sum of all sums of products of each category pair mean deviation within each picture. Thus, for example, when calculating the SS of picture A (resp. *p* = 1) each category SS, starting from A1 (category 1) to A6 (category 6) and the sum of products of each category pair (A1, A2) to (A5, A6) are summed. See Appendix A for a detailed mathematical proof of this decomposition.

We conclude from this computation that the picture-scores variances depend on the sum of their sub-variances (the variances of all item points x_{pc}) and on the sum of their covariances. But this sum of covariance will be high when categorical-covariance is generally high, which in turn leads to high categorical reliability (as can be seen with equation 1).

Reconsidering that the sum of all pair covariances is the total-test-variance minus their variances leads to the following equation for the covariance of the picture-scores:

(3a)

$$SS'_p = n \cdot Cov_p = \sum_{j=1}^{N} \sum_{p=1}^{P-1} \sum_{a=p+1}^{P} \sum_{c=1}^{C} \left[\left(x_{jpc} - \bar{x}_{pc}\right)\left(x_{jac} - \bar{x}_{ac}\right)\right] + \sum_{j=1}^{N} \sum_{p=1}^{P-1} \sum_{a=p+1}^{P} \sum_{c=1}^{C-1} \sum_{b=c+1}^{C} \left[\left(x_{jpc} - \bar{x}_{pc}\right)\left(x_{jab} - \bar{x}_{ab}\right)\right]$$

Note: j is counter of N subjects, *g* is counter of P pictures, *c* is counter of C categories.

and also to this equation for the covariance of category-scores:

$$SS'_c = n \cdot Cov_c = \sum_{j=1}^{N} \sum_{c=1}^{C-1} \sum_{p=1}^{P} \sum_{b=c+1}^{C} \left[\left(x_{jpc} - \overline{x}_{pc} \right) \left(x_{jpb} - \overline{x}_{pb} \right) \right] + \sum_{j=1}^{N} \sum_{p=1}^{P-1} \sum_{a=p+1}^{P} \sum_{c=1}^{C-1} \sum_{b=c+1}^{C} \left[\left(x_{jpc} - \overline{x}_{pc} \right) \left(x_{jab} - \overline{x}_{ab} \right) \right]$$

(3b)

Note: j is counter of N subjects, *g* is counter of P pictures, *c* is counter of C categories.

See the appendix for a detailed explanation.

Equations 3a and 3b differ only in the first term. Thus, if reliability of categories is different from the reliability of pictures, the difference issue from the first term in both equations. But this term is a component of the variance of the other one. For example, the first term in equation 3b (covariance of categories) is similar to the second term in equation 2 (variance of pictures). This finding can be a solution for the main reliability problem and the validity-reliability dilemma of projective tests. The higher the variance of the picture-scores the higher are the covariance of the categories score and the higher are the difference of the α coefficients calculated with pictures vs. calculated with category-scores (see appendix, equation 6). This indicates that the internal consistency of a valid PSE, if it was calculated with category-scores, would ever be higher as the internal consistency calculated with picture-scores. In sum, calculating reliability using categories instead of pictures would be an optimum solution. In the next step, we contrast this new method with other ideas such as using dichotomous measures (Jensen, 1959) or using item response theory for calculating reliability (Blankenship et al., 2006).

Measuring internal consistency of PSE as dichotomous data

Before calculating internal consistency using α, Kuder-Richardson Formula 20 (KR-20) was a common reliability measure used on each dichotomous single scoring unit (Kuder & Richardson, 1937). Jensen (1959) wrote about KR-20 that “still it is probably the best estimate possible of the internal consistency reliability of the TAT” (p. 123). Currently, support of this type of thinking is lacking, yet it is worthy of discussion. The negative interaction of reliability calculated using the picture-scores will be deleted by breaking down the items to the least possible scoring unit, which leads to a matrix only consisting of 0 and 1.

The advantage of calculating the internal consistency of a PSE using the categories for each picture as dichotomous items is that the value of the reliability coefficient is increased because of the higher number of items. We assume that other factors also contribute to the increase, such as higher homogeneity as a result of the high inter-correlations of the underlying categories, and also their high correlations with the overall score is an important factor (Jensen, 1959). Thus, calculating reliability on a dichotomous level will always be influenced by internal picture and categorical consistency such that high categories-as-items reliability stands in relation with low pictures-as-items reliability (see equation 6). Therefore, we can follow that in a well-constructed PSE the yield score for reliability will always be confounded by the low interaction of picture-scores. Hence, we suppose that reliability calculated on a dichotomous level is still influenced by the inhomogeneity of the picture-scores.

We consider that all measures on a dichotomous level are influenced by the individual style of crossing motive (e.g., someone likes to write more feelings, another one likes to write more instrumental activity) and on the picture which evokes the motive (e.g., someone writes a motive consisting of a story on picture one and not on picture two). But measures on a dichotomous level are based on the assumption that all items are positively correlated with each other.

With PSE generally needs are measured, but the expression of these needs could change during test situation. As it will be shown in our investigation, calculating on a dichotomous level is also influenced by these effects, but when calculating internal consistency using categories as items the problem can be solved. The scores of the categorical system are independent of the picture from which they come. If for example a person's respond fits the category "instrumental activity" for hope of success in picture A and another person's respond was influenced the same way but at picture C, because each picture reflects the individual life story of the respective reader (CAST), the categories score will not differ in consequence of that. Likewise the saw-tooth-effect assumed in the DoA theory, when the drive of writing achievement related statements is satisfied in picture B but perhaps high when writing stories for pictures A and C, will be under control when using the categories score for calculating reliability. Here it does not matter which picture story fits the criteria. So, in our opinion, the calculation of internal consistency using categories should be preferred.

Rasch-Model for higher reliability

As the last analysis, we assessed whether the assumptions of Rasch-modelling have an advantage for the estimation of PSE reliability. Tuerlinckx et al. (2002) discussed the possibility of subjecting PSE to Rasch-scaling, assuming that the tendency of giving an achievement relevant answer on each picture (scored with 0 or 1) depends on the strength of the motive of a person and the instigating force of the picture. After testing many Rasch-models, they concluded that PSE best fits a spontaneous drop out model for which some pictures force motive and some do not. Thus, the drop out hindered reliability, which in their opinion could only be solved when increasing the numbers of pictures — an option that they rejected because of practical reasons.

Likewise, Blankenship et al. (2006) found the solution of the reliability problem in using a Multifaceted Rasch Model, which is able to control confounders like the influence of the coder. Blankenship and colleagues tried to improve the test and its reliability by identifying new pictures for a better model fit and higher reliability scores. They found Cronbach's α of .78, .70 and .69 and a Person Separation Reliability (PSR), which is a Rasch-equivalent of α or KR-20 as stated by Linacre (2005 cited by Blankenship et al., 2006), between .24, .56 and .75, but they found a heterogenic result. Consistent with our theoretical suggestions, we agree with Kuhl (1978) and reject the notion that Rasch-modeling would be the best method for measuring the reliability of projective tests, because the theory and problems underlying these tests do not fit with the assumptions of local stochastic independence of the Rasch-model. Only because “the scoring criteria were […] applied independently to each PSE” as argued by Blankenship et al. (2006, p. 101), this method does not suggest that the items (i.e., the pictures) have no relationship to each other. We turn again to the DoA: If a high motive were to be stimulated by the first picture, the answer for the second picture could be based on a lower strength of the need. According to the CAST, this effect of different scores in different pictures is contingent not on the order of the pictures and a reduction of the motive drive but on their content and the interaction with the subject biography. An additional criticism of the Rasch-model is that according to Tuerlincks et al. (2002) and Blankenship et al. (2006) each picture is seen as an item, which we have shown to be the least optimum basis for measuring internal consistency. This procedure is particularly problematic when, for example, Tuerlinckx and colleagues used pictures only scored as 1 or 0. Such a procedure is not consistent with the theoretical conception of McClelland et al (1958) or Heckhausen (1963) who developed this assessment.

Expectations

Based on the arguments and the procedure that we proposed, we can formulate the following two expectations:

- Measuring reliability using category-scores will outperform methods using picture-scores as items. We should find support for this preference, because category-scores are not hindered by effects of the DoA or CAST as are picture-scores.
- Measuring reliability using categories will also be higher than measuring on the dichotomous level. Measuring on the dichotomous level is influenced by both category-

scores and picture-scores. Therefore, the saw-tooth-effect and/or the picture-cue-effect are expected to influence this type of measure.

Methods

Participants

We tested our hypothesis first with the data set of N = 35 PSE given by Heckhausen (1963) presented in his coding-manual[1], because we assume them to be most valid. Second, we used the PSE of N = 113 university students (67 female; age range 19 to 42 years; M = 23.60, SD = 3.00)[2]. Additionally, we were able to use the data set of Breidebach (2012) with N = 241 pupils of a vocational school (103 female, age range 15 to 23 years, M = 17.65, SD = 1.63)[3].

Materials

For our investigation we used the PSE of Heckhausen (1963). Heckhausen (1963) used six pictures describing a smiling man at the desktop (picture A), a man in front of the directors room (B), two men on a workbench (C), a pupil on a blackboard (D), a man at a desktop (E), two men on a machine (F), whereby three of them mainly activate hope of success (A, C, E) and three activate fear of failure (B, D, F). After having a look at the picture for 20 seconds, the subjects were instructed to answers the four questions: 1. What is going on? Who are the people? 2. What has led to this situation? What has happened before? 3. What are the people thinking about, feeling, or wanting? 4. What will happen next? How will everything turn out? For each question one minute was given. After four minutes the subjects could correct their answers for a further minute.

[1] Download: Heckhausen.dat
[2] Download: Students.dat
[3] Download: Breidebach.dat

The stories were coded with the Heckhausen coding system (1963; English language translation by Schultheiss, 2001). This coding system consists of five main categories for hope of success (HS) and six main categories for fear of failure (FF) and one weighting category for each. The main categories for HS are: expression of the need for achievement and success (NS), instrumental activity to achieve success (IS), expectation of success (ES), praise (P) and positive affect (A+). For FF the main categories are need to avoid a failure (NF) and instrumental activity to avoid failure (IF), expectation of failure (EF), negative affect (A-), criticism (C) and failure (F). When the story of a picture fits the criteria of a category, one point was given, otherwise a score of 0. For the picture-scores the points were summed up. An additional point was given, when a story is primarily "success-seeking" (ST) or "failure-avoiding" (FT). The success theme is given when NS or ES are scored and no failure category excepting A- and EF. The failure theme is given when NF and F are scored and no success category excepting IS (Schultheiss, 2001).

Analysis

For testing our hypothesis we calculated Guttman's λ_1 to λ_6 with SPSS and McDonald's ω_t with the R psych package for both, category and picture-scores, and the dichotomous data[4]. But before analysing we proofed the inter-rater-agreement of each two trained coders assessed with the a_d-coefficient by Kreuzpointner, Simon and Theis (2010) and Pearson correlations (given in brackets). In our data set a_d was .998 for HS (r = .90) and .998 for FF (r = .87), which is in both cases above the 95 % level. The inter-rater-agreement for the data of Breidebach (2012) is also very well: a_d was .999 for HS (r = .96) and .999 for FF (r = .97).

Results

[4] The psych package uses correlation instead of covariances, which do not conform exactly to our equations.

Table 3 reveals the striking finding that using categories instead of pictures as items leads to higher scores. HS reliability measures using pictures as items were $\alpha = .22$ (.12 in the student sample and .47 in the pupil sample) but with categories as items α increased to .48 (.52 in the student sample and .67 in the pupil sample). The same increase was found for FF, especially in the origin data sample of Heckhausen using categories instead of pictures which lead to an increase from a negative α (-.02) to .60. Moreover, the λ_5 reliability coefficient for HS calculated using categories was .61 (in both samples, .68 in the pupil sample), which was higher than the coefficients using pictures (.36 in the Heckhausen sample, .22 in the student-sample and .50 in Pupil sample). The same preference for categories was found for FF. λ_5 calculated for pictures was .20 in both investigations (.40 in the pupil sample), which was lower than the coefficients when calculating it for categories (.65 in the Heckhausen sample, .51 in the student sample, and .71 in the pupil sample). Regardless of the score used, the calculated reliability coefficients of λ_2 and ω_t for pictures never outperformed the coefficients calculated using categories. When having a look on the reliability scores without considering the weighting categories (ST and FT, given for stories which fitting the motive very well), we still found that the reliability coefficients calculated using categories to be higher to those using pictures (e.g. .43 for FF in the Heckhausen sample calculated by categories vs. -.05 calculated by pictures). Generally the values of the coefficients for the setting without ST and FT are mostly lower but especially for the student sample some scores are even higher.

%% table 3 about here %%

To prove if the higher values of internal consistency result from the higher intercorrelations of the categories the intercorrelations for the Heckhausen data is given in table 4.

%% table 4 about here %%

For HS (above the diagonal) the correlations of the categories are not as clearly higher as expected compared with the correlations of the picture-scores. But for FF it can be observed. On the other hand the mean (via Fisher-transformation) correlation of .03 for the HS picture-scores is clearly lower than the mean correlation of .12 for the HS category-scores. As the mean FF picture-scores correlation is .00, the mean correlation of the FF category-scores is .20. Similar results can be observed for the two other data sets.

The values of the reliability coefficients calculated on a dichotomous level are similar to the values observed for category-scores (see table 5). On this dichotomous level ω_t should be able to be calculated using a standard algorithm as an approximation. But for an exact assessment nonlinear factor models are required (McDonald, 1999, p. 102f). Both options were not available in all R-packages that we reviewed.

%% table 5 about here %%

We expected that the reliability estimated with dichotomous data would be influenced by the pictures score and the categories score reliability. Thus, this value was expected to be between category and picture reliability. The sample of Heckhausen confirmed our assumption for FF (α: .60 > .52 > -.02) but not for HS. In contrast, the pupil sample confirmed the assumption for HS (α: .76 > .70 > -.47) but not for FF. Neither pattern was found in the sample of students for HS or FF.

Conclusion and prospects

Calculating reliability of PSE has long been noted as a persistent problem, which we contend has been independent of the test: The problem was the result of treating this method as a self-report-measurement, but Picture Stories Exercises are different. The underlying phenomena, explained in the Dynamics of Action theory as saw-tooth-effect and in the Cognitive Affective System Theory as picture-cue-effect, decrease the homogeneity of items. This effects, however, does not negatively impact the coding system. Investigating the reliability of the tests on the basis of the coding system can provide a solution. We found evidence to confirm this hypothesis in three different data sets. On the one hand there are clear higher intercorrelations when the category-scores used as items compared to the picture-scores (table 4) and on the other most of and especially the preferred coefficients for internal consistency λ_2 and ω_t are higher for category-scores. In future studies the hypothesized relationship between reliability calculated using pictures or categories should be assessed with Monte Carlo simulation to confirm our theoretical assumptions and further demonstrate the superiority of calculating reliability coefficients using coding categories instead of pictures.

We strongly advise to refrain from using the α coefficient on the basis of picture-scores because of two main reasons. First, pictures are compromised by the saw-tooth-effect and/or the picture-cue-effect. Second, α is an appropriate measure for homogenous data, but not for projective tests such as PSE. λ_2, λ_5 and ω_t are more appropriate measures, because they better fit the theoretical concept of projective tests. We also dissuade from using Rasch-scaling for dichotomous data to estimate PSE reliability, because the prerequisites of stochastic independence cannot be fulfilled and the procedure does not fit the theoretical concept of PSE. On the other hand, the item response theory for ordinal data (for the category-scores) could be worth to examine in further research as a possible adequate measurement model for PSE and projective tests.The results of our study are limited to the PSE and the coding system of Heckhausen (1963). Regarding the possible dissent that categorical reliability is only higher because of the weighting categories we have shown that in both conditions, with and without weighting categories, categorical reliability always outperforms pictorial reliability. For the weighting categories do not only depend on the positive categories but also on the absence of negative categories, it is not just a lifting effect as the results for the student sample accessorily clarified. Further research is needed to replicate the effects on different projective tests, different coding systems, in different countries, and both clinical and nonclinical groups. Our method can also be adapted to other verbal-thematic projective tests for which stories or statements are produced in response to a picture and then coded by a categorical system. For example, the Fairy-Tale Test (FTT; Coulacoglou, 2008) and the Rosenzweig Picture-Frustration Test (PFT; Rosenzweig, 1945) and all modifications of TAT and PSE based on a categorical system are possible. Applying the method to sentence- and story-completing tests and drawing tests would also be appropriate, when there is a categorical system. Researchers using these tests could benefit from our method; hence further investigations are needed in this area.

Acknowledgements

We are grateful to Oliver C. Schultheiss and William Revelle for their reviews and critical comments to the earlier versions of this paper. We also want to thank Christina Hanauer for coding the data of the student sample and Guido Breidebach for the leaving of the data of his pupil sample.

References

Atkinson, J. W. & Birch, D. (1970). *The dynamics of action*. New York: Wiley.

Atkinson, J. W., Bongort, K. & Price, L. H. (1977). Explorations using computer simulation to comprehend thematic apperceptive measurement of motivation. *Motivation and Emotion*, *1*, 1-27. doi: 10.1007/BF00997578

Blankenship, V., Vega, C. M., Ramos, E., Romero, K., Warren, K., Keenan, K., … Sullivan, A. (2006). Using the multifaceted rasch model to improve the TAT/PSE measure of need for achievement. *Journal of Personality Assessment, 86*(1), 100–114. doi: 10.1207/s15327752jpa8601_11

Borsboom, D. (2005). *Measuring the mind: Conceptual issues in contemporary psychometrics.* Cambridge: University Press.

Breidebach, G. (2012). *Bildungsbenachteiligung – warum die einen nicht können und die anderen nicht wollen [Educational disadvantage – why some can't and others don't want].* Hamburg: Kovacs.

Brunstein, J. C. & Schmidt, C. H. (2004). Assessing individual differences in achievement motivation with the implicit association test. *Journal of Research in Personality, 38*(6), 536-555. doi: 10.1016/j.jrp.2004.01.003

Cramer, P. (1999). Future directions for the Thematic Apperception Test. *Journal of Personality Assessment, 72*, 74-92. doi: 10.1207/s15327752jpa7201_5

Cronbach, L. J. (1951). Coefficient α and the internal structure of tests. *Psychometrika, 16*(3), 297-334. doi: 10.1007/BF02310555

Coulacoglou, C. (2008). *Exploring the child's personality: Developmental, clinical, and cross-cultural applications of the Fairy Tale Test*. Springfield, IL: Charles C Thomas Pub LTD.

Covington, M. V. & Roberts, B. W. (1994). Self-worth and college achievement: Motivational and personality correlates. In P. R. Pintrich, D. R. Brown & C. E. Weinstein (Eds.), *Student motivation, cognition and learning* (pp. 157-188). Hillsdale, NJ: Erlbaum.

Entwisle, D. R. (1972). To dispel fantasies about fantasy-based measures of achievement motivation. *Psychological Bulletin*, *77*(6), 377. doi: 10.1037/h0020021

Fleming, J. (1982). Projective and psychometric approach to measurement: The case of fear of success. In A.J. Stewart (Ed.), Motivation and society. A volume in honor of David C. McClelland (pp. 63-96). San Francisco: Jossey-Bass.

Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika, 10*(4), 255-274. doi: 10.1007/BF02288892

Heckhausen, H. (1963). *Hoffnung und Furcht in der Leistungsmotivation.[Hope and fear in achievement motivation].* Meisenheim am Glan: Anton Hain.

Jensen, A. (1959). The reliability of projective techniques: Review of the literature. *Acta Psychologica, 16,* 198-136. doi: 10.1016/0001-6918(59)90089-7

Kenney, J. F. & Keeping, E. S. (1951). *Mathematics of statistics, part 2.* Princeton, NJ: Van Nostrand.

Kuder, G. F. & Richardson, M. W. (1937). The theory of the estimation of test reliability, *Psychometrika, 2*(3), 151-160. doi: 10.1007/BF02288391

Kuhl, J. (1978). Situations-, reaktions- und personenbezogene Konsistenz des Leistungsmotivs bei der Messung mittels des Heckhausen-TAT [Situation-, reaction- and personbased consistence of the achievement motive measuring with Heckhausen-TAT]. *Archiv für Psychologie, 130*, 37-52.

Kreuzpointner, L., Simon, P. & Theis, F. J. (2010). The a_d-coefficient as a descriptive measure of the with-group agreement of ratings. *British Journal of Mathematical and Statistical Psychology, 63* (2), 341-360. doi: 10.1348/000711009X465647

Lord, F. M. & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading MA: Addison-Welsley Publishing Company.

Lundy, A. C. (1985). The reliability of the Thematic Apperception Test. *Journal of Personality Assessment, 49*, 141-145. doi: 10.1207/s15327752jpa4902_6

McClelland, D. C. (1980). Motive dispositions. The merits of operant and respondent measures. In L. Wheeler (Ed.). *Review of personality and social psychology* (Vol. 1, pp 10-41). Beverly Hills, CA: Sage.

McClelland, D. C. (1985). *Human motivation*. Glenview, IL: Scott, Foresman.

McClelland, D. C., Atkinson, J. W., Clark, R. A. & Lowell, E. L. (1958). A scoring manual for the achievement motive. In J. W. Atkinson (Ed.). *Motives in fantasy, action, and society*. Princeton, NJ: van Nostrand.

McClelland , D.C., Koestner , R., Weinberger , J. (1989). How do self-attributed and implicit motives differ? *Psychological Review, 96,* 690-702.

McDonald, R. P. (1999). *Test theory: A unified treatment.* Mahwah, NJ: Lawrence Erlbaum.

McGrath, R. E. & Carroll, E. J. (2012). The current status of “projective” “tests”. In H. Cooper, P. M. Camic, D. L. Long, A. T. Panter, D. Rindskopf & K. J. Sher (Eds.). *APA handbook of research methods in psychology. Foundation planning, measures and psychometric* (pp. 329-348). Washington, DC: APA.

Mischel, W. & Shoda, Y. (1995). A cognitive-affective system theory of personality: Reconceptualizing situations, dispositions, dynamics, and invariance in personality structure. *Psychological Review, 102*, 246-268. doi: 10.1037/0033-295X.102.2.246

Pang, J. S. & Schultheiss, O. C. (2005). Assessing implicit motives in U.S. college students: Effects of picture type and position, gender and ethnicity, and cross-cultural comparisons. *Journal of Personality Assessment, 85* (3), 280-294. doi: 10.1207/s15327752jpa8503_04

Rae, G. (2007). A note on using stratified α to estimate the composite reliability of a test composed of interrelated nonhomogeneous items. *Psychological Methods, 12*, 177-184. doi: 10.1037/1082-989X.12.2.177

Reumann, D. (1982). Ipsative behavioural variability and the quality of thematic apperceptive measurement of the achievement motive. *Journal of Personality and Social Psychology, 43*(5), 1098-1110. doi: 10.1037/0022-3514.43.5.1098

Revelle, W. & Zinbarg, R. E. (2009) Coefficients alpha, beta, omega and the glb: comments on Sijtsma. *Psychometrika, 74* (1), 145-154. doi: 10.1007/s11336-008-9102-z

Rosenzweig, S. (1945). The picture-association method and its application in a study of reactions to frustration. *Journal of Personality, 14*, 3-23. doi: 10.1111/j.1467-6494.1945.tb01036.x

Schultheiss, O. C. (2001). *Manual for the assessment of hope of success and fear of failure. English translation of Heckhausen's need achievement measures.* Unpublished scoring manual. University of Michigan, Ann Arbor.

Schultheiss, O. C., Liening, S. & Schad, D. (2008). The reliability of a picture story exercise measure of implicit motives: Estimates of internal consistency, retest reliability, and ipsative stability. *Journal of Research in Personality, 42*, 1560-1571. doi: 10.1016/j.jrp.2008.07.008

Schultheiss, O. C. & Pang, J. S. (2007). Measuring implicit motives. In: R.W. Robins, R.C. Fraley & R.F. Krueger. *Handbook of research methods in personality psychology.* (pp. 322-345). New York: Guilford Press.

Tuerlinckx, F., De Boeck, P. & Lens, W. (2002). Measuring needs with the Thematic Apperception Test: A psychometric study. *Journal of Personality and Social Psychology*, *82*, 448-461. doi: 10.1037/0022-3514.82.3.448

Winter, D. G. (1994). *Manual for scoring motive imagery in running text (4ed.).* Unpublished manuscript. University of Michigan, Ann Arbor.

Appendix:

The variance of a scale is defined as the squared sum of each score subtracted from the mean of the score divided by *n*:

$$Var_j = \frac{1}{n}\sum_{j=1}^{N}(x_j - \bar{x}_j)^2$$

If a test consist of two pictures (p_1, p_2) and two categories (c_1, c_2), the matrix of all possible covariances can depict as in figure 1.

%% figure 1 about here %%

As obvious in figure 1 the total sum of all covariances is expressed as $Cov_c + Cov_p + Cov_{pc} + Var_{pc}$.

The sum of all subvariances (the variances of all subitems x_{pc}) is similar to the diagonal of the variance matrix (Cronbach, 1951, p. 303) and can be expressed as followed:

$$Var_{pc} = \frac{1}{n}\left[\sum\left(x_{j_{p1c1}} - \bar{x}_{j_{p1c1}}\right)^2 + \sum\left(x_{j_{p1c2}} - \bar{x}_{j_{p1c2}}\right)^2 + \sum\left(x_{j_{p2c1}} - \bar{x}_{j_{p2c1}}\right)^2 + \sum\left(x_{j_{p2c2}} - \bar{x}_{j_{p2c2}}\right)^2\right]$$

for j indicates counting up from first to last subject of the test.

The sum of covariances of categories will be the covariance of category one and two, which can be written as:

$$2 \cdot Cov_c = \frac{2}{n}\left[\sum\left(x_{j_{p1c1}} - \bar{x}_{j_{p1c1}}\right)\left(x_{j_{p1c2}} - \bar{x}_{j_{p1c2}}\right) + \sum\left(x_{j_{p2c1}} - \bar{x}_{j_{p2c1}}\right)\left(x_{j_{p2c2}} - \bar{x}_{j_{p2c2}}\right)\right]$$

The sum of covariances of pictures will be the covariance of picture one and two which can be written as:

$$2 \cdot Cov_p = \frac{2}{n}\left[\sum\left(x_{j_{p1c1}} - \bar{x}_{j_{p1c1}}\right)\left(x_{j_{p2c1}} - \bar{x}_{j_{p2c1}}\right) + \sum\left(x_{j_{p1c2}} - \bar{x}_{j_{p1c2}}\right)\left(x_{j_{p2c2}} - \bar{x}_{j_{p2c2}}\right)\right]$$

The next step demonstrates mathematically that this Formula truly represents the covariances, and that the sum $Var_{pc} + 2Cov_c + 2Cov_p + 2Cov_{pc}$ is the total test variance.

Figure 2 gives a detailed view of the total covariance-variance of an exemplary TAT- Picture-Category-Matrix.

%% figure 2 about here %%

As obvious above the total test variance expresses the mean squared deviation of the mean. For *n* is constant just the sum of squares (SS_t) are taken into account:

$$SS_t = \sum_{j=1}^{N}(x_j - \bar{x}_j)^2; \quad SS_t = \sum_{j=1}^{N}(x_j{}^2 - 2(x_j\bar{x}_j) + \bar{x}_j{}^2) \; ; \quad x_{jpc} = \sum_{p=1}^{P}\sum_{c=1}^{C}(x_{jpc})$$

$$SS_t = \sum_{j=1}^{N}\left(\sum_{p=1}^{P}\sum_{c=1}^{C}x_{jpc}\right)^2 - 2\left[\left(\sum_{p=1}^{P}\sum_{c=1}^{C}x_{jpc}\right)\left(\sum_{p=1}^{P}\sum_{c=1}^{C}\bar{x}_{jpc}\right)\right] + \left(\sum_{p=1}^{P}\sum_{c=1}^{C}\bar{x}_{jpc}\right)^2$$

Solving this equation with the theorem for squared sums

$$\left(\sum_{c=1}^{C}x_c\right)^2 = \sum_{c=1}^{C}(x_c)^2 + 2\sum_{c=1}^{C-1}\sum_{b=c+1}^{C}(x_c x_b)$$

leads to:

$$\begin{aligned}
SS_t = &\sum_{j=1}^{N}\left[\left(\sum_{p=1}^{P}\sum_{c=1}^{C}(x_{jpc})^2\right) + 2\left(\sum_{p=1}^{P-1}\sum_{a=p+1}^{P}\sum_{c=1}^{C}x_{jpc}x_{jac}\right) + 2\left(\sum_{p=1}^{P}\sum_{c=1}^{C-1}\sum_{b=c+1}^{C}x_{jpc}x_{jpb}\right) + 2\left(\sum_{p=1}^{P-1}\sum_{a=p+1}^{P}\sum_{c=1}^{C-1}\sum_{b=c+1}^{C}x_{jpc}x_{jab}\right)\right] \\
&-\sum_{j=1}^{N}\left[2\left(\sum_{p=1}^{P}\sum_{c=1}^{C}(x_{jpc}\bar{x}_{jpc})\right) + 2\left(\sum_{p=1}^{P-1}\sum_{a=p+1}^{P}\sum_{c=1}^{C}x_{jpc}\bar{x}_{jac}\right) + 2\left(\sum_{p=1}^{P}\sum_{c=1}^{C-1}\sum_{b=c+1}^{C}x_{jpc}\bar{x}_{jpb}\right) + 2\left(\sum_{p=1}^{P-1}\sum_{a=p+1}^{P}\sum_{c=1}^{C-1}\sum_{b=c+1}^{C}x_{jpc}\bar{x}_{jab}\right)\right] \\
&-\sum_{j=1}^{N}\left[2\left(\sum_{p=1}^{P-1}\sum_{a=p+1}^{P}\sum_{c=1}^{C}x_{jac}\bar{x}_{jpc}\right) + 2\left(\sum_{p=1}^{P-1}\sum_{c=1}^{C-1}\sum_{b=c+1}^{C}x_{jpb}\bar{x}_{jpc}\right) + 2\left(\sum_{p=1}^{P-1}\sum_{a=p+1}^{P}\sum_{c=1}^{C-1}\sum_{b=c+1}^{C}x_{jab}\bar{x}_{jpc}\right)\right] \\
&+\sum_{j=1}^{N}\left[\left(\sum_{p=1}^{P}\sum_{c=1}^{C}(\bar{x}_{jpc})^2\right) + 2\left(\sum_{p=1}^{P-1}\sum_{a=p+1}^{P}\sum_{c=1}^{C}\bar{x}_{jpc}\bar{x}_{jac}\right) + 2\left(\sum_{p=1}^{P}\sum_{c=1}^{C-1}\sum_{b=c+1}^{C}\bar{x}_{jpc}\bar{x}_{jpb}\right) + 2\left(\sum_{p=1}^{P-1}\sum_{a=p+1}^{P}\sum_{c=1}^{C-1}\sum_{b=c+1}^{C}\bar{x}_{jpc}\bar{x}_{jab}\right)\right]
\end{aligned}$$

$$\begin{aligned}
SS_t = &\sum_{j=1}^{N}\sum_{p=1}^{P}\sum_{c=1}^{C}x_{jpc}^2 - 2\sum_{j=1}^{N}\sum_{p=1}^{P}\sum_{c=1}^{C}(x_{jpc}\bar{x}_{jpc}) + \sum_{j=1}^{N}\sum_{p=1}^{P}\sum_{c=1}^{C}\bar{x}_{pc}^2 \\
&+2\left[\sum_{j=1}^{N}\sum_{p=1}^{P-1}\sum_{a=p+1}^{P}\sum_{c=1}^{C}(x_{jpc}x_{jac}) - \sum_{j=1}^{N}\sum_{p=1}^{P-1}\sum_{a=p+1}^{P}\sum_{c=1}^{C}(x_{jpc}\bar{x}_{jac}) - \sum_{j=1}^{N}\sum_{p=1}^{P-1}\sum_{a=p+1}^{P}\sum_{c=1}^{C}(\bar{x}_{jpc}x_{jac}) + \sum_{j=1}^{N}\sum_{p=1}^{P-1}\sum_{a=p+1}^{P}\sum_{c=1}^{C}(\bar{x}_{jpc}\bar{x}_{jac})\right] \\
&+2\left[\sum_{j=1}^{N}\sum_{p=1}^{P}\sum_{c=1}^{C-1}\sum_{b=c+1}^{C}(x_{jpc}x_{jpb}) - \sum_{j=1}^{N}\sum_{p=1}^{P}\sum_{c=1}^{C-1}\sum_{b=c+1}^{C}(x_{jpc}\bar{x}_{jpb}) - \sum_{j=1}^{N}\sum_{p=1}^{P}\sum_{c=1}^{C-1}\sum_{b=c+1}^{C}(\bar{x}_{jpc}x_{jpb}) + \sum_{j=1}^{N}\sum_{p=1}^{P}\sum_{c=1}^{C-1}\sum_{b=c+1}^{C}(\bar{x}_{jpc}\bar{x}_{jpb})\right] \\
&+2\left[\sum_{j=1}^{N}\sum_{p=1}^{P-1}\sum_{a=p+1}^{P}\sum_{c=1}^{C-1}\sum_{b=c+1}^{C}(x_{jpc}x_{jac}) - \sum_{j=1}^{N}\sum_{p=1}^{P-1}\sum_{a=p+1}^{P}\sum_{c=1}^{C-1}\sum_{b=c+1}^{C}(x_{jpc}\bar{x}_{jac}) - \sum_{j=1}^{N}\sum_{p=1}^{P-1}\sum_{a=p+1}^{P}\sum_{c=1}^{C-1}\sum_{b=c+1}^{C}(\bar{x}_{jpc}x_{jac}) + \sum_{j=1}^{N}\sum_{p=1}^{P-1}\sum_{a=p+1}^{P}\sum_{c=1}^{C-1}\sum_{b=c+1}^{C}(\bar{x}_{jpc}\bar{x}_{jac})\right]
\end{aligned}$$

$$SS_t = \sum_{j=1}^{N}\sum_{p=1}^{P}\sum_{c=1}^{C}(x_{jpc}-\bar{x}_{jpc})^2 + 2\sum_{j=1}^{N}\sum_{p=1}^{P-1}\sum_{a=p+1}^{P}\sum_{c=1}^{C}\left[(x_{jpc}-\bar{x}_{jpc})(x_{jac}-\bar{x}_{jac})\right]$$
$$+2\sum_{j=1}^{N}\sum_{p=1}^{P}\sum_{c=1}^{C-1}\sum_{b=c+1}^{C}\left[(x_{jpc}-\bar{x}_{jpc})(x_{jpb}-\bar{x}_{jpb})\right] + 2\sum_{j=1}^{N}\sum_{p=1}^{P-1}\sum_{a=p+1}^{P}\sum_{c=1}^{C-1}\sum_{b=c+1}^{C}\left[(x_{jpc}-\bar{x}_{pc})(x_{jab}-\bar{x}_{jab})\right]$$

These components are exactly the sum of square of variances (v), pictural covariance (p), categorical covariances (c) and general covariances (g), as so coloured in figure 2. But this equation also shows that the overall variance can be calculated as a sum of pictures variances and twice their covariances or category variances and twice their covariances. Let SS_p be the sum of squares for picture variance and SS_c the sum of square for category variance, and SS'_p the covariance multiplied with n for picture (SS'_p) and category (SS'_c):

$$SS_p = \sum_{j=1}^{N}(x_{jp}-\bar{x}_{jp})^2;\ \ for\ x_{jpc} = \sum_{p=1}^{P}\sum_{c=1}^{C}(x_{jpc});\ \ SS_p = \sum_{j=1}^{N}\sum_{p=1}^{P}\left[\sum_{c=1}^{C}(x_{jpc}-\bar{x}_{jpc})\right]^2$$

$$SS_p = \sum_{j=1}^{N}\sum_{p=1}^{P}\sum_{c=1}^{C}(x_{jpc}-\bar{x}_{jpc})^2 + 2\sum_{j=1}^{N}\sum_{p=1}^{P}\sum_{c=1}^{C-1}\sum_{b=c+1}^{C}\left[\left(x_{jpc}-\bar{x}_{jpc}\right)\left(x_{jpb}-\bar{x}_{jpb}\right)\right]$$

$$SS_c = \sum_{j=1}^{N}(x_{jc}-\bar{x}_{jc})^2;\ \ for\ x_{jpc} = \sum_{p=1}^{P}\sum_{c=1}^{C}(x_{jpc});\ \ SS_c = \sum_{j=1}^{N}\sum_{c=1}^{C}\left[\sum_{p=1}^{P}(x_{jpc}-\bar{x}_{jpc})\right]^2 =$$

$$SS_c = \sum_{j=1}^{N}\sum_{p=1}^{P}\sum_{c=1}^{C}(x_{jpc}-\bar{x}_{jpc})^2 + 2\sum_{j=1}^{N}\sum_{p=1}^{P-1}\sum_{a=p+1}^{P}\sum_{b=c+1}^{C}\left[\left(x_{jpc}-\bar{x}_{jpc}\right)\left(x_{jac}-\bar{x}_{jac}\right)\right]$$

$$SS'_p = \sum_{j=1}^{N}\sum_{p=1}^{P-1}\left(x_{jp}-\bar{x}_{jp}\right)\left(y_{jp}-\bar{y}_{jp}\right)\ \text{ for } y \neq x \text{ leads to } SS'_p = \sum_{j=1}^{N}\sum_{p=1}^{P-1}\sum_{a=p+1}^{P}\left(x_{jp}-\bar{x}_{jp}\right)\left(x_{ja}-\bar{x}_{ja}\right)\ \text{ for } x_j = \sum_{c=1}^{C}x_{jc}$$

$$SS'_p = \sum_{j=1}^{N}\sum_{p=1}^{P-1}\sum_{a=p+1}^{P}\left[\sum_{c=1}^{C}\left(x_{jpc}-\bar{x}_{jpc}\right)\sum_{c=1}^{C}\left(x_{jac}-\bar{x}_{jac}\right)\right]$$

$$SS'_p = \sum_{j=1}^{N}\sum_{p=1}^{P-1}\sum_{a=p+1}^{P}\left[\sum_{c=1}^{C}\left(x_{jpc}-\bar{x}_{jpc}\right)\left(x_{jac}-\bar{x}_{jac}\right)+\sum_{c=1}^{C-1}\sum_{b=c+1}^{C}\left(x_{jpc}-\bar{x}_{jpc}\right)\left(x_{jab}-\bar{x}_{jab}\right)\right]$$

$$SS'_p = \sum_{j=1}^{N}\sum_{p=1}^{P-1}\sum_{a=p+1}^{P}\sum_{c=1}^{C-1}\sum_{b=c+1}^{C}\left[\left(x_{jpc}-\bar{x}_{jpc}\right)\left(x_{jac}-\bar{x}_{jac}\right)\right] + \sum_{j=1}^{N}\sum_{p=1}^{P-1}\sum_{a=p+1}^{P}\sum_{c=1}^{C-1}\sum_{b=c+1}^{C}\left[\left(x_{jpc}-\bar{x}_{jpc}\right)\left(x_{jab}-\bar{x}_{jab}\right)\right]$$

$$SS'_c = \sum_{j=1}^{N}\sum_{c=1}^{C-1}\sum_{b=c+1}^{C}\left[\sum_{p=1}^{P}\left(x_{jpc}-\bar{x}_{jpc}\right)\sum_{p=1}^{P}\left(x_{jpb}-\bar{x}_{jpb}\right)\right]$$

$$SS'_c = \sum_{j=1}^{N}\sum_{c=1}^{C-1}\sum_{b=c+1}^{C}\left[\sum_{p=1}^{P}\left(x_{jpc}-\bar{x}_{jpc}\right)\left(x_{jpb}-\bar{x}_{jpb}\right)+\sum_{p=1}^{P-1}\sum_{a=p+1}^{P}\left(x_{jpc}-\bar{x}_{jpc}\right)\left(x_{jab}-\bar{x}_{jab}\right)\right]$$

$$SS'_c = \sum_{j=1}^{N}\sum_{p=1}^{P-1}\sum_{a=p+1}^{P}\sum_{c=1}^{C-1}\sum_{b=c+1}^{C}\left[\left(x_{jpc}-\bar{x}_{jpc}\right)\left(x_{jpb}-\bar{x}_{jpb}\right)\right] + \sum_{j=1}^{N}\sum_{p=1}^{P-1}\sum_{a=p+1}^{P}\sum_{c=1}^{C-1}\sum_{b=c+1}^{C}\left[\left(x_{jpc}-\bar{x}_{jpc}\right)\left(x_{jab}-\bar{x}_{jab}\right)\right]$$

Now taking these similarities and differences of SS'_p and SS'_c into account for the calculation of α using the category-scores and the picture-scores (see also Eq. 3a and 3b, page 11):

Into

$$\alpha = \frac{n}{(n-1)} \cdot \frac{2Ct}{Vt} \Leftrightarrow n \cdot Ct = 0.5\ (n-1)\ Vt\ \alpha$$

for categories the decomposition of C_c into the blue term (c) and the yellow term (g, see above) and for pictures into the green term (p) and the yellow term (g) as well will inserted:

$n \cdot C_c = c + g$ and $n \cdot C_p = p + g$

$c + g = 0.5\ (n-1)\ Vt\ \alpha_c$ and $p + g = 0.5\ (n-1)\ Vt\ \alpha_p$

Resolving both sides to g and equate to each other leads to the equation of α_c as function of α_p:

$$0.5\ (n-1)\ Vt\ \alpha_c - c = 0.5\ (n-1)\ Vt\ \alpha_p - p$$

$$\Leftrightarrow \alpha_c = \frac{0.5\ (n-1)\ Vt\ \alpha_p - p + c}{0.5\ (n-1)\ Vt} = \alpha_p + \frac{2\ (c-p)}{(n-1)\ Vt} \quad (6)$$

Figure 1

Variance-covariance-matrix for two pictures and two categories

Figure 2

Variance-covariance-matrix for total TAT-ratings with pictures from A to F and categories from 1 to 6.

Table 1

Example data matrix for seven subjects with sums of categories (cat1, cat2, cat3) and sums of pictures (A, B, C)

	Picture A			Picture B			Picture C									
subject	Cat 1	Cat 2	Cat 3	Cat 1	Cat 2	Cat 3	Cat 1	Cat 2	Cat 3	Sum	Cat 1	Cat 2	Cat 3	A	B	C
1	1	1	1	1	1	1	1	1	1	9	3	3	3	3	3	3
2	0	1	1	1	0	1	1	1	0	6	2	2	2	2	2	2
3	0	0	1	1	0	0	0	1	0	3	1	1	1	1	1	1
4	0	0	0	1	1	1	1	1	1	6	2	2	2	0	3	3
5	1	1	1	0	0	0	1	1	1	6	2	2	2	3	0	3
6	1	1	1	1	1	1	0	0	0	6	2	2	2	3	3	0
7	1	1	0	1	0	1	1	1	1	7	3	2	2	2	2	3

Table 2

Intercorrelations of categories (cat1, cat2, cat3) and pictures (A, B, C)

	Cat 1	Cat 2	Cat 3		A	B	C
Cat 1	1.00			A	1.00		
Cat 2	.84	1.00		B	-.13	1.00	
Cat 3	.84	1.00	1.00	C	-.12	-.12	1.00

Table 3

Reliability-coefficients (Guttman, 1945; McDonald, 1999) for categories and pictures regarding the two scales hope for success and fear of failure with weighting categories (above) and without (below)

	Hope of Success		Fear of Failure	
λ	Category	Picture	Category	Picture
1	.40/.44/.59	.18/.10/.40	.51/.27/.58	-.02/-.10/.30
2	.59/.59/.69	.36/.22/.49	.65/.38/.71	.17/.18/.39
3 = α	.48/.52/.67	.22/.12/.47	.60/.31/.68	-.02/.10/.36
4	.62/.52/.76	.28/.16/.46	.55/.37/.67	-.55/-.55/.33
5	.61/.61/.68	.36/.22/.50	.65/.51/.71	.20/.20/.40
6	.69/.57/.69	.35/.17/.45	.66/.37/.74	.16/.16/.35
ω_t	.67/.54/.84	.54/.41/.64	.69/.42/.79	.47/.28/.42
Items	6	6	7	6

	Hope of Success		Fear of Failure	
λ	Category	Picture	Category	Picture
1	.09/.46/.30	.06/.33/.10	.36/.51/.07	-.05/.16/.07
2	.26/.60/.44	.24/.46/.21	.50/.64/.19	.16/.34/.17
3 = α	.11/.57/.37	.07/.40/.12	.43/.61/.08	-.05/.19/.08
4	-.09/.59/.23	.21/.50/.11	.21/.47/.17	-.65/-.38/-.01
5	.27/.61/.47	.25/.45/.21	.50/.66/.21	.19/.35/.18
6	.22/.55/.39	.21/.44/.16	.48/.61/.15	.14/.32/.14
ω_t	.47/.32 /.63	.50/.37/.42	.59/.39 /.61	.49/.36/.63
Items	5	6	6	6

Note. The first of the three coefficients listed for each λ is from the Heckhausen data set (N = 35); the second coefficient is from the study with students (N = 113); and the third coefficient is from the pupil sample (N = 241).

Table 4

Intercorrelations of Pictures and Categories for the Heckhausen data set (N = 35)

Intercorrelations of Picture-scores									
	A	B	C	D	E	F	*M*	*SD*	r_{it}
A		-.25	.27	.09	-.26	.27	2.74	1.27	.31
B	-.10		-.01	-.06	-.04	-.30	0.11	0.32	.31
C	.04	.03		.42*	.00	.28	2.00	1.55	.20
D	-.08	-.13	.11		-.05	.10	0.06	0.24	.23
E	-.13	.01	.05	.28		-.03	1.74	1.27	.36
F	.07	-.27	-.35*	.22	.35*		0.29	0.62	-.23
M	0.14	2.20	0.26	2.00	0.40	0.89			
SD	0.36	1.45	0.61	1.68	0.91	1.08			
r_{it}	.16	-.08	-.05	.20	.25	-.17			

Note. Correlation coefficients over the diagonal refer to HS, below refer to FF, Heckhausen data set n = 35, * p < .05

Intercorrelations of Category-scores										
	NS/NF	IS/IF	ES/EF	P/C	A+/A-	/F	ST/FT	*M*	*SD*	r_{it}
NS/NF		.08	-.11	-.04	-.36*		.64**	1.09	0.95	.67
IS/IF	-.05		.26	.00	.06		.36*	2.46	0.78	.57
ES/EF	.07	-.06		-.13	.34*		.35*	0.49	0.66	.16
P/C	.22	-.21	.28		.12		-.13	0.20	0.47	.02
A+/A-	.08	.01	.20	.44**			.20	1.26	0.92	.08
/F	.03	.02	.06	.44**	.40*		-	1.46	1.17	.55
ST/FT	.43**	.02	.11	.41*	.56**	.53**		1.09	0.95	.66
M	0.43	0.57	1.14	0.23	1.91	0.77	0.83			
SD	0.61	0.98	0.97	0.43	1.04	0.84	0.89			
r_{it}	.30	.18	.15	.12	.25	.00	.42			

Note. Correlation coefficients over the diagonal refer to HS, below refer to FF, Heckhausen data set n = 35, * p < .05, ** p < .01

Table 5

Reliability-coefficients regarding to the two scales hope for success and fear of failure with dichotomous data for categories-by-pictures

	Hope of Success	Fear of Failure
λ_3 (resp. KR-20)	.50 / .56 /.70	.52 / .42 /.68
Items	36	42

Note. The first of the three coefficients listed is from the Heckhausen data set (N = 35); the second coefficient is from the study with students (N = 113); and the third coefficient is from the pupil sample (N = 241).

Figure 1
Click here to download high resolution image

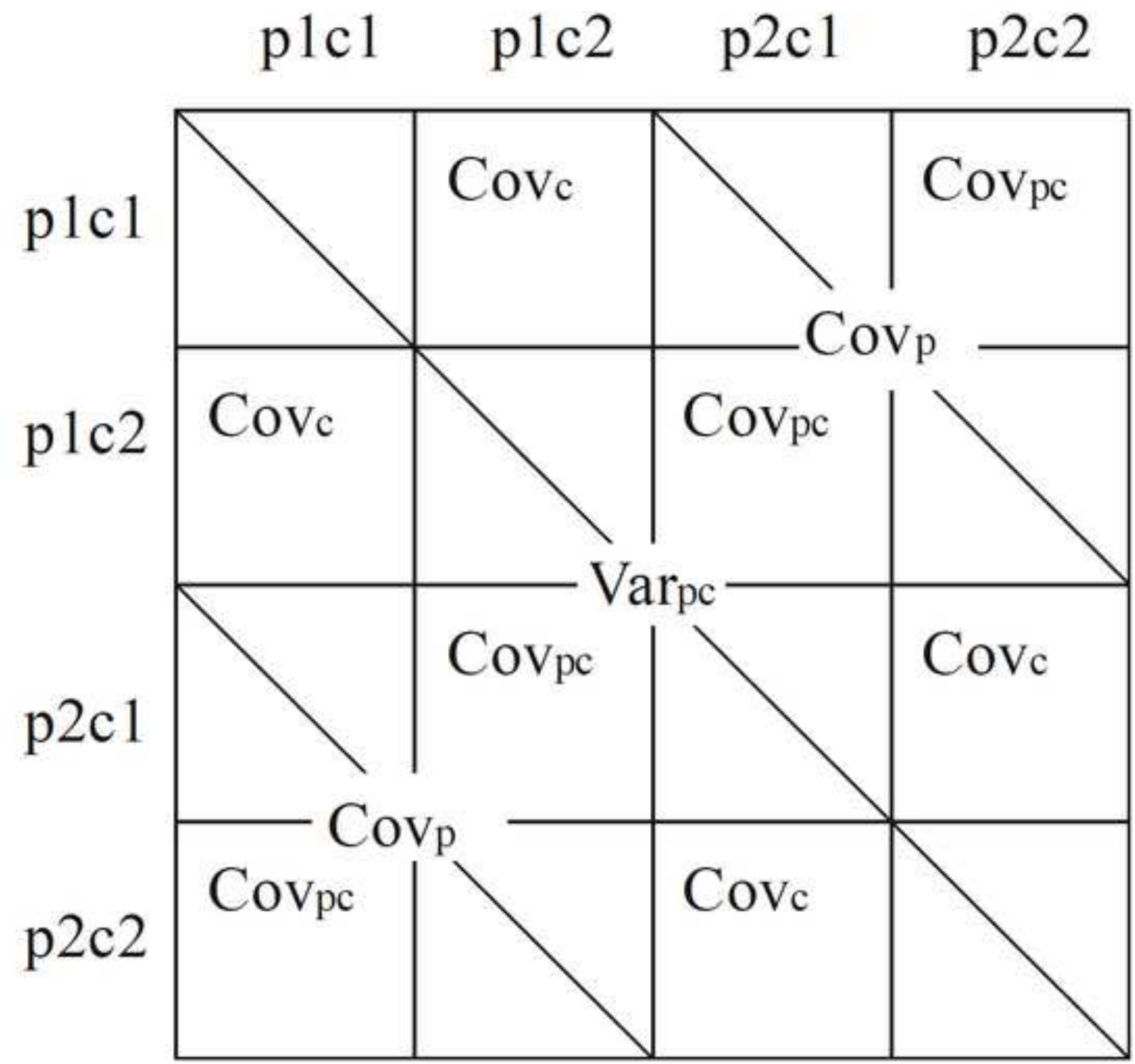
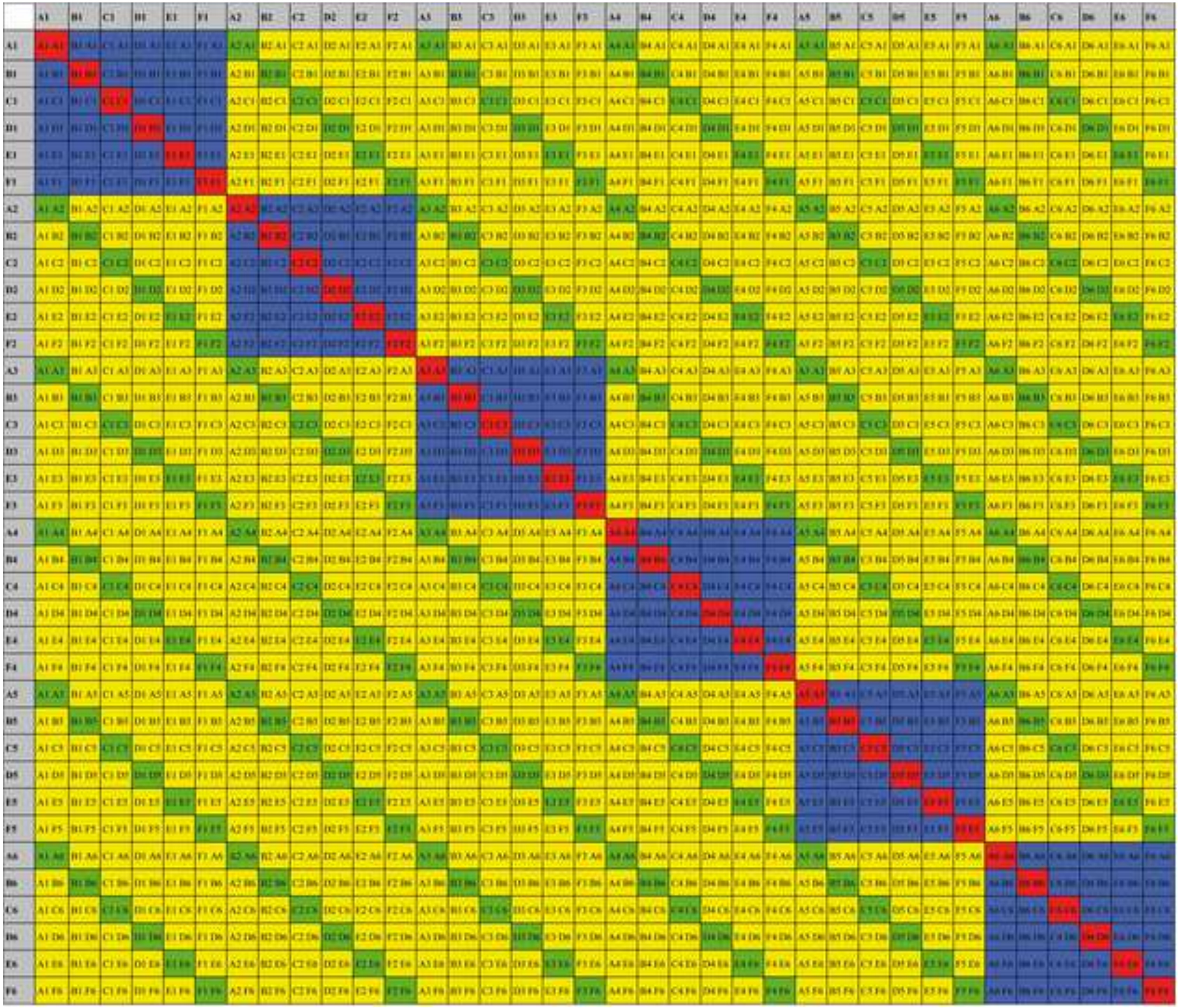


Figure 2

Click here to download high resolution image



Heckhausen data set

Click here to download Supporting Information: Heckhausen.dat

Breidebach data set

Click here to download Supporting Information: Breidebach.dat