Within-subject comparison of changes in a pretest-posttest design

Christian Hennig (University College London, Department of Statistical Science,
Daniel Müllensiefen (Goldsmiths College London, Department of Computing),
Jens Bargmann (Musikwissenschaftliches Institut, Universität Hamburg)

April 6, 2009

Abstract

A method to compare the influence of a treatment on different properties within subjects is proposed. The properties are measured by several Likert scaled items. It is shown that many existing approaches such as repeated measurement analysis of variance on sum/mean scores, a linear partial credit model and a graded response model conceptualize a comparison of changes in a way that depends on the distribution of the pretest values, while in the present paper change is measured in terms of the conditional distributions of posttest values given the pretest values. A multivariate regression/ANCOVA approach is unbiased, but shows power deficiencies in a simulation study. A new approach is suggested based on poststratification, i.e., aggregating change information conditional on each pretest value, which is unbiased and has a superior power. The approach is applied in a study that compares the influence of a certain piece of music on five different basic emotions.
Keywords: multivariate regression, repeated measurements, item response theory, graded response model, poststratified relative change scores, music and emotions

Introduction

In the present article, the analysis of data of the following form is addressed: I properties (that could be, e.g., attitudes or emotional states) of K test persons are measured by J_i, $i = 1, \ldots, I$, items (usually the J_i are the same for all properties, but this is doesn't have to be assumed) before and after a treatment. The items are scaled by P ordered categories, which should have a comparable meaning with respect to the various items. The question of interest is if one of the properties is significantly more affected by the treatment than the others. We suggest to analyze such data by a new approach based on poststratified relative change scores (PRCS). Before this approach is introduced, we discuss the application of some already existing methodology.

While there is a lot of literature on measuring change within pretest-posttest data (e.g. Cronbach and Furby, 1970, Fischer, 1976, Andersen, 1985, Embretson, 1991, Eid and Hoffmann, 1998, Dimitrov and Rumrill, 2002, Achcar et al., 2003, Fischer, 2003, further references can be found in Bonate, 2000), almost all work concerns the comparison of changes between subjects of different groups. In our setup, we want to compare changes between different variables within the same subject. The meaningfulness of such a comparison is discussed in the discussion.

Denote the random variables giving the pre- and posttest values of the items by X_{hijk}, where

- $h \in \{0, 1\}$ is 0 for a pretest score and 1 for a posttest score,

- $i \in I\!N_I = \{1, \ldots, I\}$ denotes the number of the property,
- $j \in I\!N_{J_i}$ denotes the number of an item corresponding to property i, i.e. an item is specified by the pair (i, j),
- $k \in I\!N_K$ denotes the test person number. If nothing else is said, h, i, j, and k are used as defined here.

A typical example is data from questionnaires where the measurement of different properties of the test persons is operationalized by asking J_i questions with five ordered categories for the answers with the same descriptions for all items, e.g., "strongly agree", "agree", "neither agree nor disagree", "disagree", "strongly disagree". In the data example, which is treated after the methodological sections, the aim was to find out if a piece of music from the movie soundtrack of "Alien III" affects anxiety significantly more than other emotions. The data was obtained by a questionnaire, which consisted of $J = J_i = 10$ times $i = 1, \ldots, I = 5$ questions on a $P = 5$-point scale as above corresponding to the emotions joy, sadness, love, anger and anxiety.

Such properties are frequently measured by Likert scales (Likert, 1932), i.e., the categories are treated as numbers 1, 2, 3, 4, and 5, and the mean over the values of the J_i items is taken as a score for each property (in the literature, often the sum is taken, but the mean allows unequal values of J_i): $L_{hik} = \frac{1}{J_i} \sum_{j=1}^{J_i} X_{hijk}$, **called Likert mean scores in the following**. The distribution of such mean scores is often not too far from the normal, and they allow the application of several linear models such as a repeated measures analysis of variance or an analysis of covariance (Jaccard and Wan, 1996). Instead of the analysis of covariance, we introduce a multivariate regression model, which is more general and more appropriate for the multiple properties data.

Alternatively, the data can be analyzed on the level of the single items using item response theory. This can be done by the so-called partial credit model (Masters, 1982), which is applied to the measurement of changes by Fischer and Ponocny (1994). This is the only reference known to us that can be directly applied to the data analysis problem treated in the present article. We discuss this approach along with a possible application of the graded response model (Samejima, 1969) to data of this kind. Pretest-posttest data have also been analyzed by means of structural equation models (Raykov, 1992, Steyer, Eid and Schwenkmezger, 1997, Cribbie and Jamieson, 2004). It would be possible in principle to adapt this approach to the present situation, but on the mean score level, such a method will be very similar to ANCOVA, and on the single item level, the normality assumption will be strongly violated.

Our conception of a comparison of change is based on a comparison of the conditional distributions of the posttest values given the pretest values. Our definition for "equal changes in different properties" is that for all possible pretest values these conditional distributions are equal between the properties.

Existing approaches such as the repeated measures ANOVA and the item response theory model "equality of change" in different ways, usually via parameters corresponding to differences between the pretest and posttest distribution that are interpreted as time-property interactions. We show by a toy-example that change for these approaches depends on the distribution of the pretest values and that the corresponding parameters may indicate interactions even if all conditional distributions are equal. The reason is that the lower the pretest value, the more increase is possible. For example, from a pretest value of $x = P = 5$, no further positive change can happen. The effect is similar to the so-called regression towards the mean in the pretest-posttest literature (cf. Bonate, 2000, Chapter 2). It is more serious for within-subject comparisons, because for comparisons between groups, the same theoretical distribution of the pretest values can be arranged by randomization, while this is not possible for comparisons

between different variables. However, the example is also relevant for between-groups comparison situations where the pretest distribution varies between the groups. Jamieson (1995) and Cribbie and Jamieson (2004) addressed similar effects by means of simulations.

After these sections, a new method is proposed that is more directly tailored to the specific kind of data. The proposed poststratified relative change scores (PRCS) method is based on a separate poststratification of the items of every single test person. The PRCS aggregates the differences between the posttest scores of the items corresponding to the property of interest and the mean posttest score for all other items with the same pretest score. It makes explicit use of the fact that a property is measured by aggregating the results from m items with p ordered categories (p not too large) instead of analyzing the Likert mean scores.

Poststratification based scoring has been introduced by Bajorski and Petkau (1999). These authors compute weighted sums of the P Wilcoxon rank test statistics for the posttest scores conditional on the P pretest values. As opposed to the present setup, Bajorski and Petkau (1999) deal with the comparison of two independent groups of test persons.

The Alien dataset is introduced and analyzed by the multivariate regression and the PRCS after the methodological sections.

As ANCOVA in standard setups, the multivariate regression on the pretest values in our setup is also a reasonable strategy to deal with regression towards the mean. However, the fact that it ignores the way how the Likert mean values are obtained, may result in serious power losses in some situations. This is illustrated in a small simulation study, which reveals a superiority of the PRCS approach.

Note that we do not restrict our attention to a particular model for change or treatment effects. We start with a data analytic question and compare tests derived from very different models which can be applied to give an answer. The linear regression and ANOVA approaches are based on models for $(L_{hik})_{hik}$ (denoting the vector of all Likert mean scores for all values of h, i, k), where differences between changes of different properties i_1, i_2 are modeled by parameters that specify different expected values of L_{1i_1k} and L_{1i_2k} conditional on $L_{0i_1k} = L_{0i_2k}$. In item response theory, a difference between the changes of different properties i_1, i_2 is modeled by an effect parameter for a difference between the distributions of $(X_{hi_1jk})_j$ and $(X_{hi_2jk})_j$ that does only occur for $h = 1$ but not for $h = 0$. For the PRCS approach, such a difference is understood as a difference between the expectations of the values of X_{1i_1jk} and X_{1i_2jk} conditional under $X_{0i_1jk} = X_{0i_2jk}$ in a nonparametric setup.

Most item response theory methods operate on logits or probits of probabilities, while the regression, ANOVA and PRCS methods operate on the raw Likert scores. The question of scaling is discussed, along with some other issues, in the concluding discussion.

Linear regression and ANOVA approaches

Repeated measures ANOVA

A straight forward approach to analyze the Likert mean score data is a repeated measures analysis of variance model:

$$L_{hik} = \mu + a_h + b_i + c_k + d_{hi} + e_{hik}, \tag{1}$$

where μ is the overall mean, a_h is the effect of time (pre- or posttest), b_i is the effect of the property, c_k **is the random effect of the test person**, d_{hi} is the interaction of time and property and e_{hik} is the error term, usually modeled as independently and identically distributed (i.i.d.) according to a normal distribution. If it is of interest to contrast one particular property i_0 with the others, i may take the "values" i_0 and "$-i_0$", where $L_{h-i_0k} =$

$\frac{1}{\sum_{q\neq i_0} J_q} \sum_{q\neq i_0} \sum_{r=1}^{J_q} X_{hqrk}$ is the aggregated Likert mean score of all items not belonging to property i_0 (a subscript with a minus generally denotes aggregation of all possible values at this place except of the one with the minus). The effects are assumed to be appropriately constrained for identifiability. A difference of changes between properties would be tested by testing the equality of the time-property interactions d_{hi} (equality to 0 under the usual constraints), by analogy to the case where difference of changes between groups is of interest, compare Chapter 7 of Bonate (2000).

Example 1 *We present an extreme, but simple example to demonstrate that the model Eq. 1 can indicate a time-property interaction even if for all pretest values the conditional distributions of the posttest values equal between the properties. This is caused here by different pretest value distributions for the properties,*

Assume that there are only two properties with one item for each, and these items can only take the values 0 and 1. For both items a pretest value of 0 leads to a posttest value of 0 with probability 0.1 and to a posttest value of 1 with probability 0.9. A pretest value of 1 leads always to a posttest value of 1, independently for both items. Therefore, the distributions of the changes for both items are exactly the same. The distribution of the pretest values for a test person k is assumed to be: $P\{L_{01k} = 0\} = 0.9$, $P\{L_{01k} = 1\} = 0.1$, $P\{L_{02k} = 0\} = 0.1$, $P\{L_{02k} = 1\} = 0.9$. This yields the following distribution of the posttest values: $P\{L_{11k} = 0\} = 0.09$, $P\{L_{11k} = 1\} = 0.91$, $P\{L_{12k} = 0\} = 0.01$, $P\{L_{12k} = 1\} = 0.99$. Because of the independence of the items, they could also be interpreted as items belonging to different groups. Thus, the example is also relevant for between-groups comparisons under different pretest distributions.

Some tolerance is required to apply the model Eq. 1 to this situation, because the dependent variable is only two-valued and so the error term cannot be normally distributed. Note, however, that the fact that only zeroes and ones occur as values is by no means essential for this example. The same problem as demonstrated below occurs with mixtures of normally distributed random variables or bimodally distributed Likert scores arranged so that the expected values are the same as below. The reason why we used a two-valued example is only that this makes the calculations easier (two-valued responses may be associated with techniques like logistic regression).

For the sake of simplicity, we use the constraints $a_0 = 0,\ b_2 = -b_1,\ d_{01} = d_{02} = d_{12} = 0$. We obtain from the expected values E:

$$\begin{aligned} EL_{01k} = 0.1 &\Rightarrow 0.1 = \mu + b_1, \\ EL_{02k} = 0.9 &\Rightarrow 0.9 = \mu - b_1, \\ EL_{11k} = 0.91 &\Rightarrow 0.91 = \mu + a_1 + b_1 + d_{11}, \\ EL_{12k} = 0.99 &\Rightarrow 0.99 = \mu + a_1 - b_1. \end{aligned}$$

Solving for the parameters:

$$\mu = 0.5,\ b_1 = -0.4,\ a_1 = 0.09,\ d_{11} = 0.72.$$

The interaction term d_{11} has the largest absolute value, even though the effect of time is equal for both items conditional on both possible values. The parameter models the fact that $EL_{11k} - EL_{01k}$ is much larger than $EL_{12k} - EL_{02k}$, which does not reflect a difference between the changes, but a pretest distribution of item 1 that leaves much more space for a positive change.

ANCOVA

In usual pretest-posttest setups, the related phenomenon of regression towards the mean can be handled by ANCOVA, i.e., introducing the pretest value as a covariate. The analogous

model for the present setup would be

$$L_{1ik} = \mu + b_i + c_k + \beta(L_{0ik} - \bar{L}_0) + e_{hik}, \tag{2}$$

where b_i is the effect of the property, c_k is a random within-subject effect, β is the regression coefficient for the pretest value, $\bar{L}_0 = \sum_{i=1}^{l}\sum_{k=1}^{n} L_{0ik}/(nl)$ is the overall pretest score mean, and e_{hik} is the error term. Here, absence of differences of changes is modeled by equal property effects b_i. Some suitable constraints have to be added to guarantee identifiability. The pretest scores are centered by $\bar{L}_0$ independent of i, because this makes the contribution of $\beta(L_{0ik} - \bar{L}_0)$ independent of i given the pretest score, and differences between changes manifest themselves completely in the b_i. For a more general model the regression coefficient β could be chosen dependent of i, which restricts the clear interpretation of b_i to the case $L_{0ik} = \bar{L}_0$. These models assume that the posttest value of property i is independent of the pretest values of the other properties and that the dependence between results of the same test person takes the form of an additive constant, i.e., the within-subject correlation has to be positive, which cannot be taken for granted in the present setup.

Multivariate regression

These assumptions can be avoided by a more general multivariate regression model. For ease of notation, we assume that only property i (and the aggregated score for the other properties, denoted by "$-i$") is of interest. With that,

$$\begin{pmatrix} L_{1ik} \\ L_{1-ik} \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} + \begin{pmatrix} \beta_{11} & \beta_{12} \\ \beta_{21} & \beta_{22} \end{pmatrix} \begin{pmatrix} L_{0ik} - \bar{L}_{0i} \\ L_{0-ik} - \bar{L}_{0i} \end{pmatrix} + \begin{pmatrix} e_{1k} \\ e_{2k} \end{pmatrix}, \tag{3}$$

$\bar{L}_{0i} = (\sum_{k=1}^{K}(L_{0ik} + L_{0-ik}))/(2K)$ being the overall pretest score mean. μ_1 and μ_2 are the treatment effects on property i and on the aggregate of the other properties. The regression matrix $\begin{pmatrix} \beta_{11} & \beta_{12} \\ \beta_{21} & \beta_{22} \end{pmatrix}$ specifies the influence of the pretest scores. e_{1k} and e_{2k} are error variables with zero mean independent of L_{0ik} and L_{0-ik}, but may depend on each other, which accounts for the within-subject correlation. The null hypothesis of interest is the equality of the treatment effects for L_{1ik} and L_{1-ik}, i.e., $\mu_1 - \mu_2 = 0$. This may be tested by a standard t-test of $\mu = 0$ in the univariate linear regression model

$$L_{1ik} - L_{1-ik} = \mu + \beta_1(L_{0ik} - \bar{L}_{0i}) + \beta_2(L_{0-ik} - \bar{L}_{0i}) + e_k. \tag{4}$$

While this is the most general approach, **it can be favorable in terms of the power of the test (see the simulation study) to reduce the number of free parameters by assuming**

$$\beta_{11} = \beta_{22},\ \beta_{12} = \beta_{21} = 0,\ \text{thus } \beta_2 = -\beta_1 \text{ in Eq. 4.} \tag{5}$$

This means that the difference between L_{1ik} and L_{1-ik} apart from the random error can be explained by $\mu_1 - \mu_2$ and the difference between L_{0ik} and L_{0-ik} alone. If this is not the case, the difference depends on the size of L_{0ik} and L_{0-ik} even if they are equal. The assumption Eq. 5 will often not be justified in practice, but it makes the interpretation of $\mu_1 - \mu_2$ more obvious and the real data and simulation sections demonstrate that it can improve the power of the resulting tests. The only difference between the model Eq. 2 and the multivariate regression Eq. 3 is that the latter model allows for a more general within-subject correlation structure (this could also be introduced in the models Eqs. 1 and 2 by replacing the within-subject random effect with a more complicated covariance structure among the errors).

In the setup of Example 1, it can be shown by analogous calculations that indeed $\mu_1 = \mu_2$ under Equation 3. Thus, the multivariate regression approach is superior to the repeated measurements approach in this setup. Nevertheless, the approach shows a weak power under some non-identical distributions of L_{0ik} and L_{0-ik} in the simulation study. The reason is that it ignores the nature of the Likert mean scores, which apparently leads to a violation of the linearity of the influence of the pretest scores on the posttest scores. An improvement can be attained by the PRCS introduced later, which utilizes information at the item level. First, we discuss another item-based method, which is already well-known in the literature.

Item response theory approaches

A linear partial credit model

As a contribution to item response theory, Fischer and Ponocny (1994) proposed the linear partial credit model (LPCM), which can be applied to the data treated in the present paper. Using our notation, the LPCM looks as follows:

$$\begin{aligned} P(X_{hijk} = x|\theta_k, \delta_{xhij}) &= \frac{\exp(x\theta_k + \delta_{xhij})}{\sum_{t=1}^{P} \exp(t\theta_k + \delta_{xhij})}, \\ \delta_{xhij} &= \beta_{xij} + x\tau_{hi}, \end{aligned} \tag{6}$$

where x is the item value on the P-point scale, θ_k is **a person effect**, β_{xij} is the item parameter for item (i, j), one for every possible value x, τ_{hi} for $h = 1$ specifies the change between pretest and posttest on the items of property i (a more complicated model may involve parameters τ_{hij}). To test whether "the treatment effects generalize over items", which is Fischer and Ponocny's (p. 188/189) formulation of a within-subject comparison of change, they suggest to test the equality of the τ_{1i}. Some constraints are needed to ensure the identifiability of the parameters:

$$\tau_{0i} = 0 \ \forall i, \ \delta_{1hij} = 0 \ \forall h, i, j, \ \sum_{x,h,i,j} \delta_{xhij} = 0. \tag{7}$$

In the present setup, the LPCM, as well as the model Eq. 1, can suggest a property-time-interaction via the parameters τ_{1i} even if for all pretest values the conditional distributions of the posttest values are equal between items. This can again be illustrated by means of Example 1. To apply the model to the example with only two items, we replace the pairs (i, j) in Equation 6 by a single index $i = 1, 2$. x takes the values 0 and 1 and the first constraint to the δ_{xhij} is taken as $\delta_{0hi} = 0$. Further, θ_k is assumed constant. This yields

$$\begin{aligned} P(X_{01k} = 1|\theta_k, \delta_{101}) = 0.1 = \frac{\exp(\theta_k + \delta_{101})}{1 + \exp(\theta_k + \delta_{101})} &\Rightarrow \theta_k + \delta_{101} = -2.197, \\ P(X_{02k} = 1|\theta_k, \delta_{102}) = 0.9 &\Rightarrow \theta_k + \delta_{102} = 2.197, \\ P(X_{11k} = 1|\theta_k, \delta_{111}) = 0.91 &\Rightarrow \theta_k + \delta_{111} = 2.314, \\ P(X_{12k} = 1|\theta_k, \delta_{112}) = 0.99 &\Rightarrow \theta_k + \delta_{112} = 4.595. \end{aligned}$$

Solving for the parameters,

$$\begin{gathered} \theta_k = 1.727, \ \delta_{101} = -3.924, \ \delta_{102} = 0.47, \ \delta_{111} = 0.587, \ \delta_{112} = 2.867, \\ \delta_{101} = \beta_{11}, \ \delta_{102} = \beta_{12}, \ \delta_{111} = \beta_{11} + \tau_{11}, \ \delta_{112} = \beta_{12} + \tau_{12} \\ \Rightarrow \tau_{11} = 4.511 \neq \tau_{12} = 2.397. \end{gathered}$$

The parameters describing the change are unequal, which would lead to the conclusion that the changes are different between items 1 and 2. The reason is again that the model does not separate the influence of the pretest value distribution from the comparison of changes. The τ-parameters are obtained from the differences $P(X_{1ik}=1|\theta_k,\delta_{11i})-P(X_{0ik}=1|\theta_k,\delta_{10i})$, which do not correspond to the changes alone. Thus, the LPCM is able to parametrize the given situation, but the parameters do not have the desired interpretation.

Again, it has to be emphasized that the simplicity of the example is not essential for the problem. Examples with more possible values can generate analogous problems as well as situations with more items.

There are multivariate approaches to the measurement of change (Embretson, 1991, Wang and Chyi-In, 2004) in which the within-person parameter θ_k is a multidimensional vector containing so-called modifiabilities to measure the change. Furthermore, there could be different components of θ_k corresponding to different properties. Parameters θ_{hik}, $h=0$ indicating a baseline effect and $h=1$ indicating a modifiability, $i=1,2$ indicating two different properties with one item each, would be introduced. However, in the simple example above this is equivalent to the univariate LPCM: assuming that the comparison of within-subject changes can be modelled by an additive constant, $\theta_{12k}=\theta_{11k}+\tau$, the null hypothesis $\tau=0$ would have to be tested. $\delta_{11i}=\theta_{1ik}$ and $\delta_{10i}=0$ have to grant identifiability and eventually $\tau=\tau_{12}-\tau_{11}$ above, indicating that the multivariate approaches are affected by the same problem.

A graded response model

A further class of item response models are the graded response models (Samejima, 1969). We are not aware of any literature where these models have been applied to within-subject comparisons of change. Therefore we propose our own adaptation to Example 1. There are various versions of the graded response model around. Our approach is based on the model formulation of Eid and Hoffmann (1998). The basic model is

$$P(X_{hik}\geq x|a_i,\theta_{hik},\lambda_{xi})=\Phi[a_i(\theta_{hik}-\lambda_{xi})], \tag{8}$$

where Φ is the cumulative distribution function of the standard normal distribution (distribution of the underlying latent variables), a_i is an item difficulty parameter, λ_{xi} is a cutoff parameter determining the borders between the ordered categories and θ_{hik} is an ability parameter of the subject depending on item and occasion (pre- or posttest). A suitable model for within-subject comparison of changes is $\theta_{1ik}=\theta_{0ik}+c_i$, where $c_1=c_2$ is the null hypothesis to be tested (assuming, for our toy example, that there are only two properties with one item each, $i=1,2$). In the given form, the model is heavily overparametrized. x can only take the values 0 and 1, therefore $\{X_{hik}\geq 1\}=\{X_{hik}=1\}$ and only λ_{1i} are needed. As above, there are only four equations to determine the parameters, but there are still eight parameters. Therefore four further constraints have to be made, namely $a_1=a_2=1,\ \lambda_{11}=\lambda_{12}=\lambda,\ \theta_{01k}=0$. This yields

$$\begin{aligned}
P(X_{01k}=1|\lambda) &= \Phi[-\lambda]=0.1,\\
P(X_{02k}=1|\theta_{02k},\lambda) &= \Phi[\theta_{02k}-\lambda]=0.9,\\
P(X_{11k}=1|\lambda,c_1) &= \Phi[c_1-\lambda]=0.91,\\
P(X_{12k}=1|\theta_{02k},\lambda,c_2) &= \Phi[\theta_{02k}+c_2-\lambda]=0.99.
\end{aligned}$$

Solving for the parameters:

$$\lambda=1.28,\ \theta_{02k}=2.56,\ c_1=2.62,\ c_2=1.04.$$

Again, the within-subject changes seem to differ between items, but this is only due to the different pretest value distributions.

To summarize, the discussed item response theory approaches, as well as the linear model Eq. 1, model the time-property-interaction in a way that it depends on the distribution of the pretest values and can be present even in the case that all conditional distributions of posttest values are equal between the properties. We acknowledge, however, that it may be appropriate, depending on the application, to conceptualize a comparison of changes in terms of (logit or probit scaled) differences between pretest and posttest distributions. In such a situation, the item response theory approach makes sense. It is not our intention to criticize item response theory generally or to say that it detects "wrong" interactions, but to demonstrate that these interactions do not provide a comparison of changes independently of the pretest distribution, as defined by comparing conditional distributions given the pretest values.

Poststratified relative change scores

The idea of the poststratified relative change scores (PRCS) is the aggregation of measures for the changes of the item scores belonging to the property of interest relative to the changes of the other properties conditional on their pretest values. Note that PRCS are of a non-parametric nature, i.e., they are not derived as estimators of some quantity in a parameterized model. This implies in particular that there is no "true value plus measurement error" formulation. Neither true but unobserved scores nor measurement errors are quantified. However, the resulting score values provide a directly interpretable measure of the size of the differences between within-subject changes.

The model to start is simply the whole common distribution of the random vector $(X_{hijk})_{h=0,1,\ i=1,\ldots,I,\ j=1,\ldots,J_i}$, assumed to be i.i.d. over the test persons k. The null hypothesis of no difference between the changes of the items is operationalized by

$$\begin{aligned}\mathbf{H_0}: \quad & \forall x \in \{1,\ldots,P\},\ i=1,\ldots,I,\ j=1,\ldots J_i: \\ & E(X_{1ij1}|X_{0ij1}=x) = c(x), \end{aligned} \tag{9}$$

$c(x)$ being a value that only depends on x, i.e., all item's posttest means are equal conditional under all pretest values (the condition $X_{0ij1}=x$ means that the pretest value corresponding to the posttest value X_{1ij1} is x, and the H_0 states that Equation 9 holds for all x,i,j). This hypothesis may seem rather restrictive, but note that the hypothesis $\tau_{1i}=c$ in the model Eq. 6 induces an equality between some functions of such conditional expectations as well, which are more difficult to interpret, because they depend also on the pretest value distribution. Note further that H_0 is formulated in terms of raw scores while the discussed item response theory approaches operate on logit or probit scaled probabilities. We don't claim that it is superior in general to consider the raw scores. We use the comparison of conditional expectations as a practical simplification of the comparison of the full conditional distributions, which is straighforward and easy to interpret. It would be conceivable to compare other functionals of the conditional distributions such as functions of the probability logits, but this would not lead to conventional item response approaches, as has been demonstrated in the previous section. See the last section for more discussion of the scaling issue.

PRCS enable an asymptotically unbiased test of H_0 against the alternative that all item's conditional posttest means of property i are larger or equal than the other's properties means (Eq. 10; conditioning is again on the corresponding pretest values), and that there is at least one pretest value conditional under which a nonzero difference can be observed with probability larger than 0 (Eq. 11):

$$\mathbf{H_1}: \forall x \in \{1,\ldots,P\},\ q \neq i,\ j \in \{1,\ldots,J_i\},\ r \in \{1,\ldots,J_q\}:$$

$$E(X_{1ij1}|X_{0ij1}=x) \geq E(X_{1qr1}|X_{0qr1}=x), \tag{10}$$
$$\exists x \in \{1,\ldots,P\},\ q \neq i,$$
$$j \in \{1,\ldots,J_i\},\ r \in \{1,\ldots,J_q\},\ P\{X_{0ij1}=x,\ X_{0qr1}=x\}>0:$$
$$E(X_{1i_0j1}|X_{0i_0j1}=x) > E(X_{1qr1}|X_{0qr1}=x). \tag{11}$$

Equation 10 formulates a one-sided alternative. There is no difficulty to use the same methodology for a one-sided test against the other direction, with "$\leq$" in Equation 10 and "$<$" in Equation 11, or for a two-sided test against the union of these two alternatives. However, it is not possible to replace "$\geq$" in Equation 10 by "$\neq$", because items with larger and smaller conditional expectations under property i may cancel their effects out in the computation of the PRCS.

The PRCS for a test person k and a property of interest i is defined as follows:

1. For each pretest value $x \in I\!N_P$, compute the difference between the mean posttest value over the items belonging to property i and the other properties:

$$\begin{aligned} D_{i\cdot k}(x) &= X_{1i\cdot k}(x) - X_{1-i\cdot k}(x), \text{ where} \\ X_{1i\cdot k}(x) &= \frac{\sum\limits_{j:\ X_{0ijk}=x} X_{1ijk}}{N_{0i\cdot k}(x)}, \\ X_{1-i\cdot k}(x) &= \frac{\sum\limits_{q\neq i, r:\ X_{0qrk}=x} X_{1qrk}}{N_{0-i\cdot k}(x)}, \end{aligned}$$

   $N_{0i\cdot k}(x)$ being the number of items of property i with pretest value x, and $N_{0-i\cdot k}(x)$ being the corresponding number of the other items. If one of these is equal to zero, the corresponding mean posttest value can be set to 0. Note that the sum in the definition of $X_{1i\cdot k}(x)$ is over all values j fulfilling $X_{0ijk}=x$ and the sum in the definition of $X_{1-i\cdot k}(x)$ is over all pairs q,r fulfilling $q\neq i$ and $X_{0qrk}=x$. A dot in the subscipt generally refers to an aggregation (summing up or averaging, depending on the precise definition) of all possible values at this place.

2. The PRCS $\overline{D_{i\cdot k}}$ for person k is a weighted average of the $D_{i\cdot k}(x)$, where the weight should depend on the numbers of items $N_{0i\cdot k}(x)$ and $N_{0-i\cdot k}(x)$ on which the difference is based:

$$\overline{D_{i\cdot k}} = \frac{\sum\limits_{x=1}^{P} w(N_{0i\cdot k}(x), N_{0-i\cdot k}(x)) D_{i\cdot k}(x)}{\sum\limits_{x=1}^{P} w(N_{0i\cdot k}(x), N_{0-i\cdot k}(x))}. \tag{12}$$

   The weights should be equal to 0 if either $N_{0i\cdot k}(x)$ or $N_{0-i\cdot k}(x)$ is 0, and >0 else. It is reasonable to assume that the denominator of $\overline{D_{i\cdot k}}$ is >0. Otherwise, there is no single pair of items for property i and any other property with equal pretest values, and therefore the changes of property i cannot be compared to the changes of the other properties for this test person. In this case, person k should be excluded from the analysis. The weights are suggested to be taken as

$$w(n_1,n_2) = \frac{n_1 n_2}{n_1+n_2}, \tag{13}$$

   as motivated by Lemma 1 below. Obviously it makes sense to weight up pretest values x which occur more often for the given person, because the corresponding $D_{i\cdot k}$-values are

more informative (it can easily be checked that Equation 13 achieves this). The weights may be chosen more generally as dependent also on the value of x itself, if this is suggested by prior information.

Inference can now be based on the values $\overline{D_{i\cdot k}}, k = 1, \ldots, K$.

The null hypothesis to be tested is $E\overline{D_{i\cdot k}} = 0$ with a one-sample t-test. The underlying theory of this test is presented below. In Theorem 1 it is shown that the test statistic $\sum_k \overline{D_{i\cdot k}}$ standardized by a variance estimator (see below) is asymptotically normal with variance of 1 under both hypotheses, expected value 0 under H_0 and a larger expected value under H_1. In other words, a one-sided test of $E\overline{D_{i\cdot k}} = 0$ based on the standardized statistic is asymptotically unbiased for H_0 against H_1 under the assumptions Eqs. 14-16 given below.

The t_{K-1}-distribution is asymptotically equivalent to the normal distribution and can often be expected to be a better approximation for finite samples. The reason is that the value range is bounded. If the values are not strongly concentrated far from the bounds (which may be checked by graphical methods), the distribution of $\overline{D_{i\cdot k}}$ will have lighter tails than the normal distribution. This results in heavier tails of the distribution of the test statistic than expected under the normal according to Cressie (1980). Since the t_{K-1}-distribution has heavier tails than the normal, it will match the distribution of the test statistic better in most situations. The one-sample-Wilcoxon- or sign test may also be considered as alternatives, but for finite samples it can neither be guaranteed that the distribution of $\overline{D_{i\cdot k}}$ is symmetric, nor that the median is 0 under H_0. The simulations indicate a superior power of the t-test.

Note that H_0 is fulfilled in Example 1 with $x = 0, 1,\ c(0) = 0.9,\ c(1) = 1$, assuming that the two items correspond to two different properties. Test persons would only be included in the comparison if the pretest values of the two items were equal (because there is only one item for each property; otherwise the denominator of Equation 12 is zero), which is the reason why the problems demonstrated in the previous sections don't occur. Excluding some persons may look like a drawback, but it is actually a sensible strategy, because subjects with a pretest outcome of 1 on one item and 0 on the other one don't provide useful information concerning a within-subject comparison of change. In this case, $N_{0i\cdot k}(x) = 1$ and $E(D_{i\cdot k}(x)) = 0$ because the conditional distributions given $X_{0i11} = x$ are the same for both items $i = 1, 2$.

The theory needs the following assumptions:

$$\exists x \in \{1, \ldots, P\},\ q \neq i,\ j \in \{1, \ldots, J_i\},\ r \in \{1, \ldots, J_q\} :\ \forall (x_1, x_2) \in \{1, \ldots, P\}^2 :$$
$$P\{X_{0ijk} = x,\ X_{0qrk} = x\} > 0, \tag{14}$$
$$P\{X_{1ijk} = x_1,\ X_{1qrk} = x_2\} < 1, \tag{15}$$
$$\forall x \in \{1, \ldots, P\},\ i \in \{1 \ldots, I\},\ j \in \{1, \ldots, J_i\} :\ X_{1ijk} \text{ independent of } (X_{0qrk})_{qr}$$
$$\text{conditional under } X_{0ijk} = x. \tag{16}$$

Assumption Eq. 14 ensures the existence of at least one item for property i and some other property such that the changes are comparable conditional under a given $X_{0ijk} = x$. Equation 15 excludes the case that all comparable posttest values are deterministic. In that case, statistical methods would not make sense. The only critical assumption is Eq. 16, which means that X_{1ijk} is allowed to depend on $(X_{0qrk})_{qr}$ (denoting the whole pretest result of test person k) only through X_{0ijk}. Similar restrictions are implicit in the LPCM Eq. 6 and, on the Likert mean score level, in the linear models Eqs. 1 and 2. Moreover, the PRCS test does allow for arbitrary dependence structures among the pretest values as opposed to the latter three models.

Theorem 1 *Assume Eqs. 14-16. For* $\overline{D_{i\cdot k}}$ *as defined in Equation 12,*

$$\left(\frac{\sum_{k=1}^{K} \overline{D_{i\cdot k}}}{(KS_K^2)^{1/2}} \right) \text{ converges in distribution to } \mathcal{N}(a_j, 1),\ j = 0, 1, \tag{17}$$

under H_j *with* $a_0 = 0$, $a_1 > 0$, *where* S_K^2 *is some strongly consistent variance estimator, e.g.* $S_K^2 = \frac{1}{K-1} \sum_{k=1}^{K} \left(\overline{D_{i\cdot k}} - \frac{1}{K} \sum_{q=1}^{K} \overline{D_{i\cdot q}} \right)^2$.

The proof is given in the Appendix.

An optimal choice of the weight function w depends on the alternative hypothesis. For example, if differences between the changes in property i and the other properties would only be visible conditional under a single particular pretest value of x, this x would need the largest weight, but of course such information has to be obtained independently of the observed data if it should be used in the definition of the weights.

To construct a reference alternative model, we assume that all items and pretest values behave in the same manner conditional under the pretest value. More precisely, it is assumed that for property i the conditional expectation $E_{i,x}$ of the posttest values doesn't depend on the item, and that for all other properties, all pretest values and items, the conditional expectation of the posttest value is by a fixed constant c smaller than $E_{i,x}$. All variances of the conditional posttest value distributions are assumed to be the same (conditioning is as usually on the corresponding pretest value):

$$\begin{gathered}
\forall x \in \{1, \ldots, P\},\ j \in \{1, \ldots, J_i\} : \\
E(X_{1ijk} | X_{0ijk} = x) = E_{i,x} \text{ independent of } j, \\
\forall x \in \{1, \ldots, P\},\ q \neq i, r \in \{1, \ldots, J_q\} : \\
E(X_{1qrk} | X_{0qrk} = x) = E_{i,x} - c,\ c \text{ independent of } q, r, x, \\
\forall x \in \{1, \ldots, P\},\ q = 1, \ldots, I,\ r \in \{1, \ldots, J_q\} : \\
\mathrm{Var}(X_{1qrk} | X_{0qrk} = x) = V \text{ independent of } q, r, x.
\end{gathered} \tag{18}$$

The optimality result needs a further independence condition additionally to assumption Eq. 16, namely that, given $X_{0ijk} = x$, a posttest value for a person is even independent of the posttest values of the same person for the other properties ($(X_{1qrk})_{-i}$ denoting the vector of posttest values for person k excluding property i):

$$\begin{gathered}
\forall x \in \{1, \ldots, P\},\ i \in \{1 \ldots, I\},\ j \in \{1, \ldots, J_i\} :\ X_{1ijk} \text{ independent of } (X_{1qrk})_{-i} \\
\text{conditional under } X_{0ijk} = x.
\end{gathered} \tag{19}$$

Keep in mind that the assumptions Eqs. 18 and 19 do not restrict the applicability of the PRCS method, but are only needed to define a reference alternative that can be used to find an optimal weight function (which should still be reasonable under most other alternatives).

Lemma 1 *For* $\overline{D_{i\cdot k}}$ *as defined in Equation 12 and* H_1 *fulfilling Equations 16, 18 and 19,* a_1 *from Theorem 1 is maximized by the weight function* w *given in Equation 13.*

The proof is given in the appendix.

Both the PRCS and the Likert mean scores are relatively weakly affected by missing values in single items. They can be simply left out for the computation of the means.

For exploratory purposes, the mean values of the $\overline{D_{i\cdot k}}$ may be considered for all properties $i = 1, \ldots, I$ and can be interpreted directly in terms of the category values $1, \ldots, P$ as "relative effect sizes", namely as properly weighted averages of the differences in changes. The PRCS may also be used to test the equality of changes in property i between different groups with a two-sample t-test or a more general ANOVA.

Reliability of the PRCS can be assessed by the usual split-half method. The items should be split in such a way that all properties are represented by the same number of items in both halves.

Application: effects of music on emotions

The study

At least since the days of Immanuel Kant, it is a widespread notion that music bears a close relationship to human emotions. Kant articulated in his "Kritik der Urteilskraft" that music 'speaks' through felt sensations and could therefore be seen as a language of affects (Kant, 1957, §53).

Given this important function of music, it is not surprising that in the last decades many studies in music psychology and music perception tried to clarify the relationship between music or musical features and the evocation of emotions (for an overview see for example Scherer and Zentner, 2002). Difficulties arise in this research area from the lack of a unified theoretical framework for music and emotions (e.g. Kleinginna and Kleinginna, 1981, Sloboda and Juslin, 2002) and from problems with the measurement of emotions or emotional changes caused by music listening (Madsen, 1996; McMullen, 1996; Müllensiefen, 1999; Schubert, 2002). Many empirical findings concerning music and its emotional effects are seemingly contradictory and unrelated. Among the more important reasons for this unsatisfying state of empirical knowledge are the idiosyncratic nature of emotional reactions to a wide range of aesthetic stimuli and the difficulty to control all the intervening variables in the measurement of emotional responses. To get a clearer picture of how music can induce emotional changes, a tool for the measurement of emotional change due to music listening was developed and applied in a study with high school students (Bargmann, 1998). The scope of the study has been restricted to the subjective aspects of emotions, as could be articulated verbally via a questionnaire. Physiological and gestural measurements have not been taken into account.

The original study used a semantic differential to measure the emotional states of the subjects before and after the music treatment. The semantic differential itself consisted of 50 self-referential statements that belonged to five emotional states, ten statements (items) for each state. The 50 items and their respective emotional states were selected according to the results of an extensive pretest. The results of this pretest indicated five emotional categories (called "properties" in the statistical part): joy, sadness, love, anger, and anxiety/fear. The semantic differential with its 50 items was used to evaluate the momentary state for each subject in each of the five emotional categories. The answers have been given on a five-point Likert scale as explained in the Introduction.

From a second pretest with several music examples, the instrumental piece "Bait and Chase" from the motion picture soundtrack "Alien III" was chosen as a piece that evoked strongest and most homogeneous emotions, judged by direct statements of the test persons of the pretest. It is characterized by dissonant orchestral sounds that are distorted by a lot of noise elements. It lacks an identifiable melody as well as a recognizable structure. Its associated emotional quality was anxiety/fear.

The subjects in the main study were 125 students aged 16 to 19 from two different high

schools in northern Germany. They were tested in groups in their usual classroom environment to minimize disturbing influences of laboratory testing on their emotional conditions.

The design consisted of six groups: group E with $K = 24$ was the experimental group that received the treatment (music listening) between the pretest and posttest rating of the semantic differential. Group E (Counter Demand) with $K = 20$ received the same treatment and made pretest and posttest ratings exactly like group E. The two groups differed only in the instructions given with the music example. While the instructions for group E were neutral concerning the measurement of the subjects' emotions, subjects in group E (CD) were suggested that the music example evoked joy in prior tests. The idea of the counter demand group is to evaluate the effect of the experimental instruction.

The first control group C1 with $K = 18$ received the pretest and had to complete a verbal task instead of the music treatment. As in group E, this was followed by the survey of the individual emotional state on the semantic differential in the posttest.

As in the "Solomon four group design" (Solomon, 1949), there have been three other control groups without pretest to control the effect of pretest sensitization. The statistical evaluation of these groups was by means of standard methodology and is not further discussed here.

The main research hypotheses of the experiment were that

1. the Alien III music example would increase the ratings of the items associated with anxiety/fear from pretest to posttest scores in groups E and E (CD). The increase should be stronger than any increase of the other ratings. Thus, it has been expected that the null hypothesis of no difference in the changes would be rejected.

2. the changes of the anxiety ratings should not differ from changes of the other categories from pre- to posttest in the C1 group where there was no treatment.

Furthermore, there have been comparisons between the posttest scores of the different groups. Because of a lack of time and resources, the assignment of students to the groups has not been randomized, so that a reliable conclusion from the study results to causal effects is not possible. The aim of the study was rather to demonstrate the suitability of the semantic differential. However, the results have been validated by subsequent qualitative interviews with a different group of subjects.

Results

The results of the analysis for the Alien data are as follows: In the experimental group E, the t-tests for $\mu = 0$ and i being the anxiety score under the model Eq. 4 leads to p-values of 0.00055 (unrestricted) and < 0.0001 under assumption Eq. 5. All tests are one-sided, unless indicated explicitly. The means of the PRCSs are 0.6207 (anxiety), -0.4882 (joy), -0.3450 (love), 0.1099 (sadness) and 0.2731 (anger). The t-test for the anxiety mean to be equal to zero leads to $p < 0.0001$. Not only is the change in anxiety clearly significant compared with the other changes, but the PRCS has also the largest absolute value. With the same methodology, anxiety could also be tested against every single one of the other emotions, the data being restricted to the items of two emotions for all these tests. The largest of the resulting p-values is 0.006 (anger), which is still significant at the 5%-level compared to the Bonferroni-p of 0.05/5.

In the experimental group E (Counter Demand), the regression t-test p-values for anxiety are 0.223 (unrestricted) and 0.265 under assumption Eq. 5. The means of the PRCSs are 0.0645 (anxiety), 0.2547 (joy), 0.1825 (love), -0.1488 (sadness) and -0.0324 (anger). The t-test for the anxiety mean to be equal to zero leads to $p = 0.2894$. As opposed to the research hypothesis, the different experimental instructions compared to group E seem to destroy the

effect on anxiety. The effect on joy has the largest absolute value, but it is also not significant ($p = 0.1218$).

In the control group C1, the changes in anxiety are negative, so that the one-sided tests do never reject the H_0. Thus, the two-sided p-values are reported and discussed. These are not of primary interest in the study, but they can be used to highlight differences between the PRCS and the multivariate regression method. The regression t-test for anxiety leads to $p = 0.0779$ (unrestricted) and $p = 0.0099$ under assumption Eq. 5. The means of the PRCSs are -0.1545 (anxiety), 0.8328 (joy), 0.1643 (love), -0.3065 (sadness) and 0.1102 (anger). The t-test for the anxiety mean to be equal to zero leads to $p = 0.0174$. The restricted regression test and the test based on PRCS detect a weakly significant decrease in anxiety, and it can also be shown that anxiety is significantly more decreased as in group E (CD) by a two-sample t-test applied to the PRCS (two-sided $p = 0.0495$), which could be interpreted as detecting a positive effect of "Alien III" on anxiety in the E (CD) group in comparison to no treatment. However, the difference between the effects on joy in these two groups (two-sided $p = 0.0371$) seems to be dominant.

The unrestricted regression test does not lead to a significant result in the group C1, and it may be wondered why the different tests lead to different conclusions and which result is most reliable. Table 1 gives the distributions of the posttest values conditional on all pretest values, computed separately for the items belonging to anxiety/fear and all other items. The PRCS test tests the H_0 that the two conditional distributions for each of the five pretest values are equal, against a uniformly smaller (or larger) conditional expectation for anxiety/fear. Although the table does not contain information about items belonging to the same test person, the alternative hypothesis is indicated by the smaller conditional mean posttest values and the almost uniformly stochastic smaller conditional distributions of the posttest values for anxiety/fear. This is the information on which the PRCS method operates.

In Figure 1, the pretest and posttest Likert mean scores are plotted, to which the regression methods are applied. The regression methods test the equality of the intercepts of the regression lines belonging to anxiety/fear ("F") and the other scores ("O"). This is difficult to judge by eye. The posttest values of the "O"-score show a relatively large variance even for similar pretest scores, so that the corresponding regression coefficients cannot very precisely be estimated. The assumption Eq. 5 (both regression lines are parallel) reduces the variation in the estimation of the regression coefficient by aggregating information of both scores. A likelihood ratio test to compare the model under assumption Eq. 5 with the unrestricted model yields $p = 0.446$, so that the data do not contradict assumption Eq. 5. However, the two regression lines cannot be compared very well, because many "F"-pretest scores are so small that there are no "O"-pretest scores to compare them to, and many "O"-pretest scores are concentrated at about 0.2, where there are few "F"-pretest scores. We conclude that a test based on the information in Table 1 seems to be more adequate. The difference between the distributions of the pretest values of anxiety/fear and the other properties does not bias the regression test, but seems to cause a loss of power. Note that a test of the time/property interaction in the model in Eq. 1 with $p = 0.868$ obscures any evidence for a difference in changes. This test is related to the posttest-pretest difference between the difference of the total means for fear and others. This difference is $(1.66-2.08)-(1.77-2.15) = -0.04$, which has a smaller absolute value than all five differences between the conditional means (cf. the last three columns of the "total-f/o"-lines of the Table 1).

To assess the reliability of the scores, we carried out 20 random partitions of the items in two halves exemplary for the experimental group. **We give two mean correlation coefficients for every property. The first one refers to the correlation of the test person's PRCS values of both halves, the second one refers to the test person's differences**

between pretest and posttest Likert scores of both halves. The correlations are 0.342/0.600 (anxiety), 0.609/0.553 (joy), 0.343/0.165 (love), 0.414/0.510 (sadness) and 0.434/0.621 (anger). The analogous correlations between the raw pretest Likert scores were between 0.6 and 0.87. While the reliability of the PRCS seems to be limited in this example, note that they are directly used for testing while for the tests involving the Likert scores additional variability is introduced because more parameters are estimated in the model underlying the analyses.

Simulations

We carried out a small simulation study to compare the performance of some of the proposed tests. We did not simulate the tests based on the repeated measurement model Eq. 1 and the LPCM Eq. 6, which have been shown to lead to wrong conclusions in the Example 1, and which need by far more computing time than the tests based on the PRCS approach and the multivariate regression Eq. 4. Five tests have been applied:

Regression The t-test for $\mu = 0$ in the multivariate regression Eq. 4 with unrestricted regression parameters.

RegrRestrict The t-test for $\mu = 0$ in the multivariate regression Eq. 4 under the assumption Eq. 5.

RCS-t The one-sample t-test with PRCS for $E\overline{D_{i\cdot k}} = 0$.

RCSWilcoxon The one-sample Wilcoxon test for symmetry of the distribution of the PRCS about 0.

RCSsign The sign test for $\text{Med}\overline{D_{i\cdot k}} = 0$.

All simulations have been carried out with $K = 20,\ P = 5,\ I = 5,\ J_i = 10,\ i = 1, \ldots, 5$, and property 1 has been the property of interest, i.e., a situation similar to the Alien data. Emphasis is put to the effect of different pretest distributions between property 1 and the other properties. We simulated from three different setups under the null hypothesis and three different setups under the alternative:

standard Uniform distribution on $\{1, \ldots, 5\}$ for all pretest values. Each posttest value has been equal to the corresponding pretest value with probability 0.4, all other posttest values have been chosen with probability 0.15. (H_0)

lowPre1 The pretest values for property 1 have been chosen with probabilities $0.3, 0.25, 0.2, 0.15, 0.1$ for the values $1, 2, 3, 4, 5$. The pretest values for the other properties have been chosen with probabilities $0.1, 0.15, 0.2, 0.25, 0.3$ for $1, 2, 3, 4, 5$ (lower pretest values for property 1). The posttest values and the pretest values for the other properties have been chosen as in case `standard`. (H_0)

lowPre1highPost The pretest values have been generated as in case `lowPre1`, the posttest values have been chosen equal to the pretest value with probability 0.4. The remaining probability of 0.6 for the case that the posttest values differ from the pretest values has been distributed as follows: the two highest remaining values have been chosen with probability 0.2, and the two lower values have been chosen with probability 0.1. (H_0)

highPost1 The pretest values and the posttest values for the properties 2-5 have been generated as in case `standard` (no differences in the pretest distribution), the posttest values for property 1 have been generated as in case `lowPre1highPost`. (H_1)

lowPre1highPost1 The pretest values have been generated as in case `lowPre1`, the posttest values have been generated as in case `highPost1`. (H_1)

highPre1highPost1 As for case `lowPre1highPost1`, but with pretest value probabilities of $0.1, 0.15, 0.2, 0.25, 0.3$ for $1, 2, 3, 4, 5$ for the items of property 1 and vice versa for the items of the other properties (higher pretest values for property 1).

The results of the simulation are shown in Table 2. The results for the H_0-cases do not indicate any clear violation of the nominal level. Note that this would be different using tests derived under the models Eqs. 1, 6 and 8. The sign test always appears conservative, and the regression methods are conservative for `lowPre1highPost`. The results for the H_1-cases show that different distributions for the pretest values of property 1 and the other properties result in a clear loss of power of the regression methods compared to the PRCS methods. The two nonparametric tests based on the PRCS perform a bit worse than the t-test under H_1. The linear regression test shows a better power under the assumption Eq. 5 than unrestricted in all cases.

Discussion

Some methods for comparing the changes between different properties measured on Likert scales between pretest and posttest have been discussed. Tests based on a repeated measurement model and item response theory have been demonstrated to depend on the pretest distribution, which, given our conceptualization of the comparison of change based on conditional distributions, means that they are biased under differences in the pretest value distributions. Two proposed tests based on multiple linear regression use the Likert mean scores while the PRCS tests are directly based on the item values. The advantage of the PRCS is that the effect of the pretest scores is corrected by comparing only items with the same pretest value, while the regression approach needs a linearity assumption which is difficult to justify. To work properly, the PRCS approach needs a sufficient number of items, compared with the number of categories for the answers, because the number of comparisons of items with the same pretest values within test persons determines the precision of the PRCS. If there are few items with many categories, the linear regression approach is expected to be superior.

PRCS can more generally be applied in situations where pretest and posttest data are not of the same type. The pretest data must be discrete (not necessarily ordinal) with not too many possible values, the posttest data has to allow for arithmetic operations such as computing differences and sums. Note that it may be disputable if the computation of means for five-point Likert scales is meaningful (the regression/ANOVA methods operate to an even stronger extent on the interval scale level). We think that the precise values of the PRCS have to be interpreted with care. However, no problem arises with the use of the PRCS for hypothesis tests, because this is analogous to computations with ranks as in the Spearman correlation. The only difference is that the effective difference between successive values is governed by the number of possible values in between, and not by the number of cases taking these values.

As opposed to the other methods discussed in this paper, the PRCS method is not based on a parametric model. This has the advantage that no particular distributional shape has to be assumed. On the other hand, while the PRCS values can be interpreted in an exploratory manner, the method does not provide effect parameter estimators and variance decompositions, which can be obtained from linear models. However, we have demonstrated that the effect parameters for other models may be misleading in certain situations.

For all methods, a significant difference in changes for anxiety/fear may be caused not only by the treatment affecting anxiety directly, but also if another property is changed primarily.

Therefore, it is important not only to test the changes of anxiety, but to take a look at the absolute size of the other effects. A sound interpretation is possible for a result as in group E, where the PRCS of anxiety is not only significantly different from zero, but also the largest one in absolute value.

Concern may be raised about the meaning of a comparison of measurement values for different variables (properties and items). The PRCS and the ANOVA-type analyses of Likert scores assume that it is meaningful to say that a change from "agree" to "disagree" for one item is smaller than a change from "agree" to "strongly disagree" for another item corresponding to another property (or for the same item between pretest and posttest). While we admit that this depends on the items in general (and it may be worthwhile to analyze the items with respect to this problem), we find the assumption acceptable in a setup where the categories for the answers are identical for all items and are presented to the test persons in a unified manner, because the visual impression of the questionnaire suggests such an interpretation to the test persons.

Item response theory as presented in Samejima (1969), Andrich (1978), Muraki (1990), and Fischer and Ponocny (1994) addresses such comparisons of categories between different items by model assumptions that formalize them as "difficulties" corresponding to several "abilities". While it could be interesting to apply such approaches to the present data, we think that the concept of a "true distance" between categories that can be inferred from the data is inappropriate for attitudes and feelings. Every data-analytic approach assumes implicitly that the "true distance" between categories is determined by the probability distributions of the values of the test persons belonging to these categories. This idea seems to be directly related to the idea of the difficulty of ability tests, which can indeed be adequately formalized by probabilities of subjects solving a particular task. Because in these situations the pretest distribution is meaningful in terms of the difficulty of tasks, the dependence of the measurement of change on the pretest distribution as demonstrated in Example 1 can be seen as sensible.

However, in the present setup, we don't see a clear relationship between the distribution of test person's answers to any underlying "true distance". A more reasonable way to examine such a distance would be a survey asking people directly for their subjective concepts of between-category distances.

An example for between-subjects comparisons of different variables on a different topic is the work of Liotti et al., 2000, who compare the activities in different regions of the human brain connected to the induction of certain emotions.

An important difference between the approaches discussed here is the scaling on which the methods operate. The linear models and the PRCS operate on the raw scores scale, while the item response theory approaches operate on logit or probit transformed probabilities. We do not claim that one of these scales is generally better than the other. However, we emphasize that our concept of "equality of changes" defined by the equality of all pretest value-conditional distributions of the posttest values between the properties is independent of the scale, because it doesn't depend on any scaling whether two distributions are equal or not.

The scaling issue arises for the PRCS on two levels. First, the distributions to be compared are conditional on the raw score values. This is inevitable: in probability theory distributions are always conditioned on the outcomes of random variables, not on the probabilities (however scaled) of these outcomes. Second, the alternative hypothesis of the PRCS approach models differences between expected values of raw scores, because in practice the difference between changes has to be measured by a suitable test statistic, for which we have chosen the average of raw score values. Other alternative hypotheses and other statistics, measuring the differences between the conditional distributions on other scales, are conceivable and leave opportunities for further research.

Appendix

Proof of Theorem 1: The $\overline{D_{i\cdot k}}$ are weighted averages of differences between bounded random variables and assumed to be i.i.d. over k. Therefore, $\mathrm{Var}\overline{D_{i\cdot k}} < \infty$ and $\overline{D_{i\cdot k}}, k \in I\!N_K$ i.i.d. S_K^2 converges almost surely to $\mathrm{Var}\overline{D_{i\cdot k}} > 0$ because of Equation 15. Thus, the central limit theorem ensures convergence to normality. It remains to show that $E\overline{D_{i\cdot 1}} = 0$ under H_0 and $E\overline{D_{i\cdot 1}} > 0$ under H_1.

Let $\tilde{x} \in \{1, \ldots, P\}^{J_1+\ldots+J_I}$ be a fixed pretest result. Under $(X_{0qr1})_{qr} = \tilde{x}$, define $n_i(x, \tilde{x}) = N_{0i\cdot 1}(x)$, analogously $n_{-i}(x, \tilde{x})$. Let $w_{x,\tilde{x}}$ be the corresponding value of the weight function. By assumption Eq. 16,

$$a(x, \tilde{x}) = E\left(D_{i\cdot 1}(x) | (X_{0qr1})_{qr} = \tilde{x}\right) =$$

$$= \tfrac{1}{n_i(x,\tilde{x})} \sum_{j:\ X_{0ij1}=x} E(X_{1ij1}|X_{0ij1} = x) - \frac{1}{n_{-i}(x, \tilde{x})} \sum_{(q,r)}^{n.(x,\tilde{x})} E(X_{1qr1}|X_{0qr1} = x)$$

unless $n_i(x, \tilde{x}) = 0$ or $n_{-i}(x, \tilde{x}) = 0$, in which case $w_{x,\tilde{x}} = 0$. Further,

$$E\left(\overline{D_{i\cdot 1}}\right) = E\left[E\left(\overline{D_{i\cdot 1}}|(X_{0qr1})_{qr} = \tilde{x}\right)\right] = E\left[\frac{\sum_{x=1}^{P} w_{x,\tilde{x}} a(x, \tilde{x})}{\sum_{x=1}^{P} w_{x,\tilde{x}}}\right]. \tag{20}$$

Under H_0, $a(x, \tilde{x}) = 0$ regardless of x and $\tilde{x}$. Under H_1, always $a(x, \tilde{x}) \geq 0$ and "$>$" with positive probability under the distribution of $(X_{0qr1})_{qr}$ for some x with $w(x, \tilde{x}) > 0$.

Proof of Lemma 1: The following well-known result can be shown by analogy to the Gauss-Markov theorem: If $Y_1, \ldots, Y_K$ are independent random variables with equal mean c and variances V_r, $r = 1, \ldots, K$, then the weighted mean $\frac{1}{\sum_{r=1}^{K}(1/V_r)} \sum_{r=1}^{K} \frac{Y_r}{V_r}$ has minimum variance among all unbiased estimators of c that are linear in the observations.

The notation of the proof of Theorem 1 is used. Observe $a(x, \tilde{x}) = c$ under assumption Eq. 18 regardless of x and $\tilde{x}$ unless $w_{x,\tilde{x}} = 0$. Therefore, $E\left(\overline{D_{i\cdot 1}}|(X_{0qr1})_{qr} = \tilde{x}\right) = c$ by Equation 20. From assumptions Eqs. 18 and 19,

$$\mathrm{Var}\left(D_{i\cdot 1}(x)|(X_{0qr1})_{qr} = \tilde{x}\right) = \left(\frac{1}{n_i(x, \tilde{x})} + \frac{1}{n_{-i}(x, \tilde{x})}\right) V =: V_D.$$

Since $\overline{D_{i\cdot 1}}$ is linear in the $D_{i\cdot 1}(x)$ for given $\tilde{x}$, its conditional variance is minimized by choosing the weights $w_{x,\tilde{x}} = 1/V_D = \frac{n_i(x,\tilde{x}) n_{-i}(x,\tilde{x})}{V(n_i(x,\tilde{x}) + n_{-i}(x,\tilde{x}))}$. V is independent of x and can be reduced in Equation 12. Since the conditional expectation of $\overline{D_{i\cdot 1}}$ is independent of $\tilde{x}$, separate minimization of the conditional variances for all $\tilde{x}$ also minimizes the unconditional variance of $\overline{D_{i\cdot 1}}$.

References

Achcar, J. A., Singer, J. M., Aoki, R., Bolfarine, H. (2003) Bayesian analysis of null intercept errors-in-variables regression for pretest/posttest data, *Journal of Applied Statistics 30,* 3-12.

Andersen, E. B. (1985) Estimating latent correlations between repeated testings, *Psychometrika 50,* 3-16.

Andrich, D. (1978) A binomial latent trait model for the study of Likert-style attitude questionnaires, *British Journal of Mathematical and Statistical Psychology 31,* 84-98.

Bajorski, P. and Petkau, J. (1999) Nonparametric Two-Sample Comparisons of Changes on Ordinal Responses, *Journal of the American Statistical Association, 94,* 970-978.

Bargmann, J. (1998) *Quantifizierte Emotionen - Ein Verfahren zur Messung von durch Musik hervorgerufenen Emotionen,* Master thesis, Universität Hamburg.

Bonate, P. L. (2000) *Analysis of Pretest-Posttest Designs,* Chapman & Hall, Boca Raton.

Cressie, N. (1980) Relaxing assumptions in the one-sample t-test, *Australian Journal of Statistics 22,*143-153.

Cribbie, R. A. and Jamieson, J. (2004) Decreases in Posttest Variance and the Measurement of Change, *Methods of Psychological Research Online 9,* 37-55.

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

Dimitrov, D. M. and Rumrill, P. D. (2002) Pretest-posttest designs and measurement of change, *Work 20,* 159-165.

Eid, M. and Hoffmann, L. (1998) Measuring Variability and Change with an Item Response Model for Polytomous Variables, *Journal of Educational and Behavioral Statistics 23,* 193-215.

Embretson, S. E. (1991) A multidimensional latent trait model for measuring learning and change, *Psychometrika 56,* 495-515.

Fischer, G. H. (1976) Some probabilistic models for measuring change. In: DeGruijter, D. N. M. and Van der Kamp, L. J. T. (eds.) *Advances in Psychological and Educational Measurement,* Wiley, New York, 97-110.

Fischer, G. H. and Ponocny, I. (1994) An extension of the partial credit model with an application to the measurement of change, *Psychometrika 59,* 177-192.

Fischer, G. H. (2003) The precision of gain scores under an item response theory perspective: A comparison of asymptotic and exact conditinal inference about change, *Applied Psychological Measurement 27,* 3-26.

Jaccard, J. and Wan, C. K. (1996) *LISREL approaches to interaction effects in multiple regression.* Sage Publications, Thousand Oaks.

Jamieson, J. (1995) Measurement of change and the law of initial values: A computer simulation study. *Educational and Psychological Measurement 55,* 38-46.

Kant, I. (1790) Kritik der Urteilskraft. In: Weischedel, W. (ed.) *Werke, Vol. V.* Wissenschaftliche Buchgesellschaft, Darmstadt [1957].

Kleinginna, P. R. and Kleinginna, A. M. (1981) A categorized list of emotion definitions, with suggestions for a consensual definition. *Motivation und Emotion, 5,* 345-379.

Likert, R. (1932) A Technique for the Measurement of Attitudes, *Archives of Psychology, 140,* 1-55.

Liotti, M., Mayberg, H. S., Brannan, S. K., McGinnis, S., Jerabek, P. and Fox, P. T. (2000) Differential limbic-cortical correlates of sadness and anxiety in healthy subjects: implications for affective disorders, *Biological Psychiatry 48,* 30-42.

Madsen, C. K. (1996) Empirical investigation of the 'aesthetic response' to music: musicinas and non-musicians. In: Pennycook, B. and Costa-Giomi, E. (ed.) *Proceddings of the Fourth International Conference of Music Perception and Cognition.*McGill University, Montreal. 103-110.

Masters, G. N. (1982) A Rasch model for partial credit scoring,*Psychometrika 47,*149-174.

McMullen, P.T. (1996) The musical experience and affective/aesthetic responses: a theoretical framework for empirical research. In: Hodges, D.A. (Ed.) *Handbook of Music Psychology.* IMK Press, San Antonio, 387-400.

Müllensiefen, D. (1999) Radikaler Konstruktivismus und Musikwissenschaft: Ideen und Perspektiven. *Musicae Scientiae Vol. III, 1,* 95-116.

Muraki, E. (1990) Fitting a polytomous item response model to Likert-type data, *Applied Psychological Measurement 14,*59-71.

Raykov, T. (1992) Structural models for studying correlates and predictors of change. *Australian Journal of Psychology 44,* 101-112.

Samejima, F. (1969) Estimation of latent ability using a response pattern of graded scores. *Psychometric Monograph, No. 17, 34, Part 2.*

Scherer, K. R. and Zentner, M. R. (2002) Emotional effects of music: mproduction rules. In: Juslin, P. N. and Sloboda, J. A. (ed.) *Music and Emotion: Theory and Research.* Oxford University Press, Oxford. 361-392.

Schubert, E. (2002) Continous measurement of self-report emotional response to music. In: Juslin, P. N. and Sloboda, J. A. (ed.) *Music and Emotion: Theory and Research.* Oxford University Press, Oxford. 393-414.

Sloboda, J. A. and Juslin, P. N. (2002) Psychological persepctives on music and emotion. In: Juslin, P. N. and Sloboda, J. A. (ed.) *Music and Emotion: Theory and Research.*Oxford University Press, Oxford. 71-104.

Solomon, R. L. (1949) An extension of control group design. *Psychological Bulletin 46,* 137-150.

Steyer, R., Eid, M. and Schwenkmezger, P. (1997) Modeling True Intraindividual Change: True Change as a Latent Variable, *Methods of Psychological Research Online 2,* 21-33.

Wang, W.-C. and Chyi-In, W. (2004) Gain score in item response theory as an effect size measure, *Educational and Psychological Measurement 64,* 758-780.

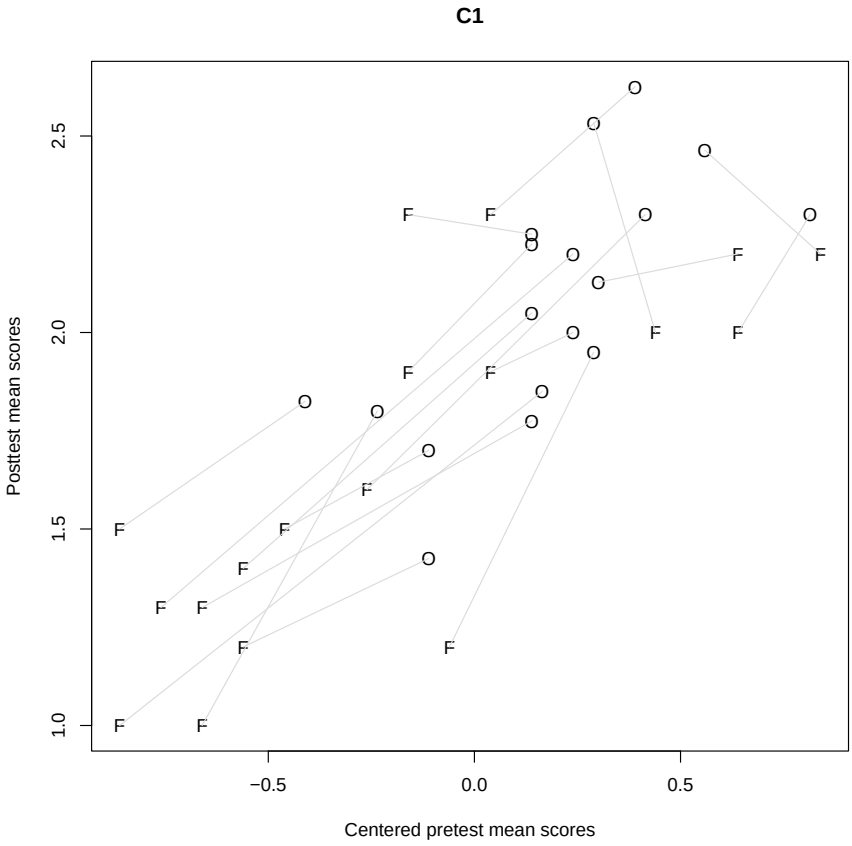

Figure 1: Pretest Likert mean scores vs. posttest Likert mean scores of all test persons of group C1. "F" indicates items belonging to anxiety/fear and "O" indicates the itmes belonging to the other categories. The "F" and the "O" of the same test person are connected by a gray line.

Pretest values	Posttest values 1		2		3		4		5		conditional mean	Pretest distributions fear	other
	n	%	n	%	n	%	n	%	n	%			
1-fear	83	81.4	16	15.7	3	2.9	0	0.0	0	0.0	1.22	56.7	39.7
1-other	207	72.6	63	22.1	10	3.5	4	1.4	1	0.4	1.35		
2-fear	11	32.4	18	52.9	3	8.8	2	5.9	0	0.0	1.88	18.0	26.1
2-other	55	29.4	92	49.2	35	18.7	5	2.7	0	0.0	1.95		
3-fear	7	24.1	8	27.6	13	44.8	1	3.4	0	0.0	2.28	16.1	17.7
3-other	19	15.0	39	30.7	53	41.7	15	11.8	1	0.8	2.53		
4-fear	1	7.7	4	30.8	5	38.5	3	23.1	0	0.0	2.77	7.2	12.7
4-other	1	1.1	18	19.8	26	28.6	37	40.7	9	9.9	3.38		
5-fear	0	0.0	0	0.0	1	50.0	0	0.0	1	50.0	4.00	1.1	3.8
5-other	0	0.0	4	14.8	3	11.1	6	22.2	14	51.9	4.11		
total-f	102	56.7	46	25.6	25	13.9	6	3.3	1	0.6	1.66	mean	mean
total-o	282	39.3	216	30.1	127	17.7	67	9.3	25	3.5	2.08	1.77	2.15

Table 1: Conditional distributions of all posttest values in the group C1 given the pretest values. Left numbers are case numbers, right numbers are percentages, computed along the rows. The percentages are computed separately for anxiety/fear and other items conditional on the pretest value. Only the percentages of the pretest distributions (last two columns) are computed along the columns. The total n is 180 for fear, 717 for others. The last two columns of the last line give the means of the pretest distributions, the "conditional mean" entry in the "total-f/o"-lines give the means of the posttest distributions.

	`Regression`	`RegrRestrict`	`RCS-t`	`RCSWilcoxon`	`RCSsign`
`standard`	0.050	0.049	0.055	0.050	0.033
`lowPre1`	0.044	0.037	0.053	0.050	0.038
`lowPre1highPost`	0.039	0.036	0.049	0.053	0.043
`highPost1`	0.517	0.574	0.516	0.491	0.340
`lowPre1highPost1`	0.107	0.115	0.468	0.444	0.301
`highPre1highPost1`	0.115	0.142	0.483	0.465	0.318

Table 2: Simulated probability of rejection of H_0 from 1000 simulation runs. The nominal level has been 0.05.