

Applied Psychological Measurement

<http://apm.sagepub.com>

Within-Subject Comparison of Changes in a Pretest-Posttest Design

Christian Hennig, Daniel Müllensiefen and Jens Bargmann

Applied Psychological Measurement 2010; 34; 291

DOI: 10.1177/0146621608329889

The online version of this article can be found at:

<http://apm.sagepub.com/cgi/content/abstract/34/5/291>

Published by:



<http://www.sagepublications.com>

Additional services and information for *Applied Psychological Measurement* can be found at:

Email Alerts: <http://apm.sagepub.com/cgi/alerts>

Subscriptions: <http://apm.sagepub.com/subscriptions>

Reprints: <http://www.sagepub.com/journalsReprints.nav>

Permissions: <http://www.sagepub.com/journalsPermissions.nav>

Citations <http://apm.sagepub.com/cgi/content/refs/34/5/291>

Within-Subject Comparison of Changes in a Pretest-Posttest Design

Applied Psychological Measurement
34(5) 291–309
© The Author(s) 2010
Reprints and permission:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0146621608329889
<http://apm.sagepub.com>



Christian Hennig¹, Daniel Müllensiefen²,
and Jens Bargmann³

Abstract

The authors propose a method to compare the influence of a treatment on different properties within subjects. The properties are measured by several Likert-type-scaled items. The results show that many existing approaches, such as repeated measurement analysis of variance on sum and 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, but in the present article, change is measured in terms of the conditional distributions of posttest values, given the pretest values. A multivariate regression and analysis of covariance approach is unbiased but shows power deficiencies in a simulation study. The authors suggest a new approach 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

In the present article, the analysis of data of the following form is addressed: I properties (e.g., attitudes or emotional states) of K test persons are measured by J_i , $i = 1, \dots, I$, items (usually the J_i are the same for all properties, but this does not 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. This article suggests analyzing such data by a new approach based on poststratified relative change scores (PRCS), first starting with the discussion on already existing methodology.

¹University College London, United Kingdom

²Goldsmiths College London, United Kingdom

³Musikwissenschaftliches Institut, Universität Hamburg, Germany

Corresponding Author:

Christian Hennig, Department of Statistical Science, University College London, Gower Street, London, WC1E 6BT, United Kingdom

Email: chrish@stats.ucl.ac.uk

Although there is a lot of literature on measuring change within pretest-posttest data (e.g., Achcar, Singer, Aoki, & Bolfarine, 2003; Andersen, 1985; Cronbach & Furby, 1970; Dimitrov & Rumrill, 2002; Eid & Hoffmann, 1998; Embretson, 1991; Fischer, 1976, 2003; further references can be found in Bonate, 2000), nearly all work concerns the comparison of changes between subjects of different groups. The intention of this article is to compare changes between different variables within the same subject, followed by the Discussion section's presentation on the meaningfulness of such a comparison.

Denote the random variables giving the pretest and posttest values of the items by X_{hikj} , where

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

$i \in IN_I = \{1, \dots, I\}$ denotes the number of the property,

$j \in IN_{J_i}$ denotes the number of an item corresponding to property i (i.e., an item is specified by the pair $[i, j]$),

$k \in IN_K$ 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 persons tested 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, and strongly disagree*). Such an example is treated after the Method sections.

Such properties are frequently measured by Likert-type scales (Likert, 1932)—that is, 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-type 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 & Wan, 1996). Instead of the analysis of covariance, a multivariate regression model is introduced, which is more general and more appropriate for the multiple properties data.

Alternatively, the data can be analyzed on the level of the single item 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. The approach is discussed 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 (Cribbie & Jamieson, 2004; Raykov, 1992; Steyer, Eid, & Schwenkmezger, 1997). 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 analysis of covariance (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 item response theory approach the "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. Using a toy example, it is shown 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 toward the mean in the pretest-posttest literature (cf. Bonate, 2000, chap. 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, whereas 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 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-type 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 deal with the comparison of two independent groups of persons tested.

The *Alien III* data set is introduced and analyzed by the multivariate regression and the PRCS after the Method sections.

Multivariate regression on the pretest values in our setup is also a reasonable strategy to deal with regression toward the mean. However, the fact that it ignores the way the Likert-type 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 our attention is not restricted to a particular model for change or treatment effects. It begins with a data analytic question and compares tests derived from very different models that 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-type mean scores for all values of h, i, k), in which differences between changes of different properties i_1, i_2 are modeled by parameters that specify different expected values of $L_{i_1,k}$ and $L_{i_2,k}$ conditional on $L_{0i_1,k} = L_{0i_2,k}$. 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 occurs 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, whereas the regression, ANOVA, and PRCS methods operate on the raw Likert-type 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 straightforward approach to analyze the Likert-type mean score data is a repeated measures analysis of variance model:

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

where μ is the overall mean, a_h is the effect of time (pretest 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-type mean score of all items not belonging to property i_0 (a subscript with a minus generally denotes aggregation of all possible values of the subscript except the value 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), analogous to the case where difference of changes between groups is of interest (cf. chap. 7 of Bonate, 2000).

Example 1

An extreme but simple example is presented to demonstrate that the model of Equation 1 can indicate a time–property interaction, even if for all pretest values, the conditional distributions of the posttest values are equal between the properties. This is caused by different pretest value distributions for the properties.

Assume that there are only two properties with one item for each, and that 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 always leads 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 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 applying the model of Equation 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-type scores arranged so that the expected values are the same as below. The reason why we used a two-valued example is that this makes the calculations easier (two-valued responses may be associated with techniques like logistic regression).

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

$$EL_{01k} = 0.1 \Rightarrow 0.1 = \mu + b_1,$$

$$EL_{02k} = 0.9 \Rightarrow 0.9 = \mu - b_1,$$

$$EL_{11k} = 0.91 \Rightarrow 0.91 = \mu + a_1 + b_1 + d_{11},$$

$$EL_{12k} = 0.99 \Rightarrow 0.99 = \mu + a_1 - b_1.$$

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 $E(L_{11k}) - E(L_{01k})$ is much larger than $E(L_{12k}) - E(L_{02k})$, which does not reflect a difference between the changes, but instead a pretest distribution of Item 1 that leaves much more space for a positive change.

ANCOVA

In pretest-posttest setups, the related phenomenon of regression toward 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 as follows:

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

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, $\bar{L}_0 = \sum_{i=1}^I \sum_{k=1}^n L_{0ik} / (nI)$ is the overall pretest score mean, and e_{hik} is the error term. Here, the absence of differences in changes is modeled by equal property effects b_i . Some suitable constraints must 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 on 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 person tested 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, it is assumed 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. Also, 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. Also, 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

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

Although 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 the following:

$$\beta_{11} = \beta_{22}, \beta_{12} = \beta_{21} = 0, \text{ thus } \beta_2 = -\beta_1 \text{ in Equation 4.} \quad (5)$$

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 of Equation 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 of Equation 2 and the multivariate regression of Equation 3 is that the latter model allows for a more general within-subject correlation structure (this could also be introduced in the models of Equations 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, this 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-type 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 uses information at the item level. First, a discussion of another item-based method is presented, 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 article. Using our notation, the LPCM is as follows:

$$P(X_{hijk} = x | \theta_k, \delta_{xhij}) = \frac{\exp(x\theta_k + \delta_{xhij})}{\sum_{t=1}^P \exp(t\theta_k + \delta_{thij})}, \quad (6)$$

$$\delta_{xhij} = \beta_{xij} + x\tau_{hi},$$

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 (1994, pp. 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

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

In the present setup, the LPCM, as well as the model of Equation 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, the pair (i, j) in Equation 6 is replaced

by a single index $i = 1, 2$. Also, 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, which yields the following:

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

Solving for the parameters,

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

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 & 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 modeled by an additive constant, $\theta_{12k} = \theta_{11k} + \tau$, the null hypothesis $\tau = 0$ would have to be tested. Also, $\delta_{11i} = \theta_{1ik}$ 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

Another item response model is the graded response model (Samejima, 1969). We are not aware of any literature in which this model has been applied to within-subject comparisons of change. Therefore, our own adaptation for Example 1 is proposed. There are various versions of the graded response model. 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 depending on item and occasion (pretest 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 overparameterized. Because x can only take the values 0 and 1, $\{X_{hik} \geq 1\} = \{X_{hik} = 1\}$, only λ_{1i} is needed. As above, there are only four equations to determine the parameters, but there are still eight parameters. Therefore, four further constraints must be imposed, namely $a_1 = a_2 = 1$, $\lambda_{11} = \lambda_{12} = \lambda$, $\theta_{01k} = 0$. This yields the following:

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

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 aforementioned item response theory approaches, as well as the linear model of Equation 1, model the time–property interaction in a way that depends on the distribution of the pretest values, and the interaction parameter can be nonzero even where all conditional distributions of posttest values are equal between the properties. 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 wrongly detects interactions, but to demonstrate that these interactions do not provide a comparison of changes independent of the pretest distribution as defined by comparing conditional distributions given the pretest values.

PRCS

The idea of the 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 nonparametric 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.

To begin with, the model is simply the whole common distribution of the random vector $(X_{hikj})_{hij}$ $h = 0, 1, i = 1, \dots, I, j = 1, \dots, J_i$, assumed to be i.i.d. over the persons tested k . The null hypothesis of no difference between the changes of the items is operationalized by the following:

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

where $c(x)$ is a value that only depends on x (i.e., all items’ posttest means are equal conditional on 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 of Equation 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. Further note that H_0 is formulated in terms of number-correct scores, whereas the discussed item response theory approaches operate on logit or probit scaled probabilities. We do not claim that it is superior in general to consider the number-correct scores but instead use the comparison of conditional expectations as a practical simplification of the comparison of the full conditional distributions. The comparison of conditional expectations is straightforward and easy to interpret. It would be conceivable to compare other functions 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 final section for more discussion of the scaling issue.

Using PRCS enables an asymptotically unbiased test of H_0 against the alternative that all items' conditional posttest means of property i are larger or equal to the other properties' means (Equation 10, again conditioned on the corresponding pretest values) and that there is at least one pretest value conditional on which a nonzero difference can be observed with a probability larger than 0 (Equation 11):

$$\mathbf{H}_1 : \forall x \in \{1, \dots, P\}, q \neq i, j \in \{1, \dots, J_i\}, r \in \{1, \dots, J_q\} : \tag{10}$$

$$E(X_{1ij1} | X_{0ij1} = x) \geq E(X_{1qr1} | X_{0qr1} = x),$$

$$\exists x \in \{1, \dots, P\}, q \neq i, \tag{11}$$

$$j \in \{1, \dots, J_i\}, r \in \{1, \dots, J_q\}, P\{X_{0ij1} = x, X_{0qr1} = x\} > 0 :$$

$$E(X_{1i0j1} | X_{0i0j1} = x) > E(X_{1qr1} | X_{0qr1} = x).$$

Equation 10 formulates a one-sided alternative. There is no difficulty in using the same methodology for a one-sided test against the opposite alternative, 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 in the computation of the PRCS.

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

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

$$D_{i \cdot k}(x) = X_{1i \cdot k}(x) - X_{1-i \cdot k}(x), \text{ where}$$

$$X_{1i \cdot k}(x) = \frac{\sum_{j: X_{0ijk} = x} X_{1ijk}}{N_{0i \cdot k}(x)},$$

$$X_{1-i \cdot k}(x) = \frac{\sum_{q \neq i, r: X_{0qrk} = x} X_{1qrk}}{N_{0-i \cdot k}(x)},$$

$N_{0i \cdot k}(x)$ 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 0, 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 subscript 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:

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

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 person tested. In this case, person k should be excluded from the analysis. The suggested weights are

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

as motivated by Lemma 1 below. Obviously, it makes sense to weight pretest values x up that 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 $\overline{D_{i \cdot k}}$, $k = 1, \dots, K$.

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 of Equations 14 through 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 distribution, it will match the distribution of the test statistic better in most situations. The one-sample Wilcoxon or the 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 that the t test has higher power.

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. Persons tested are only included in the comparison if the pretest values of the two items are 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 do not 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 do not 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:

$$\exists x \in \{1, \dots, P\}, q \neq i, j \in \{1, \dots, J_i\}, r \in \{1, \dots, J_q\} : \forall (x_1, x_2) \in \{1, \dots, P\}^2 : P\{X_{0ijk} = x, X_{0qrk} = x\} > 0, \tag{14}$$

$$P\{X_{1ijk} = x_1, X_{1qrk} = x_2\} < 1, \tag{15}$$

$$\forall x \in \{1, \dots, P\}, i \in \{1, \dots, I\}, j \in \{1, \dots, J_i\} : X_{1ijk} \text{ independent of } (X_{0qrk})_{qr} \text{ conditional under } X_{0ijk} = x. \tag{16}$$

Equation 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 in which all comparable posttest values are deterministic. In that case, statistical methods would not make sense. The only critical assumption is Equation 16, in which X_{1ijk} is allowed to depend on $(X_{0qrk})_{qr}$ (denoting the whole pretest result of person tested k) only through X_{0ijk} . Similar restrictions are implicit in the LPCM given in Equation 6 and, on the Likert-type mean score level, in the linear models of Equations 1 and 2. Moreover, the PRCS test does allow for arbitrary dependence structures among the pretest values as opposed to the models of Equations 1, 2, and 6.

Theorem 1. Assume the Equations 14 through 16. For $\overline{D_{i \cdot k}}$ as defined in Equation 12,

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

under H_m 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 given a single particular pretest value of x , then x would need the largest weight, but of course such information has to be obtained independently of the observed data if it is used in the definition of the weights.

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

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

The optimality result needs a further independence condition in addition to Equation 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):

$$\forall x \in \{1, \dots, P\}, i \in \{1, \dots, I\}, j \in \{1, \dots, J_i\} : X_{1ijk} \text{ independent of } (X_{1qrk})_{-i} \quad (19)$$

conditional under $X_{0ijk} = x$.

Keep in mind that Equations 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.

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-type 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 $\overline{D}_{i \cdot k}$ for all properties $i = 1, \dots, I$ may be inspected and can be interpreted directly in terms of the category values $i = 1, \dots, P$ 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.

The 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

In order to give a practical illustration of PRCS, excerpts of a study of Bargmann (1998) are presented. The study attempted to measure the emotional states of the subjects by a semantic differential before and after a music treatment consisting of the instrumental piece "Bait and Chase" from the *Alien III* motion picture soundtrack, among others. The semantic differential consisted of 50 self-referential statements that belonged to five emotional states (called properties generally in this article), 10 statements (items) for each state, with answers given on 5-point Likert-type scales. The states were joy, sadness, love, anger, and fear. The study design consisted of six different groups, but here only three of them are discussed, namely, "group E", with $K = 24$, which received the treatment (music listening) between the pretest and posttest rating of the semantic differential. "Group CD" ($K = 20$) worked in the same way with one exception. Whereas the instructions for group E were neutral concerning the measurement of the subjects' emotions, it was suggested to the subjects in group CD that this particular piece of music had evoked strong feelings of joy in a prior test session. In reality, it had generated feelings of fear. Actually, it was a main hypothesis of the study that the piece would evoke fear as an emotion. Consequently, the results given below focus on fear. "Group C1" with $K = 18$ received the pretest and had to complete a verbal task instead of the music treatment before the posttest. For more details, see Bargmann (1998). There is a considerable amount of literature on music-induced emotions that we cannot cover here. For literature overviews, see Juslin and Västfjäll (2008) and Scherer (2004).

In the experimental group E, the one-sided t tests for $\mu = 0$, with i being the fear score under the model of Equation 4, led to p values of .00055 (unrestricted) and smaller than .0001 assuming Equation 5. The means of the PRCS were 0.6207 (fear), -0.4882 (joy), -0.3450 (love), 0.1099 (sadness), and 0.2731 (anger). The t test for the fear mean to be equal to zero led to

Table 1 Conditional Distributions of All Posttest Values in the Group C1 Given the Pretest Values

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

Note: *n* = numbers of questions; % = pretest value-conditional percentages (sum to 100 along the rows), computed separately for fear and other; last two columns: overall pretest percentages (sum up to 100 along columns). The total *n* is 180 for fear; 717 for others; these are all questions answered by *K* = 18 persons tested. The last columns of the last two lines give overall posttest (“Conditional mean” column) and pretest means (“Pretest distribution” columns).

p < .0001. Not only was the change in fear clearly significant compared with the other changes, but the PRCS also had the largest absolute value. The interpretation of this value was that, given equal pretest values, the change in questions concerning fear was, on average (weighted depending on pretest values), 0.6207 larger than the average change in questions concerning the other properties. A referee suggested estimating the reliability of the PRCS by carrying out random splits of the items. This resulted in average correlations between test takers’ PRCS values computed on disjunct halves of the items ranging, for the five properties, from 0.342 to 0.621.

Interestingly, an application of the same tests (*t* tests based on Equation 4, PRCS, respectively) to the group CD did not give significant results, either for fear or for any other property, with PRCS means of 0.0645 (fear; *p* = .2894), 0.2547 (joy), 0.1825 (love), -0.1488 (sadness), and -0.0324 (anger).

Group C1 illustrates how PRCS take into account the pretest distribution when measuring change. A two-sided *t* test was used here, because there was no reason to expect in advance that changes in fear would have a particular direction. The means of the PRCS were -0.1545 (fear), 0.8328 (joy), 0.1643 (love), -0.3065 (sadness), and 0.1102 (anger). The *t* test for the fear mean hypothesis equal to zero led to *p* = .0174, so at the 5% level the change was significantly different from the aggregated change in the other properties. The regression *t* test for fear was no longer significant with *p* = .0779 (unrestricted; assuming Equation 5 yielded *p* = .0099). It can be seen from Table 1 that the overall mean value for the questions related to fear changed from 1.77 (pretest) to 1.66 (posttest) and the overall mean for all other questions changed from 2.15 to 2.08. The two mean changes differed by 0.04 and were therefore very similar. Note, however, that because the pretest mean value for fear was smaller than for the aggregated other properties, there was less space for the posttest value to become even smaller (this situation is somewhat opposite to the one in Example 1; the repeated measures ANOVA and

item response theory approaches would have difficulties finding significant changes here). Comparing the conditional means for the five pretest values separately reveals that all of the five differences between them were smaller than -0.04 . The significant PRCS of -0.1545 is a weighted average of the pretest value-wise differences between posttest values, preventing the different pretest distributions from obscuring the change.

Simulations

A simulation study was carried out to compare the performance of some of the proposed tests. Five tests were applied:

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

RegrRestrict. The t test for $\mu = 0$ in the multivariate regression of Equation 4 assuming Equation 5.

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

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

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

All simulations were carried out using the parameters $K = 20$, $P = 5$, $I = 5$, $J_i = 10$, $i = 1, \dots, 5$. Property 1 was the property of interest (i.e., a situation similar to the *Alien III* data). Emphasis was on the effect of different pretest distributions between Property 1 and the other properties. The simulations were based on three different setups under the null hypothesis and three different setups under the alternative:

standard. Uniform distribution on $1, \dots, 5$ for all pretest values. Each posttest value was equal to the corresponding pretest value with probability 0.4, all other posttest values were chosen with probability 0.15 (H_0).

lowPre1. The pretest values for Property 1 were 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 were 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 were chosen as in case standard (H_0).

lowPre1highPost. The pretest values were generated as in case lowPre1, the posttest values were 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 was distributed as follows: The two highest remaining values were chosen with probability 0.2, and the two lower values were chosen with probability 0.1 (H_0).

highPost1. The pretest values and the posttest values for the Properties 2 through 5 were generated as in case standard (no differences in the pretest distribution), the posttest values for Property 1 were generated as in case lowPre1highPost (H_1).

Table 2 Simulated Probability of Rejection of H_0 From 1,000 Simulation Runs

	Regression	RegrRestrict	RCS-t	RCSWilcoxon	RCSsign
Standard	0.050	0.049	0.055	0.050	0.033
LowPreI	0.044	0.037	0.053	0.050	0.038
LowPreIhighPost	0.039	0.036	0.049	0.053	0.043
HighPostI	0.517	0.574	0.516	0.491	0.340
LowPreIhighPostI	0.107	0.115	0.468	0.444	0.301
HighPreIhighPostI	0.115	0.142	0.483	0.465	0.318

lowPreIhighPostI. The pretest values were generated as in case lowPreI, the posttest values were generated as in case highPostI (H_1).

highPreIhighPostI. Same case lowPreIhighPostI 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) (H_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 described by Equations 1, 6, and 8. The sign test always appears conservative, and the regression methods are conservative for lowPreIhighPost. 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 higher power under Equation 5 than when it is unrestricted in all cases.

Discussion

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-type mean scores, whereas 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, whereas 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, and the posttest data have to allow for arithmetic operations such as computing differences and sums. Whether or not the computation of means for 5-point Likert-type scales is meaningful (the regression/ANOVA methods operate to an even stronger extent on the interval scale level) is debatable. It seems 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 exception 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 article, 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, although 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, it was demonstrated that the effect parameters for other models may be misleading in certain situations.

For all methods, a significant difference in changes for anxiety and 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 0 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-type 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). Although this depends on the items in general (and it may be worthwhile to analyze the items with respect to this problem), the assumption is acceptable in a setup in which the categories for the answers are identical for all items and are presented to the persons tested in a unified manner, because the visual impression of the questionnaire suggests such an interpretation to the persons tested.

Item response theory as presented in Samejima (1969), Andrich (1978), Muraki (1990), and Fischer and Ponocny (1994) addresses such comparisons by category by model assumptions that formalize them as difficulties corresponding to several abilities. Although it could be interesting to apply such approaches to the present data, it seems 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 persons tested 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, there does not seem to be a clear relationship between the distribution of test persons' 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 compared the activity in different regions of the human brain as related to the induction of certain emotions.

An important difference between the approaches discussed here are the scales on which the methods operate. The linear models and the PRCS operate on the number-correct scale, whereas the item response theory approaches operate on logit or probit transformed probabilities. It is not necessarily the case that one of these scales is generally better than the other, but the 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 does not depend on any scaling whether two distributions are equal or not.

The scale issue arises for the PRCS on two levels. First, the distributions to be compared are conditional on the number-correct 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 number-correct 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 number-correct 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, $\text{Var}(\overline{D_{i \cdot k}}) < \infty$ and $\overline{D_{i \cdot k}}, k \in IN_K$ i.i.d. S_K^2 converges almost surely to $\text{Var}(\overline{D_{i \cdot k}}) > 0$ 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 $\tilde{x} \in \{1, \dots, P\}^{J_1 + \dots + J_I}$ 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 Equation 16,

$$a(x, \tilde{x}) = E(D_{i \cdot 1}(x) | (X_{0qr1})_{qr} = \tilde{x}) = \frac{1}{n_i(x, \tilde{x})} \sum_{j: X_{0ij1} = x} E(X_{1ij1} | X_{0ij1} = x) - \frac{1}{n_{-i}(x, \tilde{x})} \sum_{(q,r) X_{0qr1} = x} E(X_{1qr1} | X_{0qr1} = x)$$

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

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

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, \dots, Y_K are independent random variables with equal mean c and variances $V_r, r = 1, \dots, K$, 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 of Equation 18 regardless of x and \tilde{x} unless $w_{x, \tilde{x}} = 0$. Therefore, $E(\overline{D_{i \cdot 1}} | (X_{0qr1})_{qr} = \tilde{x}) = c$ by Equation 20. From Equations 18 and 19,

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

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. Because the conditional expectation of $\overline{D_{i.1}}$ is independent of \tilde{x} , separate minimization of the conditional variances for all \tilde{x} also minimizes the unconditional variance of $\overline{D_{i.1}}$.

Declaration of Conflicting Interests

The authors had no conflicts of interest with respect to the authorship or the publication of this article.

Funding

The authors received no financial support for the research and/or authorship of this article.

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., & 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* [Quantified emotions: A procedure for measuring emotions induced by music]. Unpublished master's thesis, Universität Hamburg.
- Bonate, P. L. (2000). *Analysis of pretest-posttest designs*. Boca Raton, FL: Chapman & Hall.
- Cressie, N. (1980). Relaxing assumptions in the one-sample *t* test. *Australian Journal of Statistics*, *22*, 143-153.
- Cribbie, R. A., & Jamieson, J. (2004). Decreases in posttest variance and the measurement of change. *Methods of Psychological Research Online*, *9*, 37-55.
- Cronbach, L. J., & Furby, L. (1970). How should we measure change—or should we? *Psychological Bulletin*, *74*, 68-80.
- Dimitrov, D. M., & Rumrill, P. D. (2002). Pretest-posttest designs and measurement of change. *Work*, *20*, 159-165.
- Eid, M., & 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 D. N. M. DeGruijter & L. J. T. Van der Kamp (Eds.), *Advances in psychological and educational measurement* (pp. 97-110). New York: John Wiley.
- Fischer, G. H. (2003). The precision of gain scores under an item response theory perspective: A comparison of asymptotic and exact conditional inference about change. *Applied Psychological Measurement*, *27*, 3-26.
- Fischer, G. H., & Ponocny, I. (1994). An extension of the partial credit model with an application to the measurement of change. *Psychometrika*, *59*, 177-192.
- Jaccard, J., & Wan, C. K. (1996). *LISREL approaches to interaction effects in multiple regression*. Thousand Oaks, CA: Sage.
- Jamieson, J. (1995). Measurement of change and the law of initial values: A computer simulation study. *Educational and Psychological Measurement*, *55*, 38-46.
- Juslin, P. N., & Västfjäll, D. (2008). Emotional responses to music: The need to consider underlying mechanisms. *Behavioral and Brain Sciences*, *31*, 559-621.
- 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., & Fox, P. T. (2000). Differential limbic-cortical correlates of sadness and anxiety in healthy subjects: Implications for affective disorders. *Biological Psychiatry, 48*, 30-42.
- Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149-174.
- 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. (2004). Which emotions can be induced by music? What are the underlying mechanisms? And how can we measure them? *Journal of New Music Research, 33*, 239-251.
- Steyer, R., Eid, M., & Schwenkmezger, P. (1997). Modeling true intraindividual change: True change as a latent variable. *Methods of Psychological Research Online, 2*, 21-33.
- Wang, W.-C., & Chyi-In, W. (2004). Gain score in item response theory as an effect size measure. *Educational and Psychological Measurement, 64*, 758-780.