# Comparison of changes in a pretest-posttest design with Likert scales



**Author(s):**
Hennig, Christian; Müllensiefen, Daniel; Bargmann, Jens

**Publication date:**
2003

**Permanent link:**
https://doi.org/10.3929/ethz-a-004543804



**Originally published in:**
Research Report / Seminar für Statistik, Eidgenössische Technische Hochschule (ETH) 118

# Comparison of changes in a pretest-posttest design with Likert scales

by

Christian Hennig, [1] Daniel Müllensiefen [2] and Jens Bargmann [3]

Research Report No. 118
June 2003

Seminar für Statistik

Eidgenössische Technische Hochschule (ETH)

CH-8092 Zürich

Switzerland

[1]ETH Zürich, Seminar für Statistik and Fachbereich Mathematik-SPST der Universität Hamburg
[2]Musikwissenschaftliches Institut der Universität Hamburg
[3]Espotting Media GmbH, Hamburg

# Comparison of changes in a pretest-posttest design with Likert scales

Christian Hennig, [§] Daniel Müllensiefen [¶] and Jens Bargmann [‖]

Seminar für Statistik
ETH Zentrum
CH-8092 Zürich, Switzerland

June 2003

**Abstract**

Two methods for comparison of the influence of a treatment on different discrete variables are suggested, compared and applied to a dataset, which concerns the influence of music on emotions and stems from a questionnaire, where five emotions have been measured by ten questions for each emotion on a five-point Likert scale. The question has been if a certain piece of music induces anxiety to a significantly higher extent than other emotions. The pretest values for the emotions cannot be expected to be equally distributed, and two methods to take this into account are proposed. The first one is a linear regression on the Likert mean scores with pretest values as independent variables. The second one is a t-test on new change scores, which are derived conditional on the pretest values. It is shown that the second approach is more appropriate in the present setup.

**Keywords:** linear regression, poststratification, relative change scores, music and emotions

## 1 Introduction

In the present article, the analysis of data of the following form is addressed: $l$ properties of $n$ test persons are measured by $m_i$, $i = 1, \ldots, l$, items (usually the $m_i$ are the same for all properties) 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.

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

[§] ETH Zürich, Seminar für Statistik and Fachbereich Mathematik-SPST der Universität Hamburg
[¶] Musikwissenschaftliches Institut der Universität Hamburg
[‖] Espotting Media GmbH, Hamburg

A typical example is data from questionnaires where the measurement of different properties of the test persons is operationalized by asking $m_i$ 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". The 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 $m_i$ items is taken as a score for each property (in the literature, often the sum is taken, but the mean allows varying values of $m_i$). The new techniques are applied to data stemming from such a questionnaire, which consists of $m = m_i = 10$ times $i = 1, \ldots, l = 5$ questions on a $p = 5$-point scale as above corresponding to the emotions joy, sadness, love, anger and anxiety. The aim was to find out if a piece of music from the movie soundtrack of "Alien III" affects anxiety significantly more than the other emotions. The study is introduced in Section 2.

For pretest-posttest data like these, it is dangerous to base the inference about the changes on the differences between posttest scores and pretest scores, because these differences depend on the pretest score. For example, the difference cannot be positive if the pretest score of an item has already been maximal. Since different properties have to be compared, there is no reason to expect that the distribution of the pretest scores will be the same for all items or properties. Compare also Figure 1 of Section 5, where about the same number of the items denoted by "F" increase and decrease from pretest to posttest, but this is due to the fact that most of these items have a pretest score of 1 and the general tendency is clearly negative. Situations where a property yields larger pretest scores than the others, which causes the positive changes to be smaller for that property, can lead to the occurrence of the broadly discussed Simpson's paradox (see Samuels, 1993, and the references given therein) if the changes are compared without taking the pretest scores adequately into account.

A similar phenomenon is known also for continuous data with an unbounded value range. It is known under the term "regression towards the mean" (see Bonate, 2000, Chapter 2, and the references given therein). It means that even if pretest and posttest scores are modelled as the same "true" value plus independent errors, the observed difference between posttest and pretest value will be smaller in broad tendency if the pretest value had been large and the other way round.

A reasonable strategy to deal with regression towards the mean is analysis of covariance, where the difference between pretest and posttest scores is modeled as dependent variable, and the independent variables are the treatment factor and the pretest score (Bonate, 2000, Chapter 5). Since the changes between different emotions on the same person are to be compared, the appropriate analogue is a multivariate regression, where the pretest scores of the different emotions are the independent variables and the dependent variables are the score differences of the emotions. Inference is made about the difference between the intercepts of the emotions. Such an analysis can be carried out on the Likert mean scores

$$L_{hik} = \frac{1}{m_i} \sum_{j=1}^{m_i} X_{hijk} \text{ and } L_{h-ik} = \frac{1}{\sum_{q \neq i} m_q} \sum_{q \neq i} \sum_{r=1}^{m_q} X_{hqrk}.$$

The distribution of these scores is often not too far from the normal. The regression analysis is introduced in Section 3. It is somewhat ad hoc insofar as it ignores the way the Likert mean scores are obtained and treats them as usual continuous data.

In Section 4, a method is proposed, which is more directly tailored to the specific kind of data. Since the pretest scores for the items have only few possible values, the idea of poststratification as suggested by Bajorski and Petkau (1999) may be applied as well. 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, assuming that the distribution of the pretest scores is equal in the two groups, which will usually not hold for different emotions of the same group. Therefore, a new relative change score is defined, which is based on a separate poststratification of the items of every single test person. The relative change score aggregates the differences between the posttest scores of the items corresponding to the emotion of interest and the mean posttest score for all other items with the same pretest score. The relative change score makes explicit use of the fact that an emotion is measured by adding the results from $m$ items with $p$ ordered categories instead of analyzing the Likert mean scores.

The strategies are applied to the Alien-dataset in Section 5. It turns out that they lead to different results in some experiments. By means of a graphical data analysis it is shown that the dependence between the posttest-pretest differences and the pretest scores is stronger than would be expected for continuous variables as an effect of regression towards the mean because of the nature of the computation of the emotional scores from $m$ $p$-point scaled items. Thus, the multivariate regression method does not fully account for this dependence, and the results of the relative change score method are more reliable.

The superiority of the relative change score method is further illustrated by some small simulations in Section 6. The regression method may be applied in situations where the scores are not aggregates of as much as ten five-point scaled items per subject and property, which make the relative change score method feasible.

The paper is concluded by some discussion.

# 2 Effects of music on emotions: the Alien data

## 2.1 Music and emotion

It is a widespread conviction that music bears a close relationship to human emotions. Discussing all the emotional functions that music may have, the German musicologist Georg Knepler termed music "the language of emotions". In his opinion - which is shared by many others - music is an acoustic system of communication that can convey the meaning of inner and emotional states. In this respect music surpasses ordinary language as a means of communication for emotional conditions (Knepler, 1982, p.37). With this idea Knepler follows Kant who clearly 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 Zentner and Scherer, 1998). Difficulties arise in this research area from the lack of a unified theoretical framework for music and emotions and from problems with the measurement of emotions or emotional changes caused by music listening (Pekrun, 1985; Harrer, 1993; McMullen, 1996; Müllensiefen, 1999). 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 large 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 on a questionaire. Physiological and gestural measurements have not been taken into account.

## 2.2 Measurement instrument

The original study (Bargmann, 1998) 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 its respective emotional states (categories) were selected according to the results of an extensive pretest. In this pretest a group of 26 subjects listed emotional categories that they believed to be important with music listening and enumerated adjectives that best described these categories. This method of having subjects from a similar population define their emotional categories with music and the corresponding verbal expressions (adjectives) minimizes the possibility that the semantic differential is not apt for the intended task. Since the early days of the semantic differential as measurement tool in psychology, it is well known that the items should come from the language that the tested population usually employs for the area under study (Micko, 1962). Otherwise, items that do not seem appropriate to the subjects tend to be rated in middle categories by the subjects (Mikula and Schulter, 1970). The results of this pretest indicated five emotional categories (called "properties" in the statistical part): joy, sadness, love, anger, and anxiety/fear. These categories fit nicely with the most common and basic categories for emotional music experiences by Marx (1982) and Rösing (1993). 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.

A second pretest was conducted to find music examples that could serve as effective treatment to induce emotional changes in the subjects through listening. Eight subjects listened to 14 pieces of music from rock to classical music that were likely to represent all five emotional categories. Subjects judged the quality, intensity, unambiguity, and homogeneity of the music examples on quantitative rating scales and gave qualitative explanations for their ratings on a questionnaire. The piece that evoked strongest and most homogeneous emotions was the instrumental piece "Bait and Chase" from the motion picture soundtrack "Alien III". 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.

## 2.3 Design and sample

The subjects 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 $n = 24$ 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 $n = 20$ 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 (see Mecklenbräuker and Hager, 1986, for details).

The first control group C1 with $n = 18$ 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 so-called "Solomon four group design" (Solomon 1949), there have been three

other control groups without pretest to control the effect of pretest sensibilization (Bortz and Döring 1995, p. 502f). The statistical evaluation of these groups was by means of standard methodology and is not further discussed here.

### 2.4 Rationale

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.

### 2.5 Qualitative validation

To answer the question if and how emotional changes due to music listening could be measured adequately, a validation of the experimental results was necessary. To the authors' knowledge, the original study was the first instance of an experimental test procedure which used the Solomon design and the posttest-pretest methodology with musical stimuli. Thus, neither the outcome of significant results in accordance with the hypotheses nor the opposite could be taken as an approval for the methods employed. Therefore, as a very different methodology, subsequent qualitative interviews were run to get a confirmation from a different angle that the employed music example actually evokes anxiety/fear as hypothesized. Only in case both methods - the quantitative experimental results and the qualitative information from the interviews - yield the same results, it could be assumed that the music had the foreseen effect and the measurement with the Solomon design was correct. Interviews with six subjects ranging in age from 19 to 31 and with different music backgrounds were conducted.

## 3 Linear regression approach

A linear regression analysis of the data can be based on the Likert mean scores. Suppose that property no. $i$ is the property of interest. The changes, i.e., the differences between posttest and pretest scores $C_{ik} = L_{1ik} - L_{0ik}$, $C_{-ik} = L_{1-ik} - L_{0-ik}$, are the dependent variables and the centered pretest scores are the independent variables:

$$\begin{pmatrix} C_{ik} \\ C_{-ik} \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} + \begin{pmatrix} \beta_{11} & \beta_{12} \\ \beta_{21} & \beta_{22} \end{pmatrix} \begin{pmatrix} L_{0ik} - \bar{L} \\ L_{0-ik} - \bar{L} \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \end{pmatrix}, \tag{3.1}$$

$\bar{L} = (\sum_{k=1}^{n}(L_{0ik} + L_{0-ik}))/(2n)$ being the overall pretest score mean. $\mu_1$ and $\mu_2$ 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 and accounts for "regression towards the mean", see Bonate (2000, chapter 5). $\epsilon_1$ and $\epsilon_2$ are error variables with zero mean independent of $L_{0ik}$ and $L_{0-ik}$. The null hypothesis of interest is the equality of the treatment effects for $C_{ik}$ and $C_{-ik}$, 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

$$C_{ik} - C_{-ik} = \mu + \beta_1(L_{0ik} - \bar{L}) + \beta_2(L_{0-ik} - \bar{L}) + \epsilon. \tag{3.2}$$

For the sake of a proper interpretation, it is favourable to assume

$$\beta_{11} = \beta_{22},\ \beta_{12} = \beta_{21} = 0,\ \text{thus } \beta_2 = -\beta_1 \text{ in (3.2)}. \tag{3.3}$$

This means that the difference between $C_{ik}$ and $C_{-ik}$ apart from the random error can be explained by $\mu_1 - \mu_2$ and the difference between $L_{0ik}$ and $L_{0-ik}$ alone, while otherwise the difference depends on the size of $L_{0ik}$ and $L_{0-ik}$ even if they are equal.

It may be doubtful if assumption (3.3) is justified in practice, but it will be demonstrated in the sections 5 and 6 that it improves the quality of the results in the present setup. For the datasets treated in the present paper, standard $t$-tests did not reject (3.3) in favour of the unrestricted model, which seems to be a consequence of a high correlation between $L_{0ik}$ and $L_{0-ik}$.

It will turn out in the section 6 that the linear regression approach is not very good under some non-identical distributions of $L_{0ik}$ and $L_{0-ik}$. 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 score differences. Departures from independence of the errors could not be observed and departures from normality do not seem dangerous for the data. In the next section a methodology is presented that takes the individual items into account.

## 4 Relative change scores

The idea of the relative change scores is the aggregation of measures for the relative change of the item scores belonging to the property of interest compared with the other properties conditional on their posttest values.

The relative change score 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:

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

   where $N_{0i\cdot k}(x)$ is the number of items of property $i$ with pretest value $x$, and $N_{0-i\cdot k}(x)$ is 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.

2. The relative change score 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:

$$\overline{D_{i\cdot k}} := \frac{\sum\limits_{x=1}^{p} w(N_{0i\cdot k}(x), N_{0-i\cdot k}(x)) D_{i\cdot k}(x)}{\sum\limits_{x=1}^{p} w(N_{0i\cdot k}(x), N_{0-i\cdot k}(x))}. \tag{4.1}$$

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

$$w(n_1, n_2) := \frac{n_1 n_2}{n_1 + n_2}, \tag{4.2}$$

see Lemma 4.2 below. 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 $\overline{D_{i\cdot k}}, k = 1, \ldots, n$. The null hypothesis to be tested is $E\overline{D_{i\cdot k}} = 0$ with a one-sample $t$-test under the assumption that the test persons behave i.i.d. The simulations in section 6 indicate that a $t$-test may have a better power than the corresponding non-parametric Wilcoxon- and sign-tests. In general, the $t$-test can be expected to outperform the nonparametric tests under situations where the value range is bounded and the values are not too concentrated far from the bounds, because in such situations outliers cannot occur. However, this should be inspected graphically. Under the same circumstances, the $t$-test should also be preferable to the asymptotically equivalent normal test suggested by the central limit theorem, see Cressie (1980).

For exploratory purposes, the mean values of the $\overline{D_{i\cdot k}}$ may be considered for all properties $i = 1, \ldots, l$, and the relative change scores may also be used to test the equality of changes in property $i$ between different groups with a two-sample $t$-test.

The following theory justifies the operationalization of "equal changes between property $i$ and the others" as $E\overline{D_{i\cdot k}} = 0$. It is shown that the proposed test is (asymptotically) unbiased for the hypothesis

$$\begin{aligned} \mathbf{H_0}: \quad & \forall x \in \{1, \ldots, p\},\ q \neq i,\ j = 1, \ldots m_i,\ r = 1, \ldots, m_q: \\ & E(X_{1ij1} | X_{0ij1} = x) = E(X_{1qr1} | X_{0qr1} = x) \end{aligned} \tag{4.3}$$

(all item's posttest means are equal conditional under all pretest values) against the alternative that all item's conditional posttest means of property $i$ are larger or equal than the other's properties means and there is at least one pretest value conditional under which a nonzero difference can be observed with probability larger than 0:

$$\begin{aligned} \mathbf{H_1}: \forall x \in \{1, \ldots, p\},\ q \neq i,\ j \in \{1, \ldots, m_i\},\ r \in \{1, \ldots, m_q\}: \\ E(X_{1ij1} | X_{0ij1} = x) \geq E(X_{1qr1} | X_{0qr1} = x), \\ \exists x \in \{1, \ldots, p\},\ q \neq i, \\ j \in \{1, \ldots, m_i\},\ r \in \{1, \ldots, m_q\},\ P\{X_{0ij1} = x,\ X_{0qr1} = x\} > 0: \\ E(X_{1i_0j1} | X_{0i_0j1} = x) > E(X_{1qr1} | X_{0qr1} = x). \end{aligned} \tag{4.4}$$

The following assumptions are needed:

$$\exists x \in \{1, \ldots, p\},\ q \neq i,\ j \in \{1, \ldots, m_i\},\ r \in \{1, \ldots, m_q\}:\ \forall (x_1, x_2) \in \{1, \ldots, p\}^2:$$

$$P\{X_{0ijk} = x,\ X_{0qrk} = x\} > 0, \tag{4.5}$$

$$P\{X_{1ijk} = x_1,\ X_{1qrk} = x_2\} < 1, \tag{4.6}$$

$$\begin{aligned} \forall x \in \{1, \ldots, p\},\ i \in \{1 \ldots, l\},\ j \in \{1, \ldots, m_i\}:\ & X_{1ijk} \text{ independent of } (X_{0qrk})_{qr} \\ & \text{conditional under } X_{0ijk} = x. \end{aligned} \tag{4.7}$$

Assumption (4.5) 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 $X_{0ijk} = x$. (4.6)

excludes the case that all comparable posttest values are deterministic. In that case, statistical methods would not make sense. The technical assumption (4.7) means that $X_{1ijk}$ has to depend on $(X_{0qrk})_{qr}$ (denoting the whole pretest result of test person $k$) only through $X_{0ijk}$. This seems to be a strong restriction, but on the other hand the theory allows an arbitrary dependency structure among the pretest values $(X_{0qrk})_{qr}$.

**Theorem 4.1** *Assume (4.5)-(4.7). For $\overline{D_{i\cdot k}}$ as defined in (4.1),*

$$\left(\frac{\sum_{k=1}^{n}\overline{D_{i\cdot k}}}{(nS_n^2)^{1/2}}\right) \text{ converges in distribution to } \mathcal{N}(a_j,1),\ j=0,1, \tag{4.8}$$

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

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.

As a reference an alternative model is assumed where all items and pretest values behave in the same manner conditional under the pretest value:

$$\begin{gathered}
\forall x\in\{1,\dots,p\},\ j_1,j_2\in\{1,\dots,m_i\}:\\
E(X_{1ij_1k}|X_{0ij_1k}=x)=E(X_{1ij_2k}|X_{0ij_2k}=x)=:E_{i,x},\\
\exists c>0:\ \forall x\in\{1,\dots,p\},\ q\neq i,r\in\{1,\dots,m_q\}:\\
E(X_{1qrk}|X_{0qrk}=x)=E_{i,x}+c,\\
\forall x_1,x_2\in\{1,\dots,p\},\ q\neq i,\ j\in\{1,\dots,m_i\},\ r\in\{1,\dots,m_q\}:\\
\mathrm{Var}(X_{1ijk}|X_{0ijk}=x_1)=\mathrm{Var}(X_{1qrk}|X_{0qrk}=x_2)=:V.
\end{gathered} \tag{4.9}$$

**Lemma 4.2** *For $\overline{D_{i\cdot k}}$ as defined in (4.1) and $H_1$ fulfilling (4.9), $a_1$ from Theorem 4.1 is maximized by the weight function $w$ given in (4.2).*

The proof is given in the appendix.

Both the relative change scores 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.

## 5 Results for the Alien data

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 leads to $p$-values of 0.00055 (unrestricted) and $6.5e-5$ under (3.3). All tests are one-sided, unless indicated explicitly. The means of the relative change scores 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 $p=1.6e-6$. Not only is the change in anxiety clearly significant compared with the other changes, but the relative change score has also the largest absolute value.

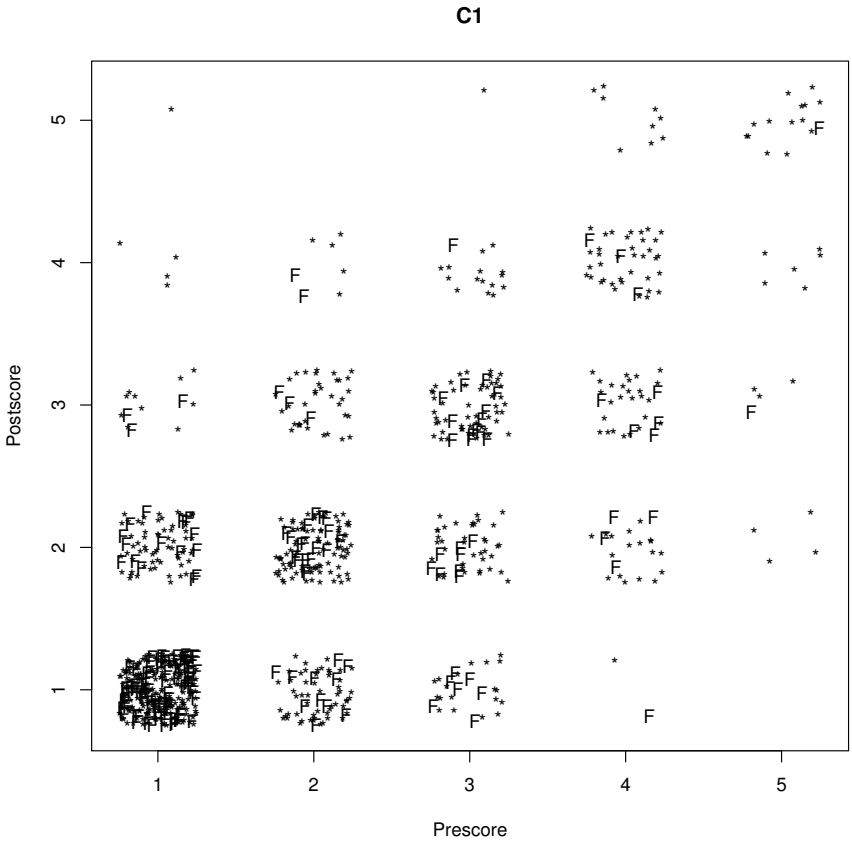


Figure 1: Pretest and posttest values of all items of all test persons of group C1. "F" indicates items belonging to anxiety/fear.

In the experimental group E (Counter Demand), the regression $t$-test $p$-values for anxiety are 0.223 (unrestricted) and 0.265 under (3.3). The means of the relative change scores 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. The regression $t$-test for anxiety are leads to $p = 0.0779$ (unrestricted) and $p = 0.0099$ under (3.3). The means of the relative change scores 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 relative change scores 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 relative change scores (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.

The unrestricted regression test does not lead to a significant result here and it may be wondered why the different tests lead to different conclusions and which result is most reliable.

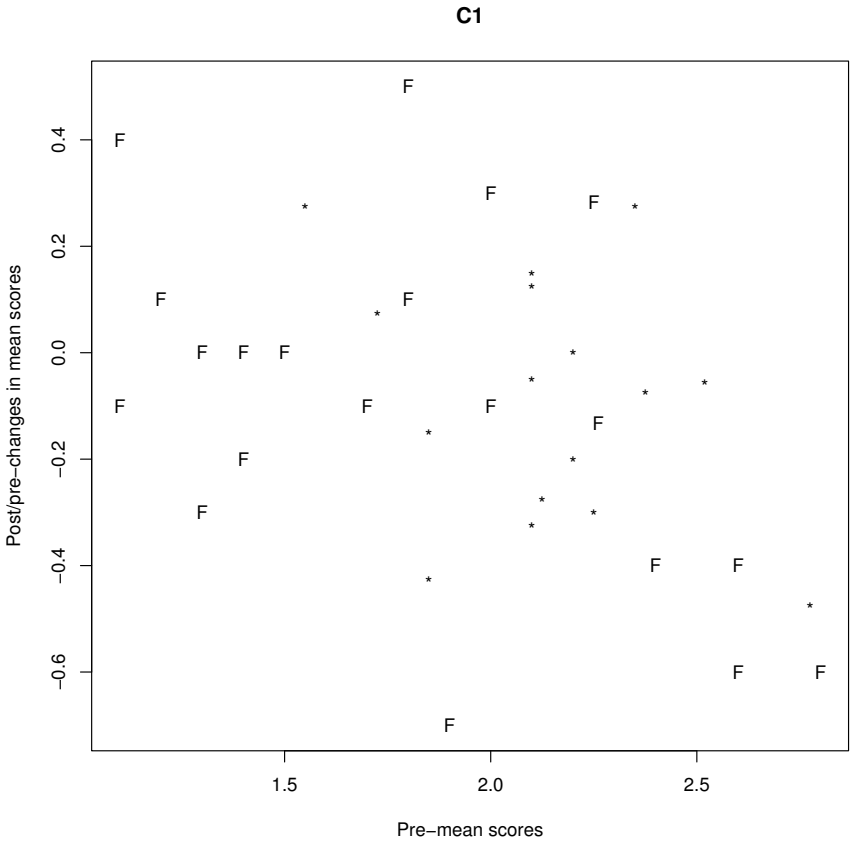


Figure 2: Pretest Likert mean scores vs. difference between post- and pretest Likert mean scores of all test persons of group C1. "F" indicates items belonging to anxiety/fear, so that there is one "F" and one "*" for each test person.

Figure 1 shows the pretest and posttest values of all items and all test persons, items belonging to anxiety/fear indicated by "F". The points are "jittered" around the true integer values to improve the clarity of the plot. It can clearly be seen that for all pretest values there is a tendency for the anxiety/fear items to produce lower posttest values. Insofar, the result of the relative change score test seems to be reliable. Note, however, that this plot does not allow to separate variation between the test persons from variation between the items of the same test person. In Figure 2, the pretest Likert mean scores, to which the regression methods are applied, are plotted vs. the difference between the posttest and pretest Likert mean scores. The unrestricted regression models the influence of the pretest scores with different slopes for anxiety/fear and the other properties. This does not lead to a significantly lower intercept estimate for anxiety. The problem here is that many of the pretest scores for anxiety are so small that there are no pretest scores for the other properties with which the anxiety scores can be compared. That anxiety leads to lower posttest scores for the same pretest scores cannot clearly be detected by use of the Likert mean scores on which the regression method is based. The restricted regression assumes the slopes for the dependency of the changes on the pretest scores as equal. This is not easy to verify. The significant result means that if the regression lines would be parallel (which may be doubted), the anxiety intercept would be below the one of the other properties. This is consistent with the result from the relative change score test,

but the evidence leading to the regression result seems to be weaker. The advantage of the relative change score method is clearly visible in this example: even if the pretest Likert mean scores of the property of interest are so low that there are no scores of the other property with which they can be compared, there may be enough items of the other properties with minimal value, so that a comparison based on the item values is more reliable than with the aggregated Likert scores.

# 6 Simulations

A small simulation study has been carried out to compare the performance of the proposed tests. Five tests have been applied:

**Regression** The $t$-test for $\mu = 0$ in (3.2) with unrestricted regression parameters.

**RegrRestrict** The $t$-test for $\mu = 0$ in (3.2) under the restriction (3.3).

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

**RCSWilcoxon** The one-sample Wilcoxon test for symmetry of the distribution of the relative change scores about 0.

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

All simulations have been carried out with $n = 20$, $p = 5$, $l = 5$, $m_q = 10$, $q = 1, \ldots, 5$, and property 1 has been the property of interest, i.e., a situation similar to the Alien data. 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 values 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 $1, 2, 3, 4, 5$. 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. Else 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`, 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.

| | 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 1: Simulated probability of rejection of $H_0$ from 1000 simulation runs. The nominal level has been 0.05.

The results of the simulation are shown in Table 1. The results for the $H_0$-cases do not indicate any clear violation of the nominal level. 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 relative change score methods. The two nonparametric tests based on relative changes scores perform worse than the $t$-test. The linear regression test shows a better power under the restriction (3.3) than unrestricted in all cases.

# 7 Discussion

Two classes of methods for comparing the changes between different properties measured on Likert scales between pretest and posttest have been proposed. The linear regression tests use the Likert mean scores while the relative change score tests are directly based on the items. The advantage of the relative change scores 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 relative change score approach needs a sufficient number of items, compared with the number of categories for the answers. If only single score values for pretest and posttest exist, the regression approach has to be chosen. Relative change scores 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), the posttest data has to allow for arithmetic operations such as computing differences and sums.

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 relative change score of anxiety is not only significantly different from 0, but also the largest one in absolute value.

Data from five-point Likert scales are not generally recognized to be of interval scale quality, but it is common practice in the social sciences to apply methods for interval scales to them. The application of such methods to ordinal data is often reasonable and robust (Jaccard and Wan, 1996), and from a statistical point of view, the relevant assumptions on statistical methods are about distributional shapes and independence, but not about scale types (Velleman and Wilkinson, 1993). Furthermore, the item values on the five-point scales are a kind of ranks, and computing sums, means and differences of ranks is crucial for some of the most common methods for ordinal data (Wilcoxon tests, Spearman correlation). There is a difference to the ordinary mathematical definition of ranks: For ordinary rank based methods, the

effective difference between two categories is determined by the number of subjects choosing the categories. For example, if there is one subject in category 1, one in category 3, one in category 4 and none in category 2, the effective difference between the categories 1 and 3 is equal to the difference between categories 3 and 4 (because the corresponding subject's ranks are 1, 2 and 3), while the former effective difference is twice the latter for all analyses based on the Likert scores (be it items, sums or means). It must be left to the interpretation of the measurement if the number of categories is more meaningful with respect to the aim of a study than the distribution of the subject's choices.

A more serious concern may be raised about the meaning of a comparison of measurement values for different variables (properties and items). The analysis presented here assumes 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. 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.

The quantification of emotion is a controversial task and we do not advocate Bargmann's (1998) approach as the definitive solution of this problem. From a statistical point of view, the measurements can be interpreted as "operational" in the sense of Hand (1996), which means roughly that "our definition of emotional change is what is measured by our instrument."

However, our concept of measureable emotions is based on the communicable subjective self-attribution of the individuals as mirrowed by the questionnaire ratings. It was confirmed by the results of the six qualitative interviews that the changes of these questionaire ratings were really caused by the induction of anxiety in the subjects through the music treatment. Half of the subjects reported that they actually felt anxiety and fear while listening to the Alien III example. These subjects described their emotional and physiological reactions for example as "negative tension", "horrifying elements", "feelings of panic" or "fear, that made me tense up". When asked about their associations with the piece of music, all of the subjects indicated terms that belong to semantic field of anxiety or horror movies, like "1000 liters of blood", "a haunted castle", "a man threatening with a knife" etc. All of the subjects declared the music as unpleasant and that they did not like the example.

So obviously all of the interviewed subjects perceived the anxiety-character of the music example, but only half of them actually experienced the corresponding emotions as their own inner states. From the explanations the subjects gave about their emotional reactions afterwards, it was concluded that the younger subjects and the subjects that had less active experiences with music showed a defence reaction to the extreme music example that they rejected aesthetically. As some of them reported, a strong feeling of rejection to the music came up first and this feeling prevented other and more specific emotional reactions. The musically more experienced subjects felt a strong dislike as well but were able to relate emotionally to the character of the music. As some of them reported, they even enjoyed aesthetically somehow the feelings of anxiety the music provoked (see Schubert (1996) for the same phenomenon).

The sketched complex interplay of the factors of personal preferences, aesthetic judgements, the possible defence mechanism and the induction of emotions may possibly explain the inhomogeneous reactions to the variety of music examples in one of the pretests. However, even the music example that induced the strongest and most homogeneous emotions in the pretest - the Alien III example - does not allow for a straight and simple stimulus-response relation, as was evidenced by the interviews.

In sum, the study reported here is an example of the meaningful and complementary interplay of quantitative and qualitative research methods. The proposed method for the treatment

of the change scores on the Likert scales made statistical testing of the hypotheses possible: emotions can be induced by music and the effect can be quantified. However, a strong influence of the experimental instructions has also been detected. The interviews in turn shed a light on how the emotional induction mechanism works and why it doesn't work in some cases. By means of this the results of the statistical analysis were differentiated and provided with additional explanatory meaning.

# Appendix

**Proof of Theorem 4.1:** The test persons are assumed to be i.i.d. and the $\overline{D_{i\cdot k}}$ are weighted averages of differences between bounded random variables. Therefore, $\mathrm{Var}\overline{D_{i\cdot k}} < \infty$ and $\overline{D_{i\cdot k}}, k \in I\!N_n$ i.i.d. $S_n^2$ converges a.s. to $\mathrm{Var}\overline{D_{i\cdot k}} > 0$ because of (4.6). 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 $\tilde{x} \in \{1,\ldots,p\}^{m_1+\ldots+m_l}$ 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 (4.7),

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

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

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

$$E\left(\overline{D_{i\cdot 1}}\right) = E\left[E\left(\overline{D_{i\cdot 1}}|(X_{0qr1})_{qr} = \tilde{x}\right)\right] = E\left[\frac{\sum_{x=1}^{p} w_{x,\tilde{x}} a(x,\tilde{x})}{\sum_{x=1}^{p} w_{x,\tilde{x}}}\right]. \tag{8.1}$$

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 4.2:** The notation of the proof of Theorem 4.1 is used. Observe $a(x,\tilde{x}) = c$ under (4.9) regardless of $x$ and $\tilde{x}$ unless $w_{x,\tilde{x}} = 0$. Therefore, $E\left(\overline{D_{i\cdot 1}}\right) = c$ by (8.1). Thus, $\mathrm{Var}\left(\overline{D_{i\cdot 1}}\right)$ must be minimized to maximize $a_1$. By (4.7) and (4.9),

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

$$\mathrm{Var}\left(\overline{D_{i\cdot 1}}|(X_{0qr1})_{qr} = \tilde{x}\right) = \frac{\sum_{x=1}^{p} w_{x,\tilde{x}}^2 \frac{n_i(x,\tilde{x}) + n_{-i}(x,\tilde{x})}{n_i(x,\tilde{x}) n_{-i}(x,\tilde{x})}}{\left(\sum_{x=1}^{p} w_{x,\tilde{x}}\right)^2} V,$$

which is minimized for given $\tilde{x}$ by $w_{x,\tilde{x}} = \frac{n_i(x,\tilde{x}) n_{-i}(x,\tilde{x})}{n_i(x,\tilde{x}) + n_{-i}(x,\tilde{x})}$.

## References

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.

Bortz, J. and Döring, N. (1995) *Forschungsmethoden und Evaluation. 2nd edition.* Springer, Berlin.

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

Hand, D. J. (1996) Statistics and the Theory of Measurement, *Journal of the Royal Statistical Society A, 159,* 445-492.

Harrer, G. (1993) Beziehung zwischen Musikwahrnehmung und Emotion. In: Bruhn, H, Oerter, R. and Rösing, H. (Eds.) *Musikpsychologie: Ein Handbuch.* Rowohlt, Reinbek, 588-599.

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

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

Knepler, G. (1982) *Geschichte als Weg zum Musikverständnis: Zur Theorie, Methode und Geschichte der Musikgeschichtsschreibung.* Reclam, Leipzig.

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

Marx, W. (1982) Das Wortfeld der Gefühlsbegriffe. *Zeitschrift für experimentelle und angewandte Psychologie XXIX, 1,* 137-146.

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.

Mecklenbräuker, S. and Hager, W. (1986) Zur experimentellen Variation von Stimmungen: Ein Vergleich einer deutschen Adaption der selbstbezogenen Velten-Aussagen mit einem Musikverfahren. *Zeitschrift für experimentelle und angewandte Psychologie XXIII, 1,* 71-94.

Micko, H. C. (1962) Die Bestimmung subjektiver Ähnlichkeiten mit dem semantischen Differential. *Zeitschrift für experimentelle und angewandte Psychologie IX,* 242-280.

Mikula, G. and Schulter, G. (1970) Polaritätsauswahl, verbale Begabung und Einstufung im Polaritätsprofil. *Zeitschrift für experimentelle und angewandte Psychologie XVII,* 371-385.

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

Pekrun, R. (1985) Musik und Emotion. In: Bruhn, H, Oerter, R. and Rösing, H. (Eds.) *Musikpsychologie: Ein Handbuch in Schlüsselbegriffen.* Urban & Schwarzenberg, Munich, 180-188.

Rösing, H. (1993) Musikalische Ausdrucksmodelle. In: Bruhn, H, Oerter, R. and Rösing, H. (Eds.) *Musikpsychologie: Ein Handbuch.* Rowohlt, Reinbek, 579-587.

Samuels, M. L. (1993) Simpson's paradox and related phenomena, *J. Amer. Stat. Assoc. 88,* 81-88.

Schubert, E. (1996) Enjoyment of negative emotions in music: An assotiative network explanation. *Psychology of Music 24,* 18-28.

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

Velleman, P. F. and Wilkinson, L. (1993) Nominal, ordinal, interval, and ratio scales typologies are misleading. *The American Statistician, 47,* 65-72.

Zentner, M and Scherer, K.R. (1998) Emotionaler Ausdruck in Musik und Sprache. In: Behne, K.-E., Kleinen, G. and de la Motte-Haber, H. (Eds.) *Musikpsychologie: Jahrbuch der deutschen Gesellschaft für Musikpsychologie, Vol. 13: Musikalischer Ausdruck.* Hogrefe, Gttingen, 8-25.