Measuring Change with Multiple Visual Analogue Scales: Application to Tense Arousal

Stéphane Vautier

Measuring Change with Multiple Visual Analogue Scales: Application to Tense Arousal

Stéphane Vautier
Université de Toulouse

Abstract

Although the visual analogue scale (VAS) may be useful for measuring change on subjective and potentially transient phenomena, there is concern about the reliability and construct validity of the associated measurement variables. The present study reports evidence for tau-equivalence of change scores associated with VASs designed for assessing tense arousal with synonymous indicators. This psychometric property allows an estimation of the true-score structure of the cross-sectional measurement variables in a longitudinal SEM model, including method effects. Findings suggest that VASs associated with synonymous indicators may yield highly reliable measurement variables. However, imperfect dynamic bipolarity was observed when data based on antonymous indicators were introduced into the analyses, a rather puzzling effect, which deserves further elaboration.

To monitor subjective and potentially transient phenomena like pain and mood in patients, sensitive and easily administered assessment techniques are required, especially when patients are requested to rate themselves repeatedly. The well-known visual analogue scale (VAS), also called the graphic rating scale (Freyd, 1923), is appealing for this reason. "A VAS is a straight line, the end anchors of which are labeled as the extreme boundaries of the sensation, feeling, or response to be measured" (Wewers & Lowe, 1990, p. 227). The respondent is asked to mark the line to indicate the intensity of the internal stimulus, and a linear function of the corresponding distance is assumed to measure it (Hofmans & Theuns, 2008; Hofmans, Theuns, & Mairesse, 2007; Krabbe, Stalmeier, Lamers, & Busschbach, 2006).

From Precision to Reliability

Clinicians using VASs may feel concerned about the precision of the outcomes (e.g., Kelly, 2001; Roach, Brown, Dunigan, Kusek, & Walas, 1997). Lack of precision may be

---

Stéphane Vautier, OCTOGONE-CERPP, Université de Toulouse, France. I am grateful to two anonymous reviewers who helped improve the paper.
Correspondence concerning this article should be sent to Stéphane Vautier, OCTOGONE-CERPP, Pavillon de la Recherche, 5 allées A. Machado, 31058 Toulouse Cedex 9, France. E-mail: vautier@univ-tlse2.fr.

framed in terms of *measurement error*. The measurement error is, by definition, the difference between the observed measurement and its associated true value. However, since pain or mood experienced in a given situation are phenomena inaccessible to objective scrutiny (Zealley & Aitken, 1969), the true values cannot be determined, and hence the measurement errors are unknown.

Classical Test Theory offers a theoretical framework from which to address the precision issue statistically, by assessing the *reliability* of a *random measurement variable*, which construes any observed score as the outcome of a postulated stochastic process (e.g., Lord & Novick, 1968; Steyer, 1989; Zimmerman, 1975). The random measurement variable is axiomatically decomposed into a true-score variable and an error variable, and its reliability is defined as a proportion of true variance. As observed measurements from a perfectly reliable random measurement variable are error free, the reliability estimate can be used to evaluate the probability that observed measurements will depart from perfect precision; the less reliable the variable, the more the measurements may lack precision (see Charter & Feldt, 2001). A psychometric task consists then of "assessing the reliability of a VAS a", that is, of assessing the reliability of the random measurement variable Y_{at} actually associated with VAS a, where t denotes the testing circumstance.[1]

In their review of the topic, Wewers and Lowe (1990) stressed appropriately that, as far as the phenomenon to be measured is dynamic, the widespread test-retest correlation approach to the reliability of a single measurement procedure is not suitable for the VAS technique. For the test-retest correlation to be interpreted as a reliability coefficient, the variables entering the correlation must be *essentially tau-parallel*, an unlikely assumption when subjective phenomena like pain or mood are to be measured.[2]

Recent work in psychometric modelling (Vautier, Steyer, & Boomsma, 2008) allows one to assess the reliability of a VAS a if at least two VASs, a and b, can be used in a test-retest design, and if the difference variables, $Y_{a2} - Y_{a1}$ and $Y_{b2} - Y_{b1}$, where 1 and 2 index the test and the retest circumstances respectively, can be assumed to be *tau-equivalent*. This is an assumption that can be tested through confirmatory factor analysis (Vautier, Gaudron, & Jmel, 2004).[3] Furthermore, if measurements are used to assess change in a subjective phenomenon (e.g., Grunhaus, Dolberg, Polak, & Dannon, 2002; Wigers, Skrondal, Finset, & Gotestam, 1997; Zealley & Aitken, 1969), clinicians could be more interested in the reliability of the *difference* variable , since the precision of a difference score is at stake (Zimmerman & Williams, 1998). Finally, it is a well known theorem of Classical Test Theory that if a set of variables can be assumed to be tau-equivalent, their sum can be used as a more reliable variable without loss of validity with respect to the true scores; the reliability coefficient

[1] Although reliability is not a property of a measurement device but of an associated random measurement variable (Thompson & Vachaa-Haase, 2000), it is customary in the assessment literature to read about the reliability of a measurement device. For the sake of simplicity I will conform to that custom, which can be understood as the concern for consensus in the use of specific measurement devices for given purposes: Reliability statistics are useful for ordering the overall precision of a set of competitive assessment procedures.

[2] One must assume that (a) true change between the test and the retest is a constant, (b) error variance is a constant, and (c) error variables at the test and the retest are uncorrelated. Assumption (a) is hardly plausible as will be shown in the Results section. See also Gaudron and Vautier (2007).

[3] Tau-equivalency specifies essential tau-parallelism by constraining the true-score variables to be equal (vs. equality up to a constant in essential tau-parallelism), and relaxes essential tau-parallelism by freeing the equality constraint of the error variances.

of the composite variable is Cronbach's alpha (1951). Thus, the theorem applies to the measurement of difference scores as well.

To further specify the problem of forming a multiple VAS device in the area of affective phenomena, one has to take into account the bipolarity issue (e.g., Cacioppo, Gardner, & Berntson, 1997, 1999; Green, Salovey, & Truax, 1999; Raufaste & Vautier, 2008; Russell & Carroll, 1999; Schimmack, 2001; Steyer & Riedl, 2004). According to Wewers and Lowe (1990), "scales that contain bipolar anchors (i.e., depression–elation) are discouraged since they compound conceptual difficulties by introducing two phenomena" (p. 233). Consider the unipolar VAS a, the anchors of which are "no depression" and "worst depression I can imagine", and the unipolar VAS b, the anchors of which are "no elation" and "elation as high as it could possibly be". Can one assume that the two VASs measure change tau-equivalently, even when one considers reversed scores on VAS b? Such an issue deserves empirical verification.

The present study draws on and extends Vautier et al.'s (2008) work to investigate from a psychometric point of view the suitability of forming a multiple VAS device to measure change in mood, starting from a set of VASs based on synonyms and antonyms (see also Vautier, Steyer, Jmel, & Raufaste, 2005). For the purpose of illustration, I will focus on the concept of tense arousal, as opposed to energetic arousal and valence (Schimmack & Reisenzein, 2002). Thus, using a set of indicators like {tendu(e), crispé(e), sous pression, décontracté(e), relaxé(e), détendu(e)},[4] the question arises as to whether the corresponding unipolar VASs may be assumed to measure change tau-equivalently, if the relevant score reversing is provided.

If tau-equivalence of the difference variables can be corroborated, a multiple VAS device can be proposed as a tool statistically grounded for the measurement of change in tense arousal: Its associated composite variable Y_{ot} can be used as a measurement variable of a latent composite at time t, where o means the summation of the measurement variables associated with each VAS, and t indexes the test or the retest circumstances. Interestingly, even if the structure of the latent composite is unknown, it is of theoretical importance that the difference variable $Y_{o2} - Y_{o1}$ can be interpreted as a measurement variable of something that is characterized by *intraindividual change measured tau-equivalently with each VAS*. Clearly, tau-equivalence in the present context means that the measurement of a latent change q is reproducible with all the VASs of the device; this is why the construct validity of the measurements rests strongly on tau-equivalence, an emerging statistical property, which can be interpreted causally as the structural effect of the phenomena to be measured (see Borsboom, Mellenbergh, & Van Heerden, 2003). Moreover, aggregating several VASs allows the resulting variable $Y_{o2} - Y_{o1}$ to exhibit higher reliability than the single-VAS difference variables.

If tau-equivalence of the difference variables is rejected, one has no empirical support for interpreting scores from the composite measurement variable $Y_{o2} - Y_{o1}$ as measurements of something that can be reproduced. Yet, the multivariate character of the measurements is useless from the viewpoint of internal validity. Although in practice there is no need to know the underlying statistical structure of a composite measurement variable to use the associated assessment technique – i.e., the process of getting answers and computing scores

[4] For an English translation of the indicators, see Appendix A (raw translation only, not suitable for application).

–, one would appreciate evidence ascertaining that the measurements reflect a consistent statistical phenomenon, viz., the structure of tau-equivalence.

From Tau-equivalence to Method Effects

Thus, it seems theoretically worthwhile, and it is practically easy, to examine the latent structure of the composite true-score difference variable associated with a multiple VAS device by fitting a suitably constrained one-factor model to the observed difference variables (e.g., Jöreskog, 1971). Once a set of VASs has been qualified for its ability to measure change tau-equivalently, it is possible to "unfold" the statistical model to assess the importance of *individual method effects*, which likely occurred at each testing circumstance (Vautier et al., 2008). An individual method effect is defined at the level of a testing circumstance as the difference between two true scores, namely the true score measured with VAS b minus the true score measured with VAS a, where VAS a serves as the measurement method of reference–for more details, see Pohl, Steyer, and Kraus (2008); Steyer (2005); Vautier and Pohl (2009); Vautier et al. (2008), and Appendix B. As individual method effects bias the measurement of the true scores associated with the reference measurement method, it is a construct validity issue to assess their statistical effects through the modeling of method components in the true-score structure.

A *method component* in the unfolded statistical model of the latent structure represents interindividual variability in the individual method effects. If the method components can be assumed to be temporally reproducible, they cancel out in difference scores, and the VASs can measure change tau-equivalently–which makes the model algebraically identified (Vautier & Pohl, 2009; Vautier et al., 2008). The statistical importance of a method component can be assessed by its effect size μ/σ. The smaller the variance of the method component, the larger its absolute effect size; the sign of μ/σ depends on the sign of μ : a positive (negative) sign means that on average, the indicator associated with the method component is less (more) difficult to endorse than the reference indicator.

In the context of the VAS technique, a method component can be interpreted as the effect of individual differences in the way respondents order the various VASs, depending on their interpretation of the semantic nuances conveyed by the corresponding indicators. For example, in the French language, the indicator *crispé(e)* seems semantically stronger than the indicator *tendue(e)*, and it can be expected that on average, the true score on *crispé(e)* is smaller than the true score on *tendue(e)*. If the individual method effects elicit individual differences, the method component will have a non-null variance, and the effect size will be negative. If one assumes that the "difficulty" of the indicator does not depend on the respondents, there are no individual differences in the individual method effects, and the variance of the method component will equate to zero, which yields an undefined effect size–if the variance tends toward zero, the effect size tends toward infinity. The concept of item difficulty in mood questionnaires can be viewed as a special case of the concept of individual method effects, that is, the measurement variables at t are tau-equivalent, which is not the case if a method component with a non-null variance does exist (for more details, see Vautier et al., 2008, Section 6). Individual differences in method effects are testable, as will be shown in the Method section.

If the method components have a non-negligible variance, the composite score formed by adding the scores from each VAS at time t will reflect uncontrolled measurement biases,

the population mean of which can be easily assessed. Concerning the interpretation of the true variance of the composite measurement variable Y_{ot}, biases due to individual method effects threaten the validity of Y_{ot}. Thus, the analysis based on the unfolded model may be helpful for those interested in having a look at the latent structure of the statistical phenomenon observed in a given testing circumstance.

Research Goals

To recapitulate, given a set of VASs based on "positive" and "negative" indicators of tense arousal, the psychometric hypothesis to be tested is that the VASs measure change tau-equivalently. Such a hypothesis is consistent with a realistic approach to the phenomena to be measured (Borsboom et al., 2003). If tau-equivalence of the change variables does not hold, an alternative model that could enable one to figure out the kind of dynamics the VASs measure would be welcome. Previous work on dynamic bipolarity suggests the alternative interpretation according to which two kinds of change could be measured, depending on the semantic polarity of the indicators (Raufaste & Vautier, 2008; Vautier & Raufaste, 2003; Vautier et al., 2005). Vautier et al. (2005) call such a feature "imperfect dynamic bipolarity", as opposed to "perfect dynamic bipolarity". Perfect dynamic bipolarity means that latent change measured by the negatively cued indicators correlates -1 with latent change measured by the positively cued indicators. In both cases, in order to assess the reliability of the relevant composite difference variable, compare it with the reliabilities of the difference variables associated with each VAS, and evaluate the importance of the method effects in terms of their means and variances, it is necessary to highlight the latent structure of the cross-sectional composite true-score variables under the test and retest circumstances.

Method

Participants, Material, and Design

The data from a sample of 909 French adults were analysed. As part of a survey, the sample had answered a 24-VAS self-rating mood questionnaire twice, with a time lag of 28.06 ± 6.95 days between test and retest. The questionnaire was adapted from the Multidimensional Mood Questionnaire (Steyer, Schwenkmezger, Notz, & Eid, 1997). The respondents were contacted by undergraduate psychology students who were instructed to conduct the interviews as part of their course. The six indicators of interest were scattered amongst other VASs corresponding to different mood indicators. The response format was a continuous 102 mm horizontal line, the extremities of which were marked *not at all* and *extremely*, with the numerical equivalent coded by the interviewers. Appendix A displays the ordered list.

Analyses

Analyses were conducted in seven steps.

Step 1: Selecting a suitable sample. The first step was aimed at selecting a suitable subsample of respondents by looking for a class of consistent response profiles and excluding potentially careless respondents (Schmitt & Stults, 1985; Woods, 2006). A consistent response profile was defined as follows. Let N_{is} and P_{jt} denote the measurement variables,

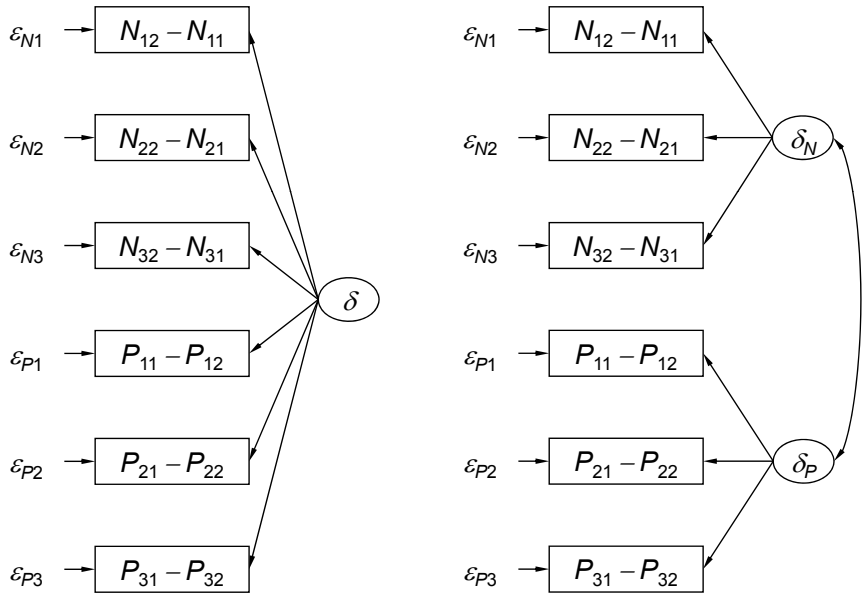

Figure 1. Path diagrams of the models used to test tau-equivalence of the difference measurement variables (left panel) and tau-equivalence of the difference measurement variables with imperfect dynamic bipolarity (alternative model, right panel). N_{it} and P_{it} respectively denote the negative and the positive manifest variable associated with the ith measurement method at the testing circumstance t; δ denotes the true-change variable. Factor loadings are set at unity; measurement intercepts are set at zero. Residual variables are uncorrelated.

where N and P respectively denote the positive and negative polarity of the corresponding indicator (see Appendix A), i and j denote the VAS number, and s and t denote the test or retest circumstance. Consistency of an individual response profile can be defined by a low profile on the following series of standardised deviations $s(N_{11},\ N_{21},\ N_{31})$, $s(P_{11},\ P_{21},\ P_{31})$, $s(N_{12},\ N_{22},\ N_{32})$, and $s(P_{12},\ P_{22},\ P_{32})$. These data were clustered via a K-means technique. Respondents belonging to the uniform lower class can hardly be suspected of carelessness in the way they treated the six VASs.

Step 2: Testing tau-equivalence of the difference variables. The second step consisted of testing tau-equivalence of the six difference measurement variables, namely $N_{12} - N_{11}$, $N_{22} - N_{21}$, $N_{32} - N_{31}$, $P_{11} - P_{12}$, $P_{21} - P_{22}$, and $P_{31} - P_{32}$. A confirmatory one-factor analytic model was specified in such a way that the variance and the mean of the factor δ were estimated, the factor loadings were fixed at one, and the measurement intercepts were fixed at zero; the residuals were uncorrelated (see Figure 1, left panel). Thus, a one-factor model such as this depicts tau-equivalence of its manifest components.

Step 3: Adding the hypothesis of imperfect dynamic bipolarity. As tau-equivalence of the difference measurement variables was rejected, the third step was aimed at fitting the data by adding the feature of imperfect dynamic bipolarity in the model. Hypothesizing imperfect dynamic bipolarity consisted of generalising the previous model by splitting the common factor into two correlated common factors, namely one factor representing interindividual variability in change measured tau-equivalently by the negative VASs on the one hand, and one factor representing interindividual variability in reversed change measured tau-equivalently by the positive VASs on the other hand (see Figure 1, right panel).

Step 4: Assessing the merit of aggregating the VASs to improve the reliability of change measurements. Once a satisfactory latent structure was found to account for the data, it was possible to specify two multiple VASs that could be considered able to provide tau-equivalent measurements of change, and thus to use Cronbach's alpha to estimate how reliably the change could be measured, with respect to how reliably each of the single VASs could measure the same change.

Step 5: Unfolding the model. The next step consisted of detailing the cross-sectional latent structure of the data, using Vautier et al.'s (2008) modeling (see also Vautier & Pohl, 2009). The modeling is depicted in Figure 2. True-score variables associated with measurement variables N_{11} and RP_{11}, where RP indicates that the variable P has been reverse scored, are denoted τ_{N11} and τ_{P11} respectively. The items *tendu(e)* and *décontracté(e)* were used as the reference methods (see Appendix B for details about the meaning of a reference method). Method components are denoted by υ_i, $i = 2,\ 3,\ 5,\ 6$. The independent variables were freely correlated, although this is not depicted in the figure for the sake of legibility, and their variances and means were freely estimated. Measurement intercepts were fixed at zero.

Step 6: Testing the method components. The sixth step consisted of refining the unfolded model by testing the need for including the method components in the model. Testing a method component means replacing it by two equal measurement intercepts, which ensures that the method component has no variance.

Step 7: Decomposing the variance of the cross-sectional composite-score variables. The final step consisted of using the parameter estimates of the final model to express the variance of the four composite-score variables entering the two multiple VAS approaches to tau-equivalent measurement of change, namely, $N_{ot} = N_{1t} + N_{2t} + N_{3t}$, and $RP_{ot} = RP_{1t} + RP_{2t} + RP_{3t}$, $t = 1,\ 2$. Such a decomposition documents the internal validity of the cross-sectional measurements. Ideally, a measurement variable should reflect a single component of true score variance . If the true score reflects the combination of a mixture, it is hard or even impossible to interpret individual differences with respect to these sources of variance. Method components act as undesirable true components in the total true variance, and its decomposition may help in assessing their relative importance.

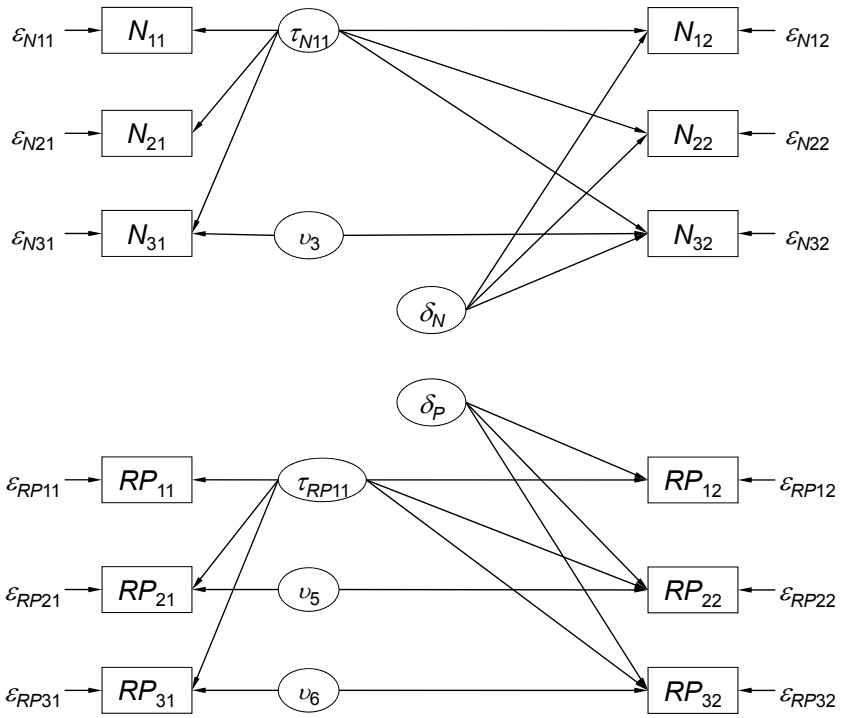

Figure 2. Path diagram of the "unfolded" model of the cross-sectional measurement variables with imperfect dynamic bipolarity. υ_i denotes the method component i associated with VAS i with respect to the reference VAS 1. Factor loadings are set at unity; measurement intercepts are set at zero. Residual variables are uncorrelated. The independent variables are freely correlated with each other.

Results

The raw data as well as the Mplus command and output files are available online.[5] (Step 1) The K-means clustering analysis suggested the selection of about half of the original sample as consistent respondents, $N = 427$. Such a sample size is sufficiently large for the subsequent SEM analyses, and I will discuss the practical implications of this finding later.

As multivariate normality of the data cannot be assumed, all models were estimated by the maximum likelihood robust estimator as implemented in Mplus (Muthén & Muthén, 2007), and the goodness of fit was evaluated by using the χ^2 statistics–the RMSEA statistics are also reported. (Step 2) The difference measurement variables did not exhibit tau-equivalence, as the one-factor model did not fit the data, $\chi^2(N = 427, df = 19) = 380.06$, $p < .0001$, $RMSEA = .211$, $p(RMSEA < .05) < .001$. (Step 3) The alternative model implementing imperfect dynamic bipolarity fit the data very well, $\chi^2(N = 427, \mathrm{df} = 16) = 17.65$, $p = .34$, $RMSEA = .014$, $p(RMSEA < .05) < .96$. The two factors correlated .82 (.03), and their estimated variances were 860.30 (79.93) for change measured by the "negative"

[5] http://w3.cerpp.univ-tlse2.fr/annuaire/vautier/LiensIndex/Publications.htm.

VASs, and 889.92 (80.39) for change measured by the "positive" VASs. Notice that essential tau-parallelism, which allows for an interpretation of the test-retest correlation as an index of reliability, is rejected, since it implies that the change factors have a null variance. (Step 4) Furthermore, the corresponding parameter estimates were used to address the reliability issue. As the unique variance corresponds to the error variance, the coefficient R^2 corresponds to the reliability coefficient of each difference measurement variable. The estimated values ranged from .82 (.02) to .89 (.02). Cronbach's alpha estimates were respectively .95 and .94.

(Step 5) The unfolded model fit the data very well, $\chi^2(N = 427, df = 34) = 36.89$, $p = .34$, $RMSEA = .014$, $p(RMSEA < .05) < .996$. The estimated variances of method components υ_2, υ_3, υ_5 were nonsignificant or marginally significant; the estimated means of method components υ_3 and υ_6 were nonsignificant. (Step 6) If the method component υ_2 has no variance, it reduces to a constant, and then the variables N_{11} and N_{21} are essentially tau-equivalent on the one hand, and N_{12} and N_{22} are essentially tau-equivalent on the other hand. This feature is modelled by introducing three changes into the original model depicted in Figure 1. (a) The method component υ_2 is removed, which entails that the variables N_{11} and N_{21} are tau-equivalent on the one hand, and that N_{12} and N_{22} are tau-equivalent on the other hand, which is too strong of an assumption. Thus (b) the measurement intercepts of the variables N_{21} and N_{22} are relaxed, which entails that the difference variables $N_{12} - N_{11}$ and $N_{22} - N_{21}$ may be essentially tau-equivalent, which is too weak of an assumption since it is assumed from Step 3 that they are tau-equivalent. This is corrected by (c) constraining the intercepts of N_{21} and N_{22} to be equal, which yields tau-equivalence of $N_{22} - N_{21}$ and $N_{12} - N_{11}$. The resulting model fit the data better, $\chi^2(N = 427, df = 42) = 44.74$, $p = .36$, $RMSEA = .012$, $p(RMSEA < .05) < .999$. That specification was thus retained in the subsequent models.

Tau-equivalence of the variables N_{i1} on the one hand, and the variables N_{i2} on the other hand was then tested by replacing the method component υ_3 by equated measurement intercepts affecting N_{31} and N_{32}. The model fit the data very well, $\chi^2(N = 427, df = 49) = 66.31$, $p = .05$, $RMSEA = .029$, $p(RMSEA < .05) < .986$. However, the appropriate χ^2 difference test was significant, $\chi^2(7) = 21.17$, $p = .004$, so the absence of individual differences in the method component υ_3 (i.e., $s^2(\upsilon_3) = 0$) was rejected.

Tau-equivalence of the variables P_{11} and P_{21} on the one hand, and the variables P_{12} and P_{22} on the other was then tested following the same method. The model fit the data very well, $\chi^2(N = 427, df = 49) = 65.09$, $p = .06$, $RMSEA = .028$, $p(RMSEA < .05) < .989$. However, the appropriate χ^2 difference test was significant, $\chi^2(7) = 21.40$, $p = .003$, so the absence of individual differences in the method component υ_5 was rejected.

Finally, tau-equivalence of the variables P_{11} and P_{31} on the one hand, and the variables P_{12} and P_{32} on the other was tested following the same method. The model fit the data very well, $\chi^2(N = 427, df = 49) = 65.23$, $p = .06$, $RMSEA = .028$, $p(RMSEA < .05) < .988$. However, the appropriate χ^2 difference test was significant, $\chi^2(7) = 21.77$, $p = .003$, so the absence of individual differences in the method component υ_6 was rejected.

The final model included all method components except υ_2, which was replaced by a constant measurement intercept, the estimate of which was -2.54 (0.39). The estimated values of the effect sizes of the remaining method components were $d(\upsilon_3) = 0.603/19.731^{0.5} = 0.136$, $d(\upsilon_5) = 3.414/17.347^{0.5} = 0.820$, and $d(\upsilon_6) = 0.683/35.035^{0.5} = 0.115$. The estimated

Table 1: Estimated Reliabilities of the Cross-sectional Measurement Variables

Variable	Reliability Estimate	Standard Error
N_{11}	.911	.013
N_{12}	.901	.013
N_{21}	.907	.013
N_{22}	.908	.013
N_{31}	.893	.017
N_{32}	.898	.015
RP_{11}	.899	.016
RP_{12}	.855	.025
RP_{21}	.909	.017
RP_{22}	.821	.029
RP_{31}	.955	.013
RP_{32}	.789	.038

reliabilities of the cross-sectional variables ranged from .789 to .955, and are displayed in Table 1.

(Step 7) Decomposing the variance of N_{o1} requires detailing its algebraic formula, which reads

$$N_{o1} = 3 \cdot \tau_{N11} - 2.537 + \upsilon_3 + \varepsilon_{N11} + \varepsilon_{N21} + \varepsilon_{N31}. \tag{1}$$

It follows that its variance decomposes as

$$s^2(N_{o1}) = 9 \cdot s^2(\tau_{N11}) + s^2(\upsilon_3) + 6 \cdot s(\tau_{N11},\ \upsilon_3) + s^2(\varepsilon_{N11}) + s^2(\varepsilon_{N21}) + s^2(\varepsilon_{N31}), \tag{2}$$

where $s^2(.)$ and $s(.,\ .)$ denote the estimated variance and covariance, respectively. The quantity $9 \cdot s^2(\tau_{N11})$, which refers to the true-score variable of reference, represents 95.92% of the total variance $s^2(N_{o1})$.

The mean bias of N_{o1} can be computed starting from the mean formula

$$m(N_{o1}) = 3 \cdot m(\tau_{N11}) - 2.537 + m(\upsilon_3), \tag{3}$$

where $m(.)$ denotes the estimated mean. Thus, it turned out that the mean bias was about -1.93 points out of 306 points.

The formula of the composite variable N_{o2} is complicated by the change factor, which yields

$$N_{o2} = 3 \cdot (\tau_{N11} + \delta_N) - 2.537 + \upsilon_3 + \varepsilon_{N12} + \varepsilon_{N22} + \varepsilon_{N32}. \tag{4}$$

Hence, its variance decomposes as follows:

$$\begin{aligned} s^2(N_{o2}) = 9 \cdot [&s^2(\tau_{N11}) + s^2(\delta_N) + 2 \cdot s(\tau_{N11},\ \delta_N)] \\ &+ s^2(\upsilon_3) + 6 \cdot [s(\tau_{N11},\ \upsilon_3) + s(\delta_N,\ \upsilon_3)] \\ &+ s^2(\varepsilon_{N12}) + s^2(\varepsilon_{N22}) + s^2(\varepsilon_{N32}). \end{aligned} \tag{5}$$

The quantity $9 \cdot [s^2(\tau_{N11}) + s^2(\delta_N) + 6 \cdot s(\tau_{N11},\ \delta_N)]$, which refers to the true-score variable of reference at the retest circumstance, represents 95.07% of $s^2(N_{o2})$. Because of the temporal

stability of the method components, the mean bias of N_{o2} has the same value as that affecting the composite N_{o1}.

Following similar calculus computations, it was found that the proportions of "good" variance in $s^2(RP_{o1})$ were about 96.24% and 93.03% in $s^2(RP_{o2})$, while the mean bias was about 4.10 points.

Discussion

Using the VAS technique for measuring change in subjective phenomena like pain or mood is a long-running practice (Aitken, 1969; Freyd, 1923). As the response format of a VAS is a continuous line, it may provide sensitive outcomes, although there is concern about potential lack of precision. For example, it has been found that the amplitude of the average minimum clinically significant difference in 100 mm VAS pain scores is about 10 mm (Kelly, 2001). Reliability analysis could be useful to document the precision issue, provided that suitable estimates are available. Particularly, one would like to be able to assess reliability of a difference measurement variable while figuring out its true-score structure, a critical aspect of construct validity. If a set of VASs measure change tau-equivalently, the reliability of (a) the difference variables associated with each VAS, (b) the composite difference variable formed by their sum, as well as the reliability of (c) the single and (d) composite cross-sectional measurement variables may be assessed by using parameter estimates of relevant structural equation models (Vautier et al., 2008).

The present study investigated measurement variables associated with a set of VASs designed for the assessment of tense arousal in a sample of French adults. The main findings can be summarized as follows: (a) The VASs based on indicators sharing the same semantic polarity can be assumed to measure change tau-equivalently. This psychometric property gives statistical evidence for a changing phenomenon that underlies the observed responses in a reproducible way: if VAS a measures a quantity of change q, then VASs b and c will measure the same quantity of change q as well. (b) The estimated reliability of the two composite difference measurement variables that can be defined with the single difference measurement variables was about .95, a high level given the cautions that are traditionally raised against the use of differences based on psychometric scores (Cronbach & Furby, 1970, but see Zimmerman & Williams, 1982). (c) The hypothesis that VASs based on positive and negative indicators measure the same and only the same changing phenomenon was rejected. In other words, change measured by a VAS based on a negative indicator could not be reproduced exactly by change measured with a VAS based on a positive indicator, and vice versa. (d) From a cross-sectional perspective, the composite-score measurement variables exhibited nice properties, as the "good" proportion of variance amounted to about 95% of the total variance, and the average bias due to individual method effects was quite limited with respect to the metric of the composite scale.

As data from "careless" respondents (Schmitt & Stults, 1985) may obscure the interpretation of factor analytic results (see also Woods, 2006), a preliminary caution consisted of selecting consistent respondents from the original sample. The present study suggests that the proportion of consistent respondents can be low, as about half of the sample was retained by a clustering analysis. Thus, depending on the experimental conditions of data collection, researchers should be well advised to take care of the quality of their dataset before analysing its moment structure. As the respondent plays the role of a subjective

gauge during the rating process, such a low proportion suggests the following research issue: Could the quality of the data be improved substantively if, before the rating task, respondents were instructed in the bipolar principle of the scale development, and on how carefully to treat the direction of the items? Vautier, Mullet, and Bourdet-Loubère (2003) showed that instructing the respondents to treat a series of nine state-anxiety items as indicators of a single construct can impact the structural properties of the data.

However, imperfect dynamic bipolarity in the data could not be imputed to careless respondents as soon as a proper subsample was selected. Such a finding raises the substantive issue of understanding how semantic bipolarization of the indicators could elicit distinct factors. As no evidence for imperfect dynamic bipolarity was found in ordered categorical data (Vautier et al., 2005) and quasi-continuous composite data (Vautier & Pohl, 2009), the present finding suggests that the assessment technique is not neutral with respect to the target phenomenon to be measured. Moreover, imperfect dynamic bipolarity raises the practical issue of which multiple VAS to choose: the one based on the negative indicators, or that based on the positive indicators?

Tau-equivalence of a set of difference measurement variables allows for the assumption of temporally stable individual method effects (Vautier & Pohl, 2009; Vautier et al., 2008). Such an assumption grounds the use of VASs at the intraindividual level. In other words, the person may be viewed as her own stable gauge of the subjective phenomenon to be assessed. Thus, two ratings from the same person using the same device can be compared to each other, and their difference can be interpreted as a measurement of change on a latent attribute that possesses some measurement invariance at the individual scale and within the time lag associated with the measurement experiments. The present findings suggest that multiple measurements based on synonyms can be useful for obtaining composite measurements of intraindividual change that exhibit internal validity. Interpreting these measurements as outcomes of an abstract random measurement variable, which is more reliable than the single random measurement variables associated with each VAS, a better precision is to be expected if change is assessed by a composite difference score.

References

Aitken, R. C. B. (1969). Measurement of feeling using visual analogue scales. *Proceedings of the Royal Society of Medicine*, *62*, 989–993.

Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2003). The theoretical status of latent variables. *Psychological Review*, *110*, 203–219.

Cacioppo, J. T., Gardner, W. L., & Berntson, G. G. (1997). Beyond bipolar conceptualizations and measures: The case of attitudes and evaluative space. *Personality and Social Psychology Review*, *1*, 3–25.

Cacioppo, J. T., Gardner, W. L., & Berntson, G. G. (1999). The affect system has parallel and integrative processing components: Form follows function. *Journal of Personality and Social Psychology*, *76*, 839–855.

Charter, R. A., & Feldt, L. S. (2001). Confidence intervals for true scores: Is there a correct approach? *Journal of Psychoeducational Assessment*, *19*, 350–364.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297–334.

Cronbach, L. J., & Furby, L. (1970). How we should measure "change"-or should we? *Psychological Bulletin*, *74*, 68–80.

Eid, M. (2000). A multitrait-multimethod model with minimal assumptions. *Psychometrika*, *65*, 241–261.

Freyd, M. (1923). The graphic rating scale. *Journal of Educational Psychology*, *14*, 83–102.

Gaudron, J. P., & Vautier, S. (2007). Estimating true short-term consistency in vocational interests: A longitudinal SEM approach. *Journal of Vocational Behavior*, *70*, 221–232.

Geiser, C., Eid, M., & Nussbeck, F. W. (2008). On the meaning of latent variables in the CT-C($M-1$) model: A comment on Maydeu-Olivares and Coffman (2006). *Psychological Methods*, *13*, 49–57.

Green, D. P., Salovey, P., & Truax, K. M. (1999). Static, dynamic, and causative bipolarity of affect. *Journal of Personality and Social Psychology*, *76*, 856–867.

Grunhaus, L., Dolberg, O. T., Polak, D., & Dannon, P. N. (2002). Monitoring the response to rTMS in depression with visual analog scales. *Human Psychopharmacology: Clinical and Experimental*, *17*, 349–352.

Hofmans, J., & Theuns, P. (2008). On the linearity of predefined and self-anchoring visual analogue scales. *British Journal of Mathematical and Statistical Psychology*, *61*, 401–413.

Hofmans, J., Theuns, P., & Mairesse, O. (2007). Impact of the number of response categories on linearity and sensitivity of self-anchoring scales. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *4*, 160–169.

Jöreskog, K. G. (1971). Statistical analysis of sets of congeneric tests. *Psychometrika*, *36*, 109–133.

Kagan, J. (1988). The meanings of personality predicates. *American Psychologist*, *43*, 614–620.

Kelly, A. M. (2001). The minimum clinically significant difference in visual analogue scale pain score does not differ with severity of pain. *Emergency Medicine Journal*, *18*, 205–207.

Krabbe, P. F. M., Stalmeier, P. F. M., Lamers, L. M., & Busschbach, J. J. V. (2006). Testing the interval-level measurement property of multi-item visual analogue scales. *Quality of Life Research*, *15*, 1651–1661.

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

McArdle, J. J., & Aber, M. (1990). Patterns of change within latent variable equation models. In A. von Eye (Ed.), *Statistical methods in longitudinal research: Vol. 1. Principles and structuring change* (pp. 151–224). San Diego, CA: Academic Press.

Muthén, L. K., & Muthén, B. O. (2007). *Mplus version 5.1*. Computer Program. Los Angeles: Authors.

Pohl, S., Steyer, R., & Kraus, K. (2008). Modelling method effects as individual causal effects. *Journal of the Royal Statistical Society: Series A*, *171*, 41–63.

Raufaste, E., & Vautier, S. (2008). An evolutionist approach to information bipolarity: Representations and affects in human cognition. *International Journal of Intelligent Systems*, *23*, 878–897.

Roach, K. E., Brown, M. D., Dunigan, K. M., Kusek, C. L., & Walas, M. (1997). Test-retest reliability of patient reports of low back pain. *Journal of Orthopaedic & Sports Physical Therapy*, *26*, 253–259.

Russell, J. A., & Carroll, J. M. (1999). On the bipolarity of positive and negative affect. *Psychological Bulletin*, *125*, 3–30.

Schimmack, U. (2001). Pleasure, displeasure, and mixed feelings: Are semantic opposites mutually exclusive? *Cognition and Emotion*, *15*, 81–97.

Schimmack, U., & Reisenzein, R. (2002). Experiencing activation: Energetic arousal and tense arousal are not mixtures of valence and activation. *Emotion*, *2*, 412–417.

Schmitt, N., & Stults, D. (1985). Factors defined by negatively keyed items: The result of careless respondents? *Applied Psychological Measurement*, *4*, 367–373.

Steyer, R. (1989). Models of classical psychometric test theory as stochastic measurement models: Representation, uniqueness, meaningfulness, identifiability, and testability. *Methodika*, *3*, 25–60.

Steyer, R. (2005). Analyzing individual and average causal effects. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *1*, 39–54.

Steyer, R., & Riedl, K. (2004). Is it possible to feel good and bad at the same time? New evidence on the bipolarity of mood-state dimensions. In K. Van Montfort, H. Oud, & A. Satorra (Eds.), *Recent developments in structural equation modeling: Theory and applications* (pp. 197–221). Amsterdam: Kluwer Academic Press.

Steyer, R., Schwenkmezger, P., Notz, P., & Eid, M. (1997). *Der mehrdimensionale befindlichkeitsfragebogen [The multidimensional mood questionnaire]*. Göttingen: Hogrefe.

Thompson, B., & Vachaa-Haase, T. (2000). Psychometrics is datametrics: The test is not reliable. *Educational and Psychological Measurement*, *60*, 174–195.

Vautier, S., Gaudron, J. P., & Jmel, S. (2004). Modèles factoriels linéaires pour l'analyse de la fidélité des variables composites [Linear factor analytic models for reliability analysis of composite variables]. *Revue d'Épidémiologie et de Santé Publique*, *52*, 441–453.

Vautier, S., Mullet, E., & Bourdet-Loubère, S. (2003). The instruction set of questionnaires can affect the structure of the data: Application to self-rated state anxiety. *Theory and Decision*, *54*, 249–259.

Vautier, S., & Pohl, S. (2009). Do balanced scales assess bipolar constructs? The case of the STAI scales. *Psychological Assessment*, *21*, 187–193.

Vautier, S., & Raufaste, E. (2003). Measuring dynamic bipolarity in positive and negative activation. *Assessment*, *10*, 49–55.

Vautier, S., Raufaste, E., & Cariou, M. (2003). Dimensionality of the Revised Life Orientation Test and the status of filler items. *International Journal of Psychology*, *38*, 390–400.

Vautier, S., Steyer, R., & Boomsma, A. (2008). The true-change model with individual method effects: Reliability issues. *British Journal of Mathematical and Statistical Psychology*, *61*, 369–399.

Vautier, S., Steyer, R., Jmel, S., & Raufaste, E. (2005). Imperfect or perfect dynamic bipolarity? The case of antonymous affective judgments. *Structural Equation Modeling*, *12*, 391–410.

Wewers, M. E., & Lowe, N. K. (1990). A critical review of visual analogue scales in the measurement of clinical phenomena. *Research in Nursing and Health*, *13*, 227–236.

Wigers, S. H., Skrondal, A., Finset, A., & Gotestam, K. G. (1997). Measuring change in fibromyalgic pain: The relevance of pain distribution. *Journal of Musculoskeletal Pain*, *5*, 29–41.

Woods, C. M. (2006). Careless responding to reverse-worded items: Implications for confirmatory factor analysis. *Journal of Psychopathology and Behavioral Assessment*, *28*, 189–194.

Zealley, A. K., & Aitken, R. C. B. (1969). Measurement of mood. *Proceedings of the Royal Society of Medicine*, *62*, 993–996.

Zimmerman, D. W. (1975). Probability spaces, Hilbert spaces, and the axioms of test theory. *Psychometrika*, *40*, 395–412.

Zimmerman, D. W., & Williams, R. H. (1982). Gain scores in research can be highly reliable. *Journal of Educational Measurement*, *19*, 149–154.

Zimmerman, D. W., & Williams, R. H. (1998). Reliability of gain scores under realistic assumptions about properties of pre-test and post-test scores. *British Journal of Mathematical and Statistical Psychology*, *51*, 343–351.

Appendix A
Ordered List of the Indicators

tendu(e)–tense	$\rightarrow$	N_1
décontracté(e)–unstrained	$\rightarrow$	P_1
crispé(e)–strained	$\rightarrow$	N_2
relaxé(e)–mellow	$\rightarrow$	P_2
sous pression–stressed	$\rightarrow$	N_3
détendu(e)–relaxed	$\rightarrow$	P_3

Appendix B
Two Approaches to Method Effects in the Tradition of Classical Test Theory

Method effects have been represented in CFA models according to two main traditions. In the factor analysis tradition, method effects are represented by group factors, that is, factors common to a subset of manifest variables to be interpreted as measurements of the same construct (e.g., Vautier, Mullet, & Bourdet-Loubère, 2003; Vautier, Raufaste, & Cariou, 2003). A drawback of the factorial tradition is that it does not ensure that a model with method factors can be deduced from a stochastic measurement model.

In the tradition of Classical Test Theory, each measurement variable is decomposed as the sum of a reliable, or true, component, and of a residual component (Lord & Novick, 1968; Zimmerman, 1975). It is acknowledged that a true component has to be interpreted in an operationalistic way as *reliable variation measured by using a given measurement method* (see Kagan, 1988). Method components have been modelled by using two techniques for decomposing the reliable variance that I will call "regressive decomposition" and "differential decomposition" (see Figure B1). Both approaches rest on a couple of true-score variables (τ_1, τ_2), which are associated with a couple of manifest measurement variables (Y_1, Y_2). The method component is defined in contrast to a reference measurement method. The choice of the reference variable is a matter of convenience. Let the first measurement method be used as the reference method.

In the regressive decomposition approach, the method component, ν_2, is the residual variable of τ_2 once τ_2 has been statistically controlled for by τ_1 (Eid, 2000; Geiser, Eid, & Nussbeck, 2008). Consequently, statements about the mean of ν_2 are not a matter of evidence because, by definition, as a residual variable, ν_2 has a null expectation. Therefore, the method component cannot account for the potential mean difference between the true-score variables. Also by definition, ν_2 does not correlate with τ_1.

In the differential decomposition approach, the method component, υ_2, is the difference variable $\tau_2 - \tau_1$ (e.g., McArdle & Aber, 1990; Pohl et al., 2008; Steyer, 2005; Vautier & Pohl, 2009; Vautier et al., 2008). Consequently, statements about the mean of υ_2 are a matter of evidence. Moreover, υ_2 is characterized by its effect size, provided that its variance is strictly positive. Also, υ_2 may correlate with τ_1. If υ_2 has no variance, it reduces to a constant parameter and Y_1 and Y_2 are *essentially* tau-equivalent – and also strictly tau-equivalent if the value of the constant is zero.

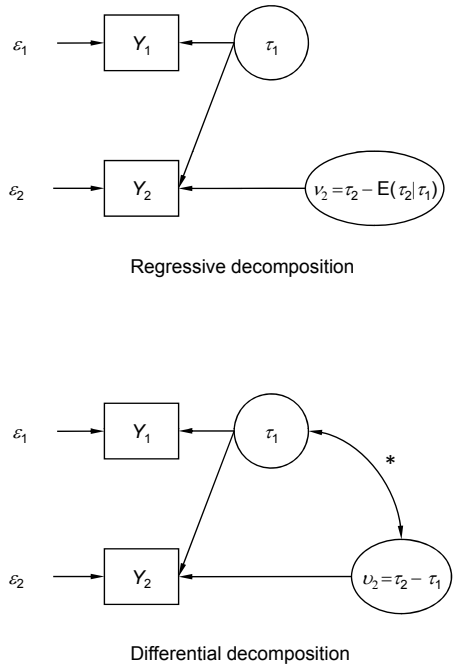

Figure B1. Formal definitions of the method component associated with the true-score variables ($\tau_1,\ \tau_2$). In the regressive decomposition approach, $\tau_2 = \tau_1 + \nu_2$; in the differential decomposition approach, $\tau_2 = \tau_1 + \upsilon_2$. * denotes a freely estimated covariance.