Is the Above-Average Effect Measurable at All? The Validity of the Self-Reported Happiness Minus Other's Perceived Happiness Construct

Stéphane Vautier
Université de Toulouse

Jean-François Bonnefon
CNRS and Université de Toulouse

Individuals routinely rate themselves higher than their peers on a number of attributes and capabilities, including their satisfaction with life. However, the construct validity of this above-average effect requires specific psychometric properties of ratings of one's contentment and ratings of other's perceived contentment. This article tests these properties with respect to the popular Satisfaction With Life Scale, through a multivariate measurement model with latent change and method effects. The model is fitted to two independent data sets (N = 597 and N = 964), and it is found twice that 4 items are suitable to compute a meaningful composite difference score. It is concluded that the above-average effect is a systematic multivariate phenomenon that can be assessed by the difference of 2 manifest, absolute evaluation scores.

The above-average effect is a robust finding in social comparative judgments (Chambers & Windschitl, 2004): Individuals routinely rate themselves higher than others on a variety of attributes and capabilities. In particular, they tend to see themselves as happier and more content with their life than their peers, to a paradoxical extent: For example, Lykken and Tellegen (1996) report that 86% of respondents placed themselves in the upper 35% contentment group.

However robust this result, Klar and Giladi (1999) draw attention in an influential article to a strong bias in the measurement of comparative contentment: When asked 'How happy are you *compared to your peers ?*' people tend to interpret the question to mean simply 'How happy are you?' People thus fail to answer the comparative happiness question, and give instead an absolute evaluation of their own happiness. This result arguably speaks for the use of the so-called *indirect technique* in social comparative judgments (Chambers & Windschitl, 2004). Applied to comparative happiness, this technique would amount to first asking for an evaluation of one's own happiness (a composite manifest variable S, as in *Self*), and then asking for an absolute evaluation of one's peers perceived happiness (a composite manifest variable P, as in *Peers*). The above-average effect would then be assessed as the positive mean of the difference $S - P$.

This procedure, however, begs an important psychometric question, which is alluded to in Klar and Giladi's (1999) conclusions (p. 594):

> Is it possible to compare one's own internal state with that of others? For generations, philosophers have argued that because of unequal access to the inner state of self and others, there is no viable way for a person to compare his or her own state of contentment—or indeed, any other internal state—with the corresponding state of another person. [...] Given this unbridgeable gap between one's self-knowledge and one's assumed knowledge of others, it should not be surprising that when participants are confronted with a comparative question, the tendency is to refer solely to their own state rather than to the difference between themselves and their peers. The question as such is, in fact, unanswerable.

From that perspective, the indirect technique is no improvement on the direct technique because, by definition, the individual differences measured by S and the individual differences measured by P are incommensurable. Due to the 'unbridgeable gap' between the knowledge that produces S and the knowledge that produces P, the (in)directness of the measurement is a moot point. An investigator who asks for a direct comparison rating, in the hope of directly obtaining $S - P$, will end up with what is really an S rating; but an investigator who asks for S and P separately, in order to compute $S - P$, will find it impossible to make any legitimate interpretation of this difference variable.

Is it still possible to save the indirect technique from this apparently devastating critique? In other words, is the above-average effect measurable at all? To be able to consider the above-average effect as a measurable phenomenon, it is necessary to formulate a measurement model that defines formally what is measured by $S - P$, and then to provide evidence for the empirical suitability of this measurement model. This is the goal of the present article.[1]

Overview

We choose as our application example the ubiquitous Satisfaction With Life Scale (SWLS; Diener, Emmons, Larsen, & Griffin, 1985). The SWLS consists of five items to which

[1] Although such an endeavor would be beyond the scope of this article, we note that our psychometric approach can also be used to investigate specific claims about comparative judgment processes, such as the various accounts reviewed in Chambers and Windschitl (2004).

respondents answer on a 7-point scale (*strongly disagree* to *strongly agree*):

1. In most ways my life is close to my ideal;
2. The conditions of my life are excellent;
3. So far I have gotten the important things I want in life;
4. If I could live my life over, I would change almost nothing;
5. I am satisfied with my life.

Here, the variable S is the sum of the five ratings that respondents give when they answer the five items from their own perspective. The variable P, in contrast, is the sum of the five ratings that respondents give when they answer from the perspective of their peers, that is, when they try to imagine the modal answer that their peers would give to each item. The question we wish to answer is whether it makes any sense to compute the difference $S - P$. On the surface, the answer from Classical Test Theory is relatively simple: The difference $S - P$ makes sense to the extent that the variable $S - P$ can be decomposed as the sum of a true-change variable and a measurement error. The key is then to identify sufficient conditions for this decomposability to hold, to implement these conditions in a testable model, and to test this model against empirical data.

This is the strategy we will follow in the rest of this article. In the next section, we define the true-score measurement model of the composite variables S and P. Within this true-score model, we identify a sufficient set of three conditions for $S - P$ to be decomposable as the sum of a true-change variable and a measurement error. We then implement these three conditions in a structural equation model, which we finally test against two data sets.

Defining the True-Score Model

Let us denote by S_i the manifest variable defined by the answers given to item i. Within the true-score model, S_i is defined as the sum of a referential true-score variable, a method variable, an intercept, and an error variable (see Eid, 2000; Eid, Lischetzke, Nussbeck, & Trierweiler, 2003):

$$S_i \equiv \lambda_{Si} f_S + (M_{Si} + \alpha_{Si}) + E_{Si} \tag{1}$$

Equation 1 expresses the assumption that any manifest variable S_i in the Satisfaction With Life Scale reflects the same underlying, true-score variable f_S.[2] Equation 1 allows for three sources of intra-individual differences in the responses given to the five items of the scale, which may convey slightly different semantic nuances. The loading λ_{Si} is there to allow a difference of one point on a given item not to be subjectively equivalent to a difference of one point on another item. The (centered) residual method effect M_{Si} and the intercept α_{Si} are there to allow deviations in the subjective anchor of each item, because strongly disagreeing with a given item may not be exactly comparable to strongly disagreeing with another item.

In order to make the true-score measurement model identifiable, it is necessary to choose a reference item, that is, to assume that one of the S_i is such that $\lambda_{Si} = 1$, $\alpha_{Si} = 0$, and M_{Si} is fixed to zero. Since item 5 ('I am satisfied with my life') summarizes the construct quite well, it is chosen as the reference item. The true-score variable f_S is thus defined as the conditional expectations of the observed variable S_5 given p_U, where p_U denotes the person variable defined on the set U of the population—formally, $f_S \equiv E(S_5|p_U)$.

Now, the composite manifest variable S is expressed as the sum of the five manifest variables S_i:

$$S = (1 + \sum_{i=1}^{4} \lambda_{Si}) \cdot f_S + \sum_{i=1}^{4} (M_{Si} + \alpha_{Si}) + \sum_{i=1}^{5} E_{Si}. \tag{2}$$

Applying the same reasoning to variable P, we arrive at the following expression of the difference variable $S - P$:

$$\begin{aligned} S - P \quad = \quad & (1 + \sum_{i=1}^{4} \lambda_{Si}) \cdot f_S - (1 + \sum_{i=1}^{4} \lambda_{Pi}) \cdot f_P \\ & + \sum_{i=1}^{4} (M_{Si} - M_{Pi} + \alpha_{Si} - \alpha_{Pi}) \\ & + \sum_{i=1}^{5} (E_{Si} - E_{Pi}). \end{aligned} \tag{3}$$

Our task is now to identify a set of sufficient conditions for the right part of Equation 3 to be reduced to the sum of a true-change variable and a measurement error.

Identifying the Sufficient Conditions

From the perspective of Classical Test Theory, the difference variable $S - P$ makes sense to the extent that it can be expressed as:

$$S - P = \lambda(f_S - f_P) + E, \tag{4}$$

where $\lambda > 0$, f_S and f_P are the true-score variables defined in the previous section, and E is a latent error variable.

Three conditions are jointly sufficient for Equation 3 to reduce to Equation 4: measurement invariance of loadings ($\forall i,\ \lambda_{Si} = \lambda_{Pi}$), measurement invariance of intercepts ($\forall i,\ \alpha_{Si} = \alpha_{Pi}$), and equality of method factors ($\forall i,\ M_{Si} = M_{Pi}$). Indeed, it can be easily checked that these three conditions reduce Equation 3 to:

$$S - P = (1 + \sum_{i=1}^{4} \lambda_{Si}) \cdot (f_S - f_P) + \sum_{i=1}^{5} (E_{Si} - E_{Pi}), \tag{5}$$

which reduces in turn to Equation 4 by letting $\lambda = 1 + \sum_{i=1}^{4} \lambda_{Si}$ and $E = \sum_{i=1}^{5} (E_{Si} - E_{Pi})$. Thus, the difference variable $S - P$ makes sense as a difference true-score variable as soon as the three conditions we have identified jointly hold. The question then becomes whether it is a plausible assumption that they do. To answer this question, we will now implement the three conditions in a structural equation model that we can test against empirical data.

[2] Note that f_S is a psychometric object rather than a psychological attribute, and should not be hastily identified with satisfaction with life itself (Borsboom, Mellenbergh, & van Heerden, 2003; Zumbo & Rupp, 2004).

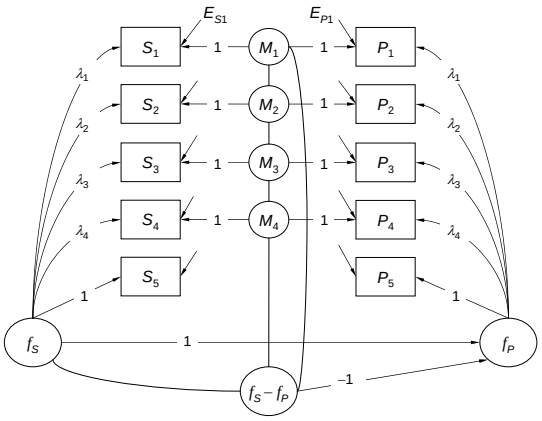

Figure 1. Path diagram of the multiple-indicator latent change model with correlated minus-1 methods. Factors M_i are the methods; factor f_S represents individual differences in self-assessed life satisfaction; factor f_P represents individual differences in the perceived life satisfaction of one's peers; factor $f_S - f_P$ represent the latent differences. Covariances are represented by curved or straight lines. M_1, M_2, M_3, M_4, and $f_S - f_P$ are allowed to covary with each other.

Implementing the Sufficient Conditions in a Testable Structural Equation Model

The structural equation model depicted in Figure 1 implements a true-score measurement model of S and P in which the three conditions we have identified jointly hold.

Drawing on Steyer, Eid, and Schwenkmezger (1997) and Raykov (1999), the difference factor $f_S - f_P$ is defined according to the tautological equation $f_P = f_S - (f_S - f_P)$. It has no residual variance and is freely correlated to the factor f_S. Drawing on Eid (2000) and Eid et al. (2003), item 5 'I am satisfied with my life' is chosen as the reference item, and the corresponding manifest variables S_5 and P_5 have no method factor. Therefore, the true variances corresponding to the variables S_5 and P_5 are identified with the factors f_S and f_P, respectively. Furthermore, fixing the loadings λ_{S5} and λ_{P5} at 1 establishes the metric of the factors as that of the reference true-score variables.

The two conditions of measurement invariance (loadings, intercepts) are specified by equating the loadings λ_{Si} and λ_{Pi}, as well as the intercepts α_{Si} and α_{Pi}. The condition of equality of the method factors is specified by the use of a method factor M_i common to the manifest variables S_i and P_i. Additional assumptions (that are usual in structural equation modeling) are that the error variables do not correlate with each others neither with other independent variables in the model.[3]

Testing the Structural Equation Model

Our structural equation model was tested against two data sets. The first set of data was collected through an Internet survey, the second set was collected through a traditional paper-and-pencil survey.

Internet survey

Method. The task was displayed on a web page. An invitation to take part to the study was sent by email to a starting list of individuals, who were invited to forward it at will. A total of 597 complete sets of answers were obtained and analyzed. For each item in the SWLS, participants first gave their own answer, then gave the answer they thought of as the one used the most frequently by other people. (All scales used the standard 7 points labeled *strongly disagree, disagree, slightly disagree, neither, slightly agree, agree,* and *strongly agree*.) The two scales appeared side by side on the screen. The order in which the items appeared on the screen was randomized for each respondent. Participants were not asked for any personal information such as name, nationality, age or gender.

The data were fitted to the measurement model with all the relevant constraints using the MLR estimator as implemented in Mplus (Muthén & Muthén, 2004). Model modification indices were used to detect the need for relaxing some equality constraints. The exact Mplus commands are available on demand from the corresponding author.

Results. The initial model fitted the data rather poorly, $\chi^2(df = 28,\ N = 597) = 131.11$, $p < .0001$, RMSEA = 0.079. The modification indices suggested to relax equality of the intercepts α_{S2} and α_{P2}. Doing this improved the fit substantially, $\chi^2(df = 27,\ N = 597) = 52.59$, $p = .002$, RMSEA = 0.040. Although the chi-square remains highly significant, the RMSEA estimate suggests a close yet not strong fit, and we used the estimated model as a workable approximation of the data. The estimated mean and variance of the difference factor $D_f = f_S - f_P$ are respectively $\hat{\mu}(D_f) = 0.555$, $SE = 0.062$ and $\hat{\text{var}}(D_f) = 1.821$, $SE = 0.207$, suggesting an above-average effect of standardized size $d = 0.41$ (d is obtained by dividing the mean difference by the standard deviation of the difference factor).

Partial invariance confines to the intercepts $\hat{\alpha}_{S2} = 1.214$, $SE = 0.250$ and $\hat{\alpha}_{P2} = 0.673$, $SE = 0.215$. This result suggests that the mean difference between the variables S_2 and P_2 is higher than predicted by the systematic comparative effect. This bias is related to the item 'The condi-

[3] A remarkable feature of the measurement model is that the variances of f_S and f_P are not necessarily equal. Indeed, letting $D_f = f_S - f_P$:

$$\begin{aligned} \text{var}(f_P) &= \text{var}(f_S - D_f) \\ &= \text{var}(f_S) + \text{var}(D_f) - 2 \cdot \text{cov}(f_S, D_f) \end{aligned}$$

Thus, for the variances of f_S and f_P to be equal, it would require that $\text{var}(f_S - f_P) = 2 \cdot \text{cov}(f_S, D_f)$, which is rather unlikely in practice.

tions of my life are excellent'. It may thus be that the use of the superlative *excellent* can distort the measurement of the above-average effect.

Because applied researchers rely on composite scores, it is of interest to assess the reliability and the raw effect size captured by the unbiased (i.e., *sans* item 2) composite difference score

$$\Delta = S_1 + S_3 + S_4 + S_5 - P_1 - P_3 - P_4 - P_5.$$

A nice feature of the measurement model is that the method factors cancel out when computing the raw difference scores. The raw composite effect size is $d = 1.958/\sqrt{30.862} \simeq 0.35$. The classical formula of the reliability coefficient ω (e.g., McDonald, 1999, p. 89) can be extended to the case of the composite difference in the following way:

$$\omega_\Delta = \frac{(\lambda_1 + \lambda_3 + \lambda_4 + 1)^2 \text{var}(D_f)}{(\lambda_1 + \lambda_3 + \lambda_4 + 1)^2 \text{var}(D_f) + E}$$

where E denotes the sum of the error variances. Here, $\hat{\omega}_\Delta \simeq .82$. The raw composite difference score thus appears to reliably capture the above average effect. It is noteworthy that if the comparative effect was constant across subjects (that is, if there was little individual variance in the above-average effect), the D_f factor would have no variance and reliability would not be defined (Raykov, 2001).

Paper and pencil survey

We were concerned that we could not control the conditions of measurement in the Internet survey, and that we had no way to prevent respondents to take part to the survey several times. We thus conducted a traditional paper-and-pencil survey in addition to the Internet survey.

Methods. Participants were 964 adult French volunteers (equal proportions of men and women, mean age = 33.3, $SD = 13.8$ for women, mean age = 30.1, $SD = 12.1$ for men). Volunteers were recruited by undergraduate psychology students, who were instructed to carefully explain the meaning of the P_i questions (the details of the recruitment procedure are available in Bonnefon & Villejoubert, 2006). The sample included a large proportion of students (39%), but the remaining 61% came from practically all professional perspectives (including 7% unemployed).

Instructions were almost the same as in the Internet survey. For each item in the SWLS, participants first gave their own answer, then gave the answer they thought was given on average by other people of their same generation, gender, and social group (a more precise definition than in the Internet survey). The two scales were presented side by side. Two counterbalanced orders of presentation were used for the five items, to reduce potential autoregressive effects (Vautier, Mullet, & Jmel, 2004).

Results. Eleven respondents failed to answer all questions, and were removed from the analyses. The model with all constraints fitted the data rather closely, $\chi^2(df = 28,\ N = 953) = 46.52$, $p = .015$, RMSEA = 0.026. However, the modification indices suggested to remove the same equality constraint as in the Internet sample, that is, $\alpha_{S2} = \alpha_{P2}$, yielding very strong fit, $\chi^2(df = 27,\ N = 953) = 24.96$, $p = .58$, RMSEA = 0.000.

The estimated mean and variance of the difference factor are respectively $\hat{\mu}(D_f) = 0.736$, $SE = 0.045$ and $\hat{\text{var}}(D_f) = 1.393$, $SE = 0.101$, suggesting an above-average effect of standardized size $d = 0.62$. Partial invariance confines to the intercepts $\hat{\alpha}_{S2} = 1.281$, $SE = 0.203$ and $\hat{\alpha}_{P2} = 1.028$, $SE = 0.172$. This result suggests again that the mean difference between the variables S_2 and P_2 is higher than predicted by the systematic comparative effect. The bias is nevertheless smaller than in the Internet sample. The raw effect size captured by the unbiased composite difference score Δ is $d = 2.494/\sqrt{22.067} \simeq 0.53$, and the composite reliability is again high, $\hat{\omega}_\Delta \simeq .77$.

Discussion

As demonstrated by Klar and Giladi (1999), the above-average effect in life satisfaction cannot be directly assessed through a comparative question. When asked for a rating of comparative happiness, respondents only give a rating of their own personal happiness. An intuitive way out of this difficulty is to use the so-called indirect measurement technique. This technique amounts to using two ratings instead of just one: one rating for personal happiness, and one rating for other's perceived happiness; and to use the difference of these two ratings as a measure of comparative happiness. If several items are used, as in the Satisfaction With Life Scale, the indirect technique amounts to computing a raw composite variable $S - P$, where S is the composite variable from judgments of personal happiness, and P is the composite variable from judgments of others' perceived happiness. The mean and variance of $S - P$ can then be estimated, and a standardized effect size can be computed for the above-average effect.

However, this statistic is a meaningful estimate of the average comparative effect *only if* the variable $S - P$ can be defined as a meaningful measurement variable. In terms of a true-score model, the difference $S - P$ should reflect a true-score difference as formulated in Equation 4. (Some readers could object that because the present study rests on the untested assumption of interval scale variables, our findings cannot be used as a proof for the psychometric meaningfulness of the difference $S - P$. We discuss that technical point in the Appendix to this article.) Starting from a latent change model with method effects, we demonstrated that three conditions are jointly sufficient to arrive at a meaningful psychometric formulation of $S - P$: measurement invariance of the loadings, measurement invariance of the intercepts, and equality of the method factors. We implemented these three properties in a testable structural equation model, which we applied to judgments collected with the popular Satisfaction With Life Scale (Diener et al., 1985). Results repeatedly sug-

gested that the three conditions plausibly held, except for the invariance of intercepts associated to the item 'The conditions of my life are excellent,' which can be removed from the composite score in order to insure measurement invariance.

Overall, the above-average effect was shown to act as a structural change. Our modeling allowed us to detect partial measurement invariance (Byrne, Shavelson, & Muthén, 1989) of the intercepts, and then to remove the faulty manifest variables from the formula of the composite difference variable $S - P$, to arrive at an unbiased estimate Δ of the latent difference score. The reliability of Δ was reasonably high (.82 in the Internet survey, .77 in the paper and pencil survey) and allowed to assess above-average effects of moderate size (0.35 in the Internet survey, 0.53 in the paper and pencil survey). Considering the numerous invariance constraints in the model (4 equalities for loadings, 4 equalities on the intercepts), these findings provide strong support to the use of the indirect technique in the measurement of the above-average effect.

Indeed, the indirect technique was already known to bypass a number of potential biases in the assessment of the above-average effect (Chambers & Windschitl, 2004; Klar & Giladi, 1999). Yet, there was no evidence until now that the manifest composite difference score $S - P$ could yield a reliable, meaningful, unbiased estimation of a true-score difference. Now that we have provided evidence to that effect, and a method for detecting items that might compromise the reliability of the estimation, we do not see any counterindication to the use of the indirect technique, applied to the raw, observed composite scores.

References

Bonnefon, J. F., & Villejoubert, G. (2006). Tactful or doubtful? expectations of politeness explain the severity bias in the interpretation of probability phrases. *Psychological Science*, *17*, 747–751.

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

Byrne, B. M., Shavelson, R. J., & Muthén, B. O. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin*, *105*, 456–466.

Chambers, J. R., & Windschitl, P. D. (2004). Biases in social comparative judgments: The role of nonmotivated factors in above-average and comparative-optimism effects. *Psychological Bulletin*, *130*, 813–838.

Diener, E., Emmons, R. A., Larsen, R. J., & Griffin, S. (1985). The Satisfaction With Life Scale. *Journal of Personality Assessment*, *49*, 71–76.

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

Eid, M., Lischetzke, T., Nussbeck, F. W., & Trierweiler, L. (2003). Separating trait effects from trait-specific method effects in multitrait-multimethod models: A multiple-indicator CT-C($M-1$) model. *Psychological Methods*, *8*, 38–60.

Fischer, G. H. (1987). Applying the principles of specific objectivity and of generalizability to the measurement of change. *Psychometrika*, *52*, 565–587.

Fischer, G. H., & Ponocny, I. (1994). An extention of the Partial Credit Model with an application of the measurement of change. *Psychometrika*, *59*, 177–192.

Klar, Y., & Giladi, E. (1999). Are most people happier than their peers, or are they just happy? *Personality and Social Psychology Bulletin*, *25*, 586–595.

Lykken, D., & Tellegen, A. (1996). Happiness is a stochastic phenomenon. *Psychological Science*, *7*, 186–189.

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

Muthén, B. O., & Asparouhov, T. (2002). *Latent variable analysis with categorical outcomes: Multiple-group and growth modeling in Mplus.* Mplus web notes: No. 4. Los Angeles: Muthén & Muthén.

Muthén, L. K., & Muthén, B. O. (2004). *Mplus 3.11*. Computer Program. Los Angeles: Authors.

Raykov, T. (1999). Are simple change scores obsolete? An approach to studying correlates and predictors of change. *Applied Psychological Measurement*, *23*, 120–126.

Raykov, T. (2001). On the use and utility of the reliability coefficient in social and behavioral sciences. *Quality and Quantity*, *35*, 253–263.

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. *Psychological Bulletin*, *114*, 552–566.

Steyer, R., Eid, M., & Schwenkmezger, P. (1997). Modeling true intraindividual change: True change as a latent variable. *Methods of Psychological Research Online*, *2*, 21–33.

Takane, Y., & de Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika*, *52*, 393–408.

Vautier, S., Mullet, E., & Jmel, S. (2004). Assessing the structural robustness of self-rated satisfaction with life: A SEM analysis. *Social Indicators Research*, *86*, 225–238.

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.

Zumbo, B. D., & Rupp, A. A. (2004). Responsible modeling of measurement data for appropriate inferences: Important advances in reliability and validity theory. In D. Kaplan (Ed.), *The SAGE handbook of quantitative methodology for the social sciences* (pp. 73–92). Thousand Oaks, CA: Sage.

Appendix
Linear Vs. Probit Regression

Our modeling relies on linear regression to link the latent variables to the manifest variables. This assumes, for estimation reasons, that the manifest variables follow a multivariate normal distribution. This is hardly true, as the manifest variables result from 7-point Likert scales. As a consequence, our analyses must be understood as approximative accounts of the data. It could be tempting then to replace linear regression by a nonlinear function such as the probit regression in the graduated response model (Takane & de Leeuw,

1987), which is available in Mplus, or to consider the approach of Fischer (1987) and Fischer and Ponocny (1994), who addressed the issue of measurement of change in the framework of Item Response Theory. In this appendix, we wish to clarify why we find the probit approach inappropriate.

The probit formulation of the graduated response model requires the use of underlying continuous outcome variables, the variances of which are equated arbitrarily (see Muthén & Asparouhov, 2002). However, systematic comparative effects generally involve homologous continuous outcome variables with different variances (Zimmerman & Williams, 1998). If the homologous outcome variables have different variances, then standardizing the underlying variables and constraining the invariance of the homologous thresholds can be misleading. Suppose that two homologous outcome variables have different variances, and that the homologous thresholds are equal; standardizing the variances will violate the thresholds equalities, and equating the homologous standardized thresholds would result in modeling heterogeneous unstandardized thresholds. Moreover, admitting that the outcome variables have equal variances, equating the homologous loadings to test the equality of the measurement units would be suitable only if the homologous error variances were equal, which is a strong assumption.

In sum, the equality constraints on the loadings and the thresholds that are usual in the context of multigroup designs (Reise, Widaman, & Pugh, 1993) are suitable to test measurement invariance only if it can be assumed that the variances of the homologous continuous variables are equal, and that homologous error variances are equal. It happens that the hypothesis of equal variances of the homologous manifest variables can be rejected with respects to both our data sets. In the Internet survey, the model specifying appropriate equality constraints on the variances fitted the data badly, $\chi^2(N = 597, 5) = 212.57$, $p < .0001$, RMSEA = 0.264. This was also true for the paper and pencil survey, $\chi^2(df = 5, N = 953) = 267.09$, $p < .0001$, RMSEA = 0.235. For these reasons, in spite of the obvious limits of the linear approach to model discrete and bounded scores, it seems more appropriate here than the use of a probit regression with arbitrarily standardized continuous variables. The fact that this modeling is not entirely satisfactory is, however, an incitement to replace usual Likert scores by ratings that would yield unbounded and continuous measurements.