# Do Others Judge My Humor Style as I Do?

# Self-Other Agreement and Construct Validity of the Humor Styles Questionnaire

Sonja Heintz

University of Zurich, Department of Psychology, Personality and Assessment

Binzmuehlestrasse 14/7, 8050 Zurich, Switzerland

Phone: +41 44 635 7574

E-mail: s.heintz@psychologie.uzh.ch

This manuscript was published as

Heintz, S. (2019). Do others judge my humor style as I do? Self-other agreement and construct validity of the Humor Styles Questionnaire. *European Journal of Psychological Assessment*, *35*, 625–632. https://doi.org/10.1027/1015-5759/a000440

# Abstract

Humor research has intensified in psychology over the last two decades, with the Humor Styles Questionnaire (HSQ) being the most prevalent measure. Still, the construct validity of its four scales (affiliative, self-enhancing, aggressive, and self-defeating) has not received univocal support. The present study uses a multitrait-multimethod approach to test the self-other agreement of the four HSQ scales with 202 targets and two knowledgeable informants per target. Employing a multilevel multiple-indicator correlated trait-correlated (method-1) (ML-CT-C[M-1]) model informed on the construct validity of the HSQ. Discriminant validities were sufficient for all scales. Convergent validity was supported for three of the four HSQ scales, except for the self-defeating scale. Similarly, the overlaps of the self- and other-reported HSQ scales with maladaptive personality as external criteria converged for all HSQ scales except for the self-defeating scale. Taken together, the present findings suggest that the self-defeating scale does not measure the maladaptive humor style it is supposed to measure.

Keywords: Humor Styles Questionnaire; self-other agreement; construct validity; multitrait-multimethod analysis; ML-CT-C(M-1) model

# Do others judge my humor style as I do? Investigating the self-other agreement and construct validity of the Humor Styles Questionnaire with the ML-CT-C(M-1) model

## Introduction

Humor is studied in many areas of psychology. Investigating self-other agreement as one aspect of construct validity is especially important for humor measures, as humor is largely a social phenomenon. Thus, others should judge our humor similarly as we do, and deviations between the two perceptions could potentially have undesirable consequences (like failing at achieving social support or offending others with one's humor). The present study focuses on the Humor Styles Questionnaire (HSQ; Martin, Puhlik-Doris, Larsen, Gray, & Weir, 2003), as it is the most frequently used instrument to assess individual differences in the sense of humor and as validity evidence for the HSQ is scarce and partially conflicting. Investigating the self-other agreement should yield an in-depth picture of the construct validity and the social reality of the HSQ, revealing whether the scales actually measure what they are supposed to measure (i.e., different humor styles).

### Multitrait-Multimethod Approach

Two aspects of construct validity are relevant for the present study (see Campbell & Fiske, 1959): Convergent validity, the extent to which self-other agreement of the HSQ scales is established, and discriminant validity, the extent to which the HSQ scales can be distinguished from one another. Both aspects of construct validity can be simultaneously tested in a multitrait-multimethod approach, which separates the variance due to trait, method, and measurement error. Recent multitrait-multimethod approaches based on structural equation models overcome the limitations of investigating observable correlations between the traits and methods by disentangling these different sources of variance (for overviews, see Eid, Lischetzke, Nussbeck, & Trierweiler, 2003; Eid et al., 2008).

Specifically, the multilevel multiple-indicator correlated trait-correlated (method-1) (ML-CT-C[M-1]) model was proposed as a fruitful option for multitrait-multimethod data involving structurally different methods (like self- and other-reports) and interchangeable methods (like raters; Carretero Dios, Eid, & Ruch, 2011). The multilevel portion of the model contains at level 1 the raters, who are supposed to be drawn from a larger pool of possible raters of the target (and are thus interchangeable). As several raters judge the same target, they are nested within the target. Targets, including their self-reports and the average other-report, are modeled at level 2. The CT-C(M-1) portion of the model contains one latent factor for each trait in the confirmatory factor analysis. One method factor is dropped from the model, as one reference method is chosen (the self-reports in the present study) against which the other non-reference methods are compared (the other-reports in the present study). At level-2, the trait factors thus represent the variance common to the self-reported HSQ scales, and the method factors represent the deviations of the average other-report from the trait factor (i.e., the residual).

The ML-CT-C(M-1) model provides several advantages (see Carretero Dios et al., 2011; Eid et al., 2003, 2008). First, measurement error can be separated from method and trait variance, as originally envisioned by Campbell and Fiske (1959); in other words, the true-score variances of traits and methods are investigated. Second, modeling the multilevel structure of raters nested in targets allows separating potential biases of individual raters at level-1 (unique method specificity) from the potential biases of the average other-report at level 2 (common method specificity). Third, trait-specific method effects can be estimated, as method biases might often vary across the different traits. Previous applications empirically supported the usefulness of the ML-CT-C(M-1) model (e.g., Carretero Dios et al., 2011). The present study applies the model for the first time to self- and other-reports of the HSQ.

## Construct Validity of the HSQ

The HSQ measures four trait-like humor styles, defined as “the interpersonal and intrapsychic functions that humor is made to serve by individuals in their everyday lives, and particularly those functions that are considered most relevant to psychosocial well-being” (Martin et al., 2003, p. 51). Two humor styles are supposed to be adaptive (affiliative and self-enhancing), and two are supposed to be potentially maladaptive (aggressive and self-defeating).

Four studies thus far compared self and other-reports of the HSQ scales. Martin et al. (2003) investigated in the construction article the agreement of students' self-reported HSQ scales and reports by their dating partners on one item of each HSQ scale (*N* = 165). Convergent validities were small to medium for the four HSQ scales, and discriminant validities were close to zero except for a positive correlation between the affiliative and self-enhancing scales. Findlay and Jones (2005) investigated the agreement between self-, partner-, and friend-reports of the HSQ scales in 80 students and found "moderate agreement between the three judgments" (p. 204). Cann, Zapata, and Davis (2011) investigated the agreement between self-reports and partner-reports in sample of 82 couples, and they found small to medium convergent validities for each HSQ scale. Finally, Zeigler-Hill, Besser, and Jett (2013) investigated the self-peer agreement of the HSQ in 257 students. The peer-reports of three items for each HSQ scale were aggregated across several peers, resulting in small (self-enhancing) to medium convergent validities (affiliative, aggressive, and self-defeating). Again, small to medium positive relationships emerged between the affiliative and self-enhancing scales. Overall, these studies yielded small to medium convergent validities of the HSQ scales, and lower discriminant validities among the affiliative and self-enhancing scales. However, this might have been due to suboptimal methodologies used; that is, no structural equation modeling approaches were employed, and the other-reports were either measured with short scales or they were not averaged across raters.

Additionally, two recent studies investigated the construct validity of the HSQ scales using self-reports. Comparing the HSQ scales to the definitions and the construct descriptions of the humor styles by employing a single-indicator CT-C(M-1) model, large agreements were found between the three sources for all HSQ scales, and a small overlap was found between the HSQ self-enhancing scale and its definition (Heintz & Ruch, 2015). The largest overlap (i.e., lowest discriminant validity) was found between affiliative and self-enhancing. Another approach (Ruch &

Heintz, 2017) involved experimentally separating the construct-relevant content (i.e., humor) and construct-irrelevant context within the HSQ items. Correlating these manipulated versions with the original HSQ scales showed that the affiliative and self-enhancing scales were mainly determined by their construct-relevant content, while the self-defeating scale was mainly determined by the construct-irrelevant context entailed in its items (e.g., going overboard).

Overall, these studies yielded partially conflicting and preliminary results, making further investigations of the construct validity and especially the self-other agreement of the HSQ scales necessary. As the HSQ is usually studied in the context of psychosocial well-being, the present study additionally investigates whether the previous self-report findings can be replicated with the other-reports of the HSQ in terms of maladaptive personality (see Zeigler-Hill, McCabe, & Vrabel, 2016). This would further support the social reality and relevance of the HSQ for our psychosocial well-being, in addition to its construct validity.

## Materials and Methods

### Participants

**Targets.** Overall, 468 participants agreed to take part in the study, of which 306 had complete and usable scores (65.4%). Only participants who for whom two other-reports were available were considered in the final sample, which was the case for 202 (72.3% female, 27.7% male) participants. They were on average 26.37 years old (*SD* = 11.00, range 18–75 years) and they were primarily Swiss (84.2%), German (9.9%), or had another nationality (5.9%). Two-thirds of them were college or university students (59.9%), 17.3% had passed tertiary education, 16.3% had a high school diploma, and 6.4% completed an apprenticeship.

**Raters.** Overall, 489 raters (knowledgeable informants) agreed to take part in the study, of whom 404 (82.6%) provided complete and usable scores (56.2% female, 43.8% male). Overall, there were 202 dyads (two independent raters for each target). They were on average 33.87 years old (*SD* = 14.87, range 18–72 years). Most raters indicated that they were a friend (38.9%) or a relative (child, sibling, or parent; 38.9%) of the target, 17.1% were romantic partners, and 5.2%

indicated other types of relationships (e.g., work colleague). The raters were very familiar with the targets: The average relationship length was 14.04 years (*SD* = 9.89, range 1–60), and raters on average indicated that they knew the person very well (*M* = 6.32, *SD* = 0.82, range 3–7) on a Likert-type scale from *very little knowledge* (1) to *excellent knowledge about the person* (7).[1]

## Instruments

**Humor Styles Questionnaire (HSQ; Martin et al., 2003; German version by Ruch & Heintz, 2016).** The HSQ measures four humor styles with eight items each. Sample items are "I usually don't laugh or joke around much with other people" (affiliative), "If I am feeling depressed, I can usually cheer myself up with humor" (self-enhancing), "If someone makes a mistake, I will often tease them about it" (aggressive), and "I don't often say funny things to put myself down" (self-defeating). The instrument employs a seven-point Likert scale from *totally disagree* (1) to *totally agree* (7). Internal consistencies (McDonald's omega) were sufficient (from .75 for aggressive to .88 for self-enhancing).

**Humor Styles Questionnaire – Other-Report Form (adapted for this study).** The instrument consists of the same 32 items as the HSQ, yet they were rephrased to refer to another person instead of oneself. Specifically, the possessive pronoun "my" was replaced by "her/his" (adapted to the target's gender), and the pronoun "I" was replaced by the targets' first name. Sample items are "[Name] usually doesn't laugh or joke around much with other people" (affiliative), "If [Name] is feeling depressed, he/she can usually cheer himself/herself up with humor" (self-enhancing), "If someone makes a mistake, [Name] will often tease them about it" (aggressive), and

[1] The rank correlations between the two measures of familiarity (relationship length and knowledge) and accuracy (computed as the squared Euclidian distances between the self- and other-reports, separate for each HSQ scale) were small and mostly nonsignificant ($-.13 \leq \rho \leq .15$). Thus, differences in familiarity across the different target-rater dyads did not influence the accuracy of the other-reports.

"[Name] doesn't often say funny things to put himself/herself down" (self-defeating). It employs the same seven-point Likert scale as the HSQ. McDonald's omega of the other-reports (aggregated across two raters) was good (from .82 for aggressive to .89 for affiliative).

**Personality Inventory for DSM-5—Brief Form—Adult (PID-5-BF; Krueger, Derringer, Markon, Watson, & Skodol, 2013; German version by Zimmermann, Krueger, Markon, & Leising, 2012).** The PID-5-BF assesses five maladaptive personality factors (negative affectivity, detachment, antagonism, disinhibition, and psychoticism) described in the Diagnostic and Statistical Manual of Mental Disorders (American Psychiatric Association, 2013) with five items each. Sample items are "I worry about almost everything" (negative affectivity), "I often feel like nothing I do really matters" (detachment), "I use people to get what I want" (antagonism), "People would describe me as reckless" (disinhibition), "I have seen things that weren't really there" (psychoticism). Items were answered on a four-point scale ranging from *very false or often false* (0) to *very true or often true* (3). McDonald's omega was sufficient (ranging from .67 for antagonism to .73 for psychoticism).

## Procedure

The study was conducted online (www.unipark.info) and in accord with the local ethical guidelines. All participants (targets and raters) declared their online informed consent. After completing the self-reports, targets were provided with a link to the HSQ Other-Report Form, which they should forward to at least two people who knew them well. The link included a unique identifier number to match the other-reports anonymously with the self-reports. Other variables were assessed that are not relevant for the present study.

## Statistical Analyses

The ML-CT-C(M-1) model was computed with MPlus 5.1 (Muthén & Muthén, 1998–2008). The model was run separately for each pairwise combination of HSQ scales to reduce the complexity of the model to achieve an optimal ratio of free parameters to the (level-2) sample size.[2] Four two-item parcels were created for each HSQ scale using a balancing approach. A power of .82 with an alpha level of .05 was achieved for two-tailed correlations of .20.

## Results

### Observed Self-Other and Inter-Rater Agreement

The four HSQ scales exhibited on average large inter-rater agreements ($ICC_{\text{mean}}$ = .51, range .44–.62) and self-other agreements ($r_{\text{mean}}$ = .48, range .31–.59). (Table S1 in Electronic Supplement Material 1 additionally shows the means and standard deviations and the agreement for each HSQ item and scale.) Additionally, item-profile agreement was computed by averaging the correlation between each self-other and each rater dyad across the eight HSQ items of each scale. Inter-rater item-profile agreements were large for affiliative ($r_{\text{mean}}$ = .76) and self-defeating ($r_{\text{mean}}$ = .47), medium to large for aggressive ($r_{\text{mean}}$ = .38), and medium for self-enhancing ($r_{\text{mean}}$ = .30). Self-other item-profile agreements were medium for self-enhancing ($r_{\text{mean}}$ = .31 and self-defeating ($r_{\text{mean}}$ = .30), and medium to large for affiliative ($r_{\text{mean}}$ = .41) and aggressive ($r_{\text{mean}}$ = .35).

### ML-CT-C(M-1) Model

Table 1 shows the fit indices of the six estimated ML-CT-C(M-1) models (one for each pairwise combination of HSQ scales).

As shown in Table 1, most models showed an overall good to accepting fit. The SRMR for level-1 was always good, while the SRMR of level-2 was unsatisfactory. Figure 1 illustrates the ML-CT-C(M-1) model including the standardized factor loadings (The means, unstandardized factor loadings, residual variances, and reliabilities of the model indicators are shown in Table S2 in Electronic Supplement Material 1.)

[2] Results were highly similar when the four traits were included in one model simultaneously.

Table 1

*Fit Indices of the Six Multilevel Multiple Indicator Correlated Trait-Correlated (Method-1) Models*

| Model | $\chi^2$ | $\chi^2/df$ | CFI | TLI | RMSEA | SRMR 1 | SRMR 2 |
|---|---|---|---|---|---|---|---|
| AF-SE[a] | 295.05 | 2.06 | .94 | .94 | .05 | .03 | .11 |
| AF-AG[a] | 217.58 | 1.52 | .96 | .96 | .04 | .03 | .12 |
| AF-SD[a] | 277.66 | 1.94 | .94 | .94 | .05 | .05 | .10 |
| SE-AG[b] | 253.33 | 1.77 | .94 | .94 | .04 | .03 | .12 |
| SE-SD | 295.65 | 2.07 | .93 | .93 | .05 | .04 | .11 |
| AG-SD[b] | 228.89 | 1.60 | .95 | .95 | .04 | .03 | .11 |

*Note. df* = 143. SRMR 1 = SRMR at level 1, SRMR 2 = SRMR at level 2.

[a] The models in which affiliative was included were used to determine convergent validity; results only differed slightly in the other models.

[b] The latent variable covariance matrix was not positive definite for these models.

Focusing on the results at level-2 (*N* = 202), Figure 1 shows that the loadings between the observed indicators of the self-reports (“Self 1”–“Self 4”) and the latent trait factors were large for the four HSQ scales (ranging from .59 to .85). Squaring these loadings yields the level-2 reliabilities of the self-reports; they ranged from .35 for aggressive to .66 for self-enhancing (*Mdn* = .58). The latent correlations between the average other-reports (across raters), modeled as a latent variable at level-2, and the latent trait factors were high for three of the HSQ scales (.77–.90) and medium to large for self-defeating (.46). Squaring these latent correlations yields the convergent validities at level 2, which are interpreted as the amount of variance in the average other-reports that can be explained by the self-reports. These convergent validities were high for affiliative (for which 65.6% of the variance of the average other-reports was explained by the self-reports), self-enhancing (59.3%), and aggressive (81.0%), and lower for self-defeating (21.2%).

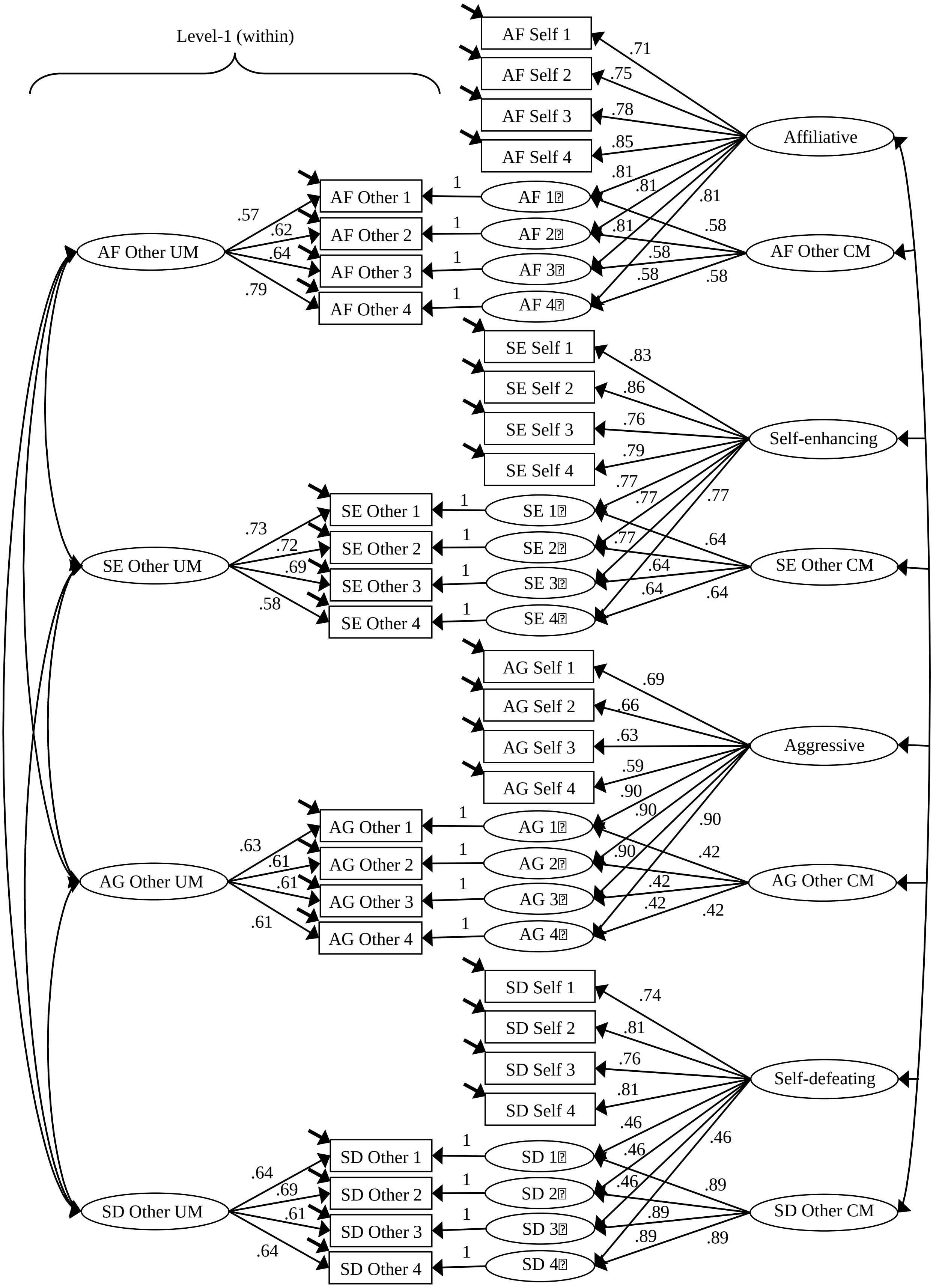

*Figure 1*. Multilevel multiple-indicator correlated-trait-correlated (method-1) model with standardized loadings (CM = common method, UM = unique method).

The common method factor loadings between the average other-reports and the common method factors (“Other CM”) represent the extent to which the average other-report cannot be explained by the self-reports (i.e., the residuals). These loadings ranged from .42–.89. Squaring these common method factor loadings represents the amount of variance specific to the average other-report (i.e., not shared with self-reports). They were smaller than the convergent validities at level-2, ranging from 17.6% (aggressive) to 41.0% (self-enhancing), with the exception of self-defeating (79.2%).

The observed other-reports (“Other 1”–“Other 4”) at level-1 (*N* = 404) can be subdivided into four different variance components: The variance due to the unique method factor (level-1), the common method factor (level-2), the trait factor (level-2) and measurement error (level-1). The loadings between the observed other-reports and the unique method factors (“Other UM”) capture the deviation of one rater from the average other-report of the same target. Squaring these loadings yields the level-1 reliabilities, which were mostly low (range from .32–.62, *Mdn* = .40).

Using the formulas by Eid et al. (2008; see Carretero Dios et al., 2011, for another exemplification of these formulas in the context of the ML-CT-C[M-1] model), the observed other-reports can be separated into different coefficients based on the four variance components: Consistency (convergent validities between self-reports and other-reports that consider the average other-report and the individual view of each rater), common method specificity (common view of both raters that is not shared with the target), unique method specificity (individual view of each rater that is not shared with the target and the other rater), and reliability (amount of true score variance in comparison to observed variance). For each HSQ scale, the largest proportion of variance was due to the individual view of each rater (unique method specificity): 49% (affiliative), 57% (self-defeating), 65% (aggressive), and 69% (self-enhancing). The second largest proportion of the variance was explained by the consistencies (convergent validities) for affiliative (33%), self-enhancing (19%), and aggressive (28%). The third largest amount of variance was explained by the common view of the raters (common method specificity): 18% for affiliative, 13% for self-enhancing, and 7% for aggressive. For self-defeating, the variance due to the common method specificity (34%) was larger than the variance due to the consistency (9%). The level-2 reliabilities of the observed indicators of the other-reports ranged from .47–.77 (*Mdn* = .53).

Discriminant validities are indexed by the intercorrelations of the trait factors in the ML-CT-C(M-1) model. (Variances and correlations of the trait and method factors are shown in Table S3 in Electronic Supplement Material 1.) The overlap ranged from 0.4% (self-enhancing and aggressive) to 22% (affiliative and self-enhancing), with a median of 6%. The second largest overlap emerged between self-enhancing and self-defeating (11%). The latent correlations among the unique and common method factors additionally showed that method effects did not generalize across each trait.

## Overlap with Maladaptive Personality

Correlations and partial correlations (controlling for gender and age) were computed between the self- and other-reported HSQ scales and the self-reported maladaptive personality factors. (Table S4 in in Electronic Supplement Material 1 shows each zero-order and partial correlation.) In line with previous findings of the self-reported HSQ, affiliative ($r_{\text{mean}}$ = -.12, $r_{p\ \text{mean}}$ = -.14) and self-enhancing (-.17/-15) correlated on average negatively with maladaptive personality, while aggressive (.22/.19) and self-defeating (.19/.19) correlated positively with it. Affiliative (-.07/-.12) and self-enhancing (-.18/-.19) remained adaptive in the other-reports, and aggressive (.10/.07) remained maladaptive. The correlations with self-defeating (-.03/-.02) were close to zero in the other-reports.

# Discussion

The present study aimed at testing the self-other agreement of the four HSQ scales, yielding information on the degree of their construct validity. Discriminant validities were high for all HSQ scales, with the largest overlap occurring between affiliative and self-enhancing (22.1% shared true-score variance), which is similar to the previous findings (Martin et al., 2003; Heintz & Ruch, 2015; Zeigler-Hill et al., 2013). The other-reported HSQ scales were mainly determined by unique method specificity (i.e., the individual view of each rater). The second largest source of variance in the other-reported affiliative, self-enhancing, and aggressive scales were the consistencies (i.e., the convergent validities of the self- and other-reports), followed by the common method specificity (i.e., the common view of the raters not shared with the target). Obtaining larger consistencies than common method specificities (ratio 1.5–4:1) supports the convergent validities of these three HSQ scales. For the self-defeating scale, the common method variance exceeded the consistencies (ratio 3.8:1). Similar convergent validities were found when considering the latent correlations at level-2: The amount of variance explained by the self-reports in the average other-reports was large, and larger than the residual for three of the HSQ scales (ratio 1.4–4.6:1), and this effect was reversed for self-defeating (ratio 3.7:1). Thus, the other-reported HSQ self-defeating scale was more strongly determined by method than by trait variance, failing to support convergent validity for this scale.

Furthermore, the relationships of the self-reported HSQ scales with maladaptive personality were replicated in the other-reports for all scales except for self-defeating (which was found to be neutral in the other-reports). Thus, the relationships to external criteria generalized across other-reports for three of the four scales, further supporting their social reality.

What are possible explanations for the low convergent validity and social reality of the self-defeating scale? As indicated by the large common method specificity found in the ML-CT-C(M-1) model and large observed inter-rater agreements, the two raters agreed on their judgments of the target's self-defeating score more than they agreed with the target's perspective. Stated differently, both raters differed systematically from the target's perspective. This finding reminds of Martin et al.'s (2003) assertion that "Although individuals who are high on this humor dimension may be seen as quite witty or amusing […], there is an element of emotional neediness, avoidance, and low self-esteem underlying their use of humor" (p. 54). According to this view, the raters might have put more emphasis on the observable humor behaviors incorporated in this scale (like letting others laugh at oneself and saying funny things that put oneself down), while they might have missed (or put less emphasis on) the underlying negativity than the targets did (like getting carried away or going overboard when showing these behaviors).

However, given that raters were very familiar with the targets, a more likely explanation might be that the construct validity and social reality of the self-defeating scale is indeed impaired. This implies that the self-defeating scale does not adequately measure the self-defeating humor style and might thus lead to misleading conclusions (see also Ruch & Heintz, 2017). It seems likely that the self-defeating scale is less a measure of humor than of negative self-evaluation, potentially changing the meaning of the construct and the interpretations of existing findings on the scale. In other words, low self-esteem would not only underlie this humor style (as suggested by Martin et al., 2003), but it might actually be the active ingredient that causes the previously established negative correlations between the self-defeating scale and psychosocial well-being. The other-reports obtained in the present study might have been less influenced by this bias: They showed that the self-defeating scale was not negative in terms of maladaptive personality and also positively related to the self-enhancing scale. The same results were found in studies that focused on the humor entailed in this scale (Ruch & Heintz, 2013, 2017).

## Limitations

First, investigating the self-other agreement and construct validity of the HSQ in other languages, cultures, and samples with different demographic backgrounds would be advisable to test to which extent the present findings can be generalized. Second, level-2 reliabilities were lower than .60 for the indicators of the aggressive scale, and level-1 reliabilities were low in general. As the ML-CT-C(M-1) model separates this error variance from the trait and method variance, these lower reliabilities have likely not influenced the present findings substantially. Still, using indicators that consist of more than two items would be desirable in future studies to enhance reliability (as for example shown in Carretero Dios et al., 2011; Koch et al., 2015). Similarly, using at least three instead of two raters per target would be desirable. Third, the present study is purely correlational, and future studies need to go beyond correlations to study the cause and effect in the relationships between the HSQ and psychosocial well-being (e.g., by employing experimental designs or survey testing techniques like cognitive interviews).

## Suggestions for Future Research

The present findings highlight the importance of comprehensively investigating the construct validity of humor measures, best during the process of test construction. Testing whether a measure and/or a construct needs further revision at an early stage (i.e., before a measure becomes widely used in research and/or application) would likely avoid interpretation problems. If problems are detected at a later stage (as is the case for the self-defeating scale), either the scale or the construct or both need to be revised, which is a task for future studies.

Employing the ML-CT-C(M-1) model yields information about the extent with which different assessment methods converge with the "golden standard" and which sources of variance contribute to non-convergence. Misalignments can then be resolved by specifically adapting the relevant aspects of the construct and/or measure. Additionally incorporating relevant criteria can show whether the supposed nomological network generalizes across different methods. Overall, the ML-CT-C(M-1) model can be recommended as a useful framework for conducting multitrait-multimethod analyses for psychometric measures if (a) structurally different and interchangeable methods are investigated, (b) a reference method (i.e., a golden standard) can be rationally determined, and (c) some methods are nested within other methods. Recent extensions of the ML-CT-C(M-1) model also allow incorporating structurally different methods at level-1 (Koch, Schultze, Burrus, Roberts, & Eid, 2015) and longitudinal designs (Koch, Schultze, Eid, & Geiser, 2014), making the model suitable for wide range of methods (e.g., different types of raters, observable behaviors such as facial displays of emotion and laughter, or experience sampling data).

## Summary and Conclusions

The present study investigated the self-other agreement of the HSQ as an aspect of construct validity, using a large sample of targets and raters (two knowledge informants). Employing the ML-

CT-C(M-1) model allowed separating the different variance components entailed in the multitrait-multimethod data. Convergent validity was supported and correlations with external criteria (maladaptive personality) were replicated for all HSQ scales except for self-defeating, while discriminant validity was sufficient for all HSQ scales. The current findings thus suggest that the self-defeating scale might not measure what is supposed to measure (i.e., a maladaptive humor style), potentially changing the interpretation of past findings and cautioning against a further usage of this scale.

## Acknowledgments

I would like to thank Claudia Hürzeler for her valuable help in constructing and pre-testing the other-report form of the HSQ, in setting up the online study, and in recruiting and managing participants. I would also like to thank Willibald Ruch for his feedback on the study design and on earlier versions of this manuscript, Hugo Carretero Dios for his advice on the ML-CT-C(M-1) model, and Nadine Eggimann for her comments on earlier versions of this manuscript.

## Electronic Supplementary Material

ESM 1. Tables S1–S4 (SupplementaryTables.pdf). The tables show descriptive statistics, inter-rater agreement and self-other agreement of the HSQ items and scales and the unstandardized loadings, correlations, variances, and reliabilities of the ML-CT-C(M-1) models as well as the correlations of the HSQ with maladaptive personality.

ESM 2. MPlus codes for the multilevel multiple-indicator correlated trait-correlated (method-1) models (MPlusCodes.txt).

## References

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders (5th ed.)*. Arlington, VA: American Psychiatric Association.

Cann, A., Zapata, C.L., & Davis, H.B. (2011). Humor style and relationship satisfaction in dating couples: Perceived versus self-reported humor styles as predictors of satisfaction. *Humor: International Journal of Humor Research, 24*, 1–20. http://dx.doi.org/10.1515/humr.2011.001

Campbell, D.T., & Fiske, D.W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*, 81–105. http://dx.doi.org/10.1037/h0046016

Carretero Dios, H., Eid, M., & Ruch, W. (2011). Analyzing multitrait-mulitmethod data with multilevel confirmatory factor analysis: An application to the validation of the State-Trait Cheerfulness Inventory. *Journal of Research in Personality, 45*, 153–164. http://dx.doi.org/10.1016/j.jrp.2010.12.007

Eid, M., Lischetzke, T., Nussbeck, F.W., & Trierweiler, L.I. (2003). Separating trait specific effects from trait-specific method effects in multitrait-multimethod models: A multiple-indicator CT-C(M-1) model. *Psychological Methods, 8*, 38–69. http://dx.doi.org/10.1037/1082-989X.8.1.38

Eid, M., Nussbeck, F.W., Geiser, C., Cole, D.A., Gollwitzer, M., & Lischetzke, T. (2008). Structural equation modeling of multitrait-multimethod data: Different models for different types of methods. *Psychological Methods, 13*, 230–253. http://dx.doi.org/10.1037/a0013219

Findlay, B.M., & Jones, C. (2005). I think I've got a sense of humour, but do others think so? *Australian Journal of Psychology, 57*, S1, 204. http://dx.doi.org/10.1080/00049530600940010

Heintz, S., & Ruch, W. (2015). An examination of the convergence between the conceptualization and the measurement of humor styles: A study of the construct validity of the Humor Styles

Questionnaire. *Humor: International Journal of Humor Research, 28*, 611–633. https://dx.doi.org/10.1515/humor-2015-0095

Koch, T., Schultze, M., Burrus, J., Roberts, R.D., & Eid, M. (2015). A multilevel CFA-MTMM model for nested structurally different methods. *Journal of Educational and Behavioral Statistics, 40*, 477–510. https://dx.doi.org/10.3102/1076998615606109

Koch, T., Schultze, M., Eid, M., & Geiser, C. (2014). A longitudinal multilevel CFA-MTMM model for interchangeable and structurally different methods. *Frontiers in Psychology, 5*:311. https://dx.doi.org/10.3389/fpsyg.2014.00311

Krueger, R.F., Derringer, J., Markon, K.E., Watson, D., & Skodol, A.E. (2013). *The personality inventory for DSM-5—brief form (PID-5-BF)—adult.* Arlington, VA: American Psychiatric Association.

Martin, R.A., Puhlik-Doris, P., Larsen, G., Gray, J., & Weir, K. (2003). Individual differences in uses of humor and their relation to psychological well-being: Development of the Humor Styles Questionnaire. *Journal of Research in Personality, 37*, 48–75. http://dx.doi.org/10.1016/S0092-6566(02)00534-2

Muthén, L.K., & Muthén, B.O. (1998–2008). *Mplus 5.1. Statistical analysis with latent variables. User's guide.* Los Angeles, CA: Muthén and Muthén.

Ruch, W., & Heintz, S. (2013). Humour styles, personality, and psychological well-being: What's humour got to do with it? *European Journal of Humour Research, 1*, 1–24. https://doi.org/10.7592/EJHR2013.1.4.ruch

Ruch, W., & Heintz, S. (2016). The German version of the Humor Styles Questionnaire: Psychometric properties and overlap with other styles of humor. *Europe's Journal of Psychology, 12*, 434–455. http://dx.doi.org/10.5964/ejop.v12i3.1116

Ruch, W., & Heintz, S. (2017). Experimentally manipulating items informs on the (limited) construct and criterion validity of the Humor Styles Questionnaire. *Frontiers in Psychology, 8*:616. http://dx.doi.org/10.3389/fpsyg.2017.00616

Zeigler-Hill, V., Besser, A., & Jett, S.E. (2013). Laughing at the looking glass: Does humor style serve as an interpersonal signal? *Evolutionary Psychology, 11*, 201–226. http://dx.doi.org/10.1177/147470491301100118

Zeigler-Hill, V., McCabe, G.A., & Vrabel, J.K. (2016). The dark side of humor: DSM-5 pathological personality traits and humor styles. *Europe's Journal of Psychology, 12*, 363–376. http://dx.doi.org/10.5964/ejop.v12i3.1109

Zimmermann, J., Krueger, R.F., Markon, K.E., & Leising, D. (2012). *German translation of the personality inventory for DSM-5—brief form (PID-5-BF)—adult.* Unpublished manuscript, University of Kassel, Germany.