Self-Compassion as a Facet of Neuroticism? A Reply to the Comments of Neff, Tóth-Király, and Colosimo (2018)

MATTIS GEIGER[1], STEFAN PFATTHEICHER[1,2]*, JOHANNA HARTUNG[1], SELINA WEISS[1], SIMON SCHINDLER[3] and OLIVER WILHELM[1]

[1]*Ulm University, Germany*
[2]*Aarhus University, Denmark*
[3]*Kassel University, Germany*

Abstract: In this paper, we respond to comments by Neff et al. (2018) made about our finding that the negative dimensions of self-compassion were redundant with facets of neuroticism (rs ≥ 0.85; Pfattheicher et al., 2017) and not incrementally valid. We first provide epistemological guidance for establishing psychological constructs, namely, three hurdles that new constructs must pass: theoretically and empirically sound measurement, discriminant validity, and incremental validity—and then apply these guidelines to the self-compassion scale. We then outline that the critique of Neff et al. (2018) is contestable. We question their decisions concerning data-analytic methods that help them to circumvent instead of passing the outlined hurdles. In a reanalysis of the data provided by Neff et al. (2018), we point to several conceptual and psychometric problems and conclude that self-compassion does not overcome the outlined hurdles. Instead, we show that our initial critique of the self-compassion scale holds and that its dimensions are best considered facets of neuroticism. © 2018 European Association of Personality Psychology

Key words: jangle fallacy; neuroticism; rebuttal; self-compassion

INTRODUCTION

Science is a cumulative journey. In this regard, we appreciate the response of Neff, Tóth-Király, and Colosimo (2018); (hereby collectively referred to as 'NTC') to our critical analysis of the concept 'self-compassion' and the self-compassion scale (SCS) (Pfattheicher, Geiger, Hartung, Weiss, & Schindler, 2017). In this rebuttal, we will show that several major conclusions of NTC's response are problematic from both a conceptual and an empirical stance. Through reanalysis, we demonstrate that self-compassion can be seen as a facet of neuroticism and that it has essentially no incremental validity above established personality dimensions. We hence repeat, expand, and elaborate our initial critique and conclude by pointing out future research directions that might proof to be more fruitful.

Self-compassion is defined as the tendency to be understanding and kind to oneself when confronted with negative experiences (Neff, 2003). Self-compassion incorporates three positive and three negative dimensions. The three positive dimensions are self-kindness (i.e. being supportive, caring, and understanding to oneself), common humanity (i.e. all humans are connected, and personal failures are part of the human existence), and mindfulness (i.e. accepting the present [negative] moment in a non-judging and non-reacting manner). The three negative dimensions that reflect the absence of self-compassion consist of critical self-judgment (i.e. harsh self-criticism), isolation (i.e. loneliness during negative experiences), and over-identification (i.e. rumination about thoughts and emotions). These six dimensions are argued to be crucial when individuals experience negative affect (cf. Barnard & Curry, 2011; Neff, 2003; Neff & Dahm, 2015; Reyes, 2012). Together, they create the SCS (Neff, 2003), an instrument designed for and used to measure self-compassion (Barnard & Curry, 2011; Neff & Dahm, 2015).

When proposing a new psychological construct such as self-compassion, three hurdles need to be overcome in order to infer that the new construct provides utility (e.g. Wilhelm, Kaltwasser, & Hildebrandt, 2017). The first hurdle requires us to prove that a construct has been measured adequately. This requirement includes demonstrating that a sound measurement model can account for the data: proposed factors account for a non-trivial share of variance in indicators assigned to them, the pattern of loadings is meaningful, competing models show poorer fit, etc. (Borsboom, Mellenbergh, & van Heerden, 2004). The second hurdle is that a new construct must not be a linear function of established constructs. For example, constructs such as Need for Cognition can be shown to be completely redundant with similar concepts like Intellectual Engagement (von Stumm & Ackerman, 2013). The third and quintessential hurdle is that a new construct

*Correspondence to: Stefan Pfattheicher, Department of Psychology and Behavioral Sciences, Aarhus University, 8000 Aarhus C, Denmark.
E-mail: sp@psy.au.dk

must predict meaningful outcomes, not only directly, but *incrementally* over and above established constructs.

When reviewing the literature on self-compassion, we found that these hurdles were not properly overcome. Concerning the first hurdle, in our initial comment to Neff et al., we stressed that SCS items can hardly be distinguished from neuroticism items. We doubt that researchers unfamiliar with specific items but knowledgeable about the construct definitions can reliably assign items such as (i) 'When something upsets me I get carried away with my feelings'; (ii) 'I get overwhelmed by emotions'; (iii) 'I judge myself more harshly than others do'; or (iv) 'I'm disapproving and judgmental about my own flaws and inadequacies'[1] to either neuroticism facets as, for example, provided in the NEO-PI-R (Costa Jr. & McCrae, 1992) or the SCS scales from Neff et al. (see Muris, van den Broek, Otgaar, Oudenhoven, & Lennartz, in press, for a similar point). Additionally, measurement models should reflect essential distinctions proposed on a conceptual level. Although proponents of the SCS stress the relevance of positive and negative dimensions, this distinction is not reflected in advocated measurement models (see also Neff et al., in press).

We are somewhat sceptical concerning the second hurdle, because the contents of the SCS are so akin to neuroticism that we doubt that a clear separation is possible. Prior research has not shown that self-compassion traits are distinct from neuroticism and its facets (Pfattheicher et al., 2017). Instead, we suggest that self-compassionate traits should be considered traits below the established and much broader trait dimension of neuroticism.

Whether the SCS overcomes the third hurdle—incremental validity of new constructs—has hardly been tested before now. Obviously, incremental validity is only tested meaningfully if the variables over which validity should be established are not strawmen. A meaningful incremental validation should be taken into consideration when deciding whether conceptually similar but more established predictors are available within a field. Apparently, SCS dimensions seem to be highly redundant with various facets of neuroticism (Pfattheicher et al., 2017). Additionally, in many applied fields, there is a set of established predictors for relevant criteria. A first step towards incremental validity should be to show that SCS dimensions predict criteria over and above facets of neuroticism and other established predictors in a field. Obviously, *ceteris paribus* criteria should reflect meaningful outcomes. Unfortunately, for the SCS, most validation studies are limited to other self-reports that might themselves be deemed as facets of established constructs (see Table 1).

Given these concerns, we concluded that self-compassion did not pass the test for new constructs, and we summarized our concerns in a recent paper (Pfattheicher et al., 2017). NTC reanalysed our data and provided two additional data sets. NTC also conclude that 'although self-compassion overlaps with neuroticism, the two constructs are distinct' (p. 2).

Indeed, self-compassion keeps being applied in hundreds of studies. In these studies, self-compassion is considered an established construct and used as outcome for interventions. In what follows, we argue that this situation is scientifically questionable. In order to evaluate whether or not the application of self-compassion and the SCS is premature, we applied systematic tests for the three hurdles outlined earlier. We will discuss the SCS relative to these hurdles one by one.

Hurdle 1: Sound measurements procedures

The task of measurement models in psychological assessment is typically to provide a parsimonious account of individual differences in a measure. To this end, competing measurement models can be tested against each other or against saturated or null models. Apart from the fit of such models to the data, there are additional aspects that need consideration when accepting and interpreting a measurement model. The factors we specify in measurement models are meant to account for individual differences in the indicators regressed on them. If the loadings on a factor are very small, vary substantially in size, or—worst case scenario—vary in sign, odds are that the factor is useless as explanatory variable (for a discussion on problematic and poorly defined factors, e.g. in bifactor models, see Eid, Geiser, Koch, & Heene, 2017).

Similarly, the factor should have non-trivial variance. Given the factor variance and the unstandardized loadings, it is easy to compute how much change in a manifest variable is associated with an increase in one standard deviation of the factor. If the factor has next to no variance—which is often the result of incoherent and poor loading patterns—odds are that the factor does not determine relevant changes in manifest variables. Indeed, NTC repeatedly argue that their general factor accounts for about 95% of the reliable variance captured with the SCS score. This argument implies that the remaining six factors account for less than 1% of the systematic variance on average. Closer inspection of the loading patterns confirms this worry. Unfortunately, we have many more worries about the measurement models proposed by Neff et al. in various publications, as well as in NTC. Here is a collection of some of these concerns:

First, whereas Neff and colleagues argue in favour of models with a general factor of self-compassion, this factor structure has been heavily debated recently (Brenner, Heath, Vogel, & Credé, 2017; López et al., 2015; Muris, 2016; Muris, Meesters, Pierik, & de Kock, 2016; Muris, Otgaar, & Petrocchi, 2016; Muris & Petrocchi, 2017; Neff, 2003; Neff, 2016; Neff et al., in press; Neff, Whittaker, & Karl, 2017).

Second, the ESEM (Asparouhov & Muthén, 2009, exploratory structural equation modelling) approach favoured by NTC is not the right choice to establish a measurement model. Typical approaches for examining the factor structure of a new measure are testing competing measurement models derived from sound theoretical considerations or exploring the factorial space with exploratory methods, such as exploratory factor analysis or ESEM, and then testing an exploratory derived structure with confirmatory factor analysis (CFA). In ESEM, all indicators are allowed to load on all factors, and factors are usually allowed to correlate. Obviously,

---

[1](i) is taken from the mindfulness subscale of the SCS; (ii) is taken from the IPIP item pool (Goldberg, 1999; item code H950) NEO-IPIP scale neuroticism dimension, facet vulnerability; (iii) is taken from the IPIP item pool (item code X252), subscale calmness; and (iv) is taken from the self-judgment subscale of the SCS.

Table 1. A content analysis of outcome measures of Study 3 in Neff et al. (2018) and their overlap with established constructs of personality as measured with the NEO-PI-R (Costa Jr. & McCrae, 1992), IPIP items (Goldberg, 1999), and HEXACO-PI items (Lee & Ashton, 2004)

Construct of outcome measure	Outcome measure example items (item code)	Related personality construct	Related personality example items (item code)
Optimism (Life Orientation Test; LOT; Scheier, Carver, & Bridges, 1994)	'If something can go wrong for me, it will'. (LOT3)	Anxiety (neuroticism)	'I often worry about things that might go wrong'. (NEO-PI-R151)
	'I'm always optimistic about my future'. (LOT4)	Anxiety (neuroticism)	'I'm seldom apprehensive about the future'. (NEO-PI-R121)
	'Overall, I expect more good things to happen to me than bad'. (LOT10)	Anxiety (neuroticism)	'I am not a worrier'. (NEO-PI-R001)
Curiosity (Curiosity and Exploration Inventory, CEI; Kashdan, Rose, & Fincham, 2004)	'I frequently find myself looking for new opportunities to grow as a person (e.g. information, people, and resources)'. (CEI3)	Openness to actions (openness)	'I prefer to spend my time in familiar surroundings'. [R] (NEO-PI-R138)
	'I am not the type of person who probes deeply into new situations or things'. (CEI4)	Intellect	'Will not probe deeply into a subject'. (IPIP H1392, X54)
	'Everywhere I go, I am out looking for new things or experiences'. (CEI7)	Experience-seeking	'Love to learn new things'. (IPIP H1265) 'Try out new things'. (IPIP H1269)
Personal Initiative (Personal Growth and Initiative Scale; PI; Robitschek, 1998)	'I have a good sense of where I am headed in life'. (PI2)	Conscientiousness	'Do not plan ahead'. [R] (IPIP E104)
	'I have a specific action plan to help me reach my goals'. (PI6)	Hope/optimism	'Have no plan for my life five years from now'. (IPIP V206)
	'If I want to change something in my life, I initiate the transition process'. (PI3)	Achievement striving (conscientiousness)	'When I start a self-improvement program, I usually let it slide after a few days'. [R] (NEO-PI-R80)
Self-esteem (Rosenberg Self-Esteem Scale; RSE; Rosenberg, 1979)	'I feel that I have a number of good qualities'. (RSE3)	Competence (conscientiousness)	'I'm a very competent person'. (NEO-PI-R185)
	'I feel that I'm a person of worth'. (RSE7)	Depression (neuroticism)	'Sometimes I feel completely worthless'. (NEO-PO-R41)
Psychological Wellbeing (Psychological Well-being Scales; PWB; Ryff & Keyes, 1995) — Autonomy	'I am able to do things as well as most other people'. (RSE4)	Modesty (honesty-humility)	'Don't think that I'm better than other people'. [R] (HEXACO-PI Q175)
	'My decisions are not usually influenced by what everyone else is doing'. (PWB7)	Negative valence	'Conform to others' opinions'. [R] (IPIP H1056)
Environmental Mastery	'In general, I feel I am in charge of the situation in which I live'. (PWB2)	Assertiveness (extraversion)	'Feel that many things are outside my control'. [R] (IPIP Q237)
Personal Growth	'I do not enjoy being in new situations that require me to change my old familiar ways of doing things'. (PWB27)	Self-consciousness (neuroticism)	'Am comfortable in unfamiliar situations'. [R] (IPIP X197)
Positive Relations	'Most people see me as loving and affectionate'. (PWB4)	Warmth (extraversion)	'I'm known as a warm and friendly person'. (NEO-PI-R62)
Purpose in Life	'I live life one day at a time and don't really think about the future'. (PWB5)	Order (conscientiousness)	'I would rather keep my options open than plan everything in advance'. [R] (NEO-PI-R10)
Self-acceptance	'In general, I feel confident about myself'. (PWB12)	Modesty (agreeableness)	'I have a very high opinion of myself'. (NEO-PI-R144)

Note: [R] indicates that this item must be understood reversed to match the corresponding item.

ESEM can hardly be deemed a parsimonious model. Instead, ESEM is tailored to help accomplishing acceptable fit for self-report measures. In both the CFA and ESEM cases, an essential part of evaluating how well a construct overcomes hurdle 1 is to test competing models against each other.

Third, consistency is required in model comparison. NTC did not consider parsimony when choosing ESEM over CFA, but they should have. They then reject a better-fitting model [Model 4(b) with one general SCS factor versus Model 5(b) which is better fitting with two general SCS factors], although this model is superior in fit [Study 1: $\chi^2(7) = 77$, $p < 0.001$; Study 2: $\chi^2(7) = 47$, $p < 0.001$].[2] However, the better model is rejected, because of one unreliable factor that has nine non-significant loadings (out of 13). We must assume that this factor distortion is an artefact produced by over-parametrization with seven loadings per item necessary for ESEM. This assumption is supported by exploring the parallel regular CFA solution (Model 5a) that has no such problems in defining this factor. Instead, all its loadings are of large magnitude (Study 1: $\lambda = 0.71$–0.90; Study 2: $\lambda = 0.70$–0.80).

Finally, for all of the models NTC accept, there are problematic factor loadings and subsequent distortions. For instance, consider NTC's NEO-PI-R model: many main loadings are either hovering around zero or vary in sign. For example, five out of eight self-consciousness items have either negative loadings or at least one cross-loading that is larger than the main loading. With three out of six facets having major definition problems, the validity of the higher order neuroticism model introduced by NTC and all factor scores derived from this model are of questionable validity. We conclude that factors from such models have dubious quality. This result also calls the decision favouring the ESEM models into question.

We therefore conclude that confirmatory decisions about the measurement model of the SCS should be based on comparing regular CFA models rather than ESEM models. With this in mind, NTC's decision on the measurement model 4(b) does not overcome the first hurdle and instead other model specifications should be preferred. As Pfattheicher et al. (2017) have argued, the SCS should be modelled with two general factors and as a regular CFA model.

Hurdle 2: Discriminant validity from established constructs

Conceptual overlap and high correlations of manifest SCS and neuroticism scores clearly indicate that there is substantial overlap between the constructs. As neuroticism is the more established and much broader construct, self-compassion must demonstrate that it is distinct from neuroticism (and not vice versa). As Pfattheicher et al. (2017) pointed out, there was only little research among the large body of studies on self-compassion that explored basic relations of self-compassion and neuroticism. Furthermore, the

little research that was available suffered from two major methodological limitations.

First, studies relating the SCS with the Big Five before Pfattheicher et al. (2017) only used second-level scores of personality, such as total neuroticism scores. This is also true for Study 3 presented in NTC. However, the Big Five are best understood as a hierarchical model with narrower facets on the first level and very broad dimensions on the second level (Saucier & Ostendorf, 1999). The facets usually have specific variance, i.e. variance that is not accounted for by the second-order dimensions. Discriminant validity is not only achieved by discriminating a construct from an overarching dimension but from other established constructs that are related in content. Consequently, if an overarching second-order personality dimension is found to relate to a proposed new construct, it is wise to compare the new candidate with affiliated facets *jointly*.

Second, measurement error is typically substantial for self-report instruments. Unreliability for both the predictors and the predicted constructs should therefore take attenuation into account. This necessity demands latent modelling and thus large sample sizes, which were often available among previous self-compassion research. However, researchers typically still use total scores or factor scores that are always riddled by measurement error.

Pfattheicher et al. (2017) addressed these issues, for the first time in self-compassion research, by testing the discriminant validity of the SCS against *multiple facets* of neuroticism on a latent level. They found that there is a large overlap of both constructs. NTC, however, in their reanalysis (Study 1) and additional studies (Studies 2 and 3) went two steps backwards and used factor scores (Studies 1 and 2) or total scores (Study 3) and single instead of multiple predictors. Furthermore, factor scores were derived from problematic measurement models, as pointed out earlier. NTC argue for the use of factor or total scores because of model complexity. However, this argument is misguided because model complexity was driven by the use of ESEM and a model that, as also presented earlier, is not the appropriate measurement model for passing the first hurdle. As we demonstrate later, with regular CFA models, running structural equation models (SEMs) and thereby accounting for measurement error is possible in all three studies. Additionally, they restrict their regression analysis to individual predictors when they should have and could have used multiple predictors. Overall, we must therefore conclude that NTC's analysis of discriminant validity favours the self-compassion constructs inadequately and is burdened by shortcomings. We infer that the SCS did not pass the second hurdle.

Hurdle 3: Incremental validity over established constructs

Predictive and incremental validity are only interpretable if (i) both predicting and predicted constructs or outcomes are theoretically sound constructs and (ii) their measurement instruments passed the first hurdle. Predicted outcomes should be stable, sound, and meaningful, such as level of education, income, physical health, or satisfaction with life. The satisfaction with life scale (Diener, Emmons, Larsen, & Griffin, 1985),

---

[2]This χ^2-difference test was not reported by NTC but conducted by us using the Mplus 7.11 DIFFTEST function based on syntax by NTC taken from Open Science Framework (https://osf.io/6ayhs).

as chosen by Pfattheicher et al. (2017) in Study 1, currently has 20 012 citations on Google Scholar (retrieved June 4, 2018) and is among the most prevalent outcome measures in psychology. With high convergent validity (sum score correlations of $r = 0.61$ to 72, Lyubomirsky & Lepper, 1999) to the satisfaction with life scale, the happiness scale (Lyubomirsky & Lepper, 1999) can be considered a proxy to life satisfaction. Other outcome variables introduced by NTC are more problematic, because they suffer from serious measurement problems or predictor-criterion contamination. We will discuss these problems in detail in the following paragraphs.

The Difficulties in Emotion Regulation Scale (DERS; Gratz & Roemer, 2004) arguably measures difficulties in various aspects of emotion regulation. The term 'difficulties' suggests that the construct refers to a disability to regulate emotions. However, ability (or disability) constructs should be measured with tasks provoking maximal effort rather than typical behaviour (cf. Cronbach, 1949). Items such as 'I know exactly how I felt' (DERS10) or 'When I'm upset, I have difficulty controlling my behavior' (DERS31) clearly aim at ability measurement; however, they ask the test taker to respond in self-reports. We are not aware of an exception to the rule that efforts to measure ability constructs with typical behaviour approaches fail (Wilhelm, 2005; Wilhelm, Witthöft, & Schipolowski, 2010). Furthermore, the low objectivity of self-reported ability is reflected by typically zero to small correlations with objective measures of maximal ability (e.g. Jacobs & Roodenburg, 2014). Problems of the DERS go beyond the self-report of abilities. Other self-report items are worded to measure typical behaviour ('When I'm upset, I start to feel very bad about myself'; DERS34) or beliefs ('When I'm upset, I believe that my feelings are valid and important'; DERS21). Overall, content validity of the questionnaire is questionable.

Similar to emotion regulation ability is wisdom: an aspect of crystallized intelligence and an ability that should be measured via performance tests of maximal effort. However, in Study 3 in NTC, wisdom was assessed via a self-report measure (Three-dimensional Wisdom Scale; Ardelt, 2003). The scale uses items that are written like ratings of typical behaviour (e.g. 'I am hesitant about making important decisions after thinking about them'), preferences (e.g. 'I prefer just to let things happen rather than try to understand why they turned out that way'), or opinions/beliefs ('People are either good or bad'). Again, content validity of the questionnaire seems questionable to us.

Optimism (assessed with the Life Orientation Test; Scheier, Carver, & Bridges, 1994), curiosity (assessed with the Curiosity and Exploration Inventory; Kashdan, Rose, & Fincham, 2004), personal initiative (assessed with the Personal Growth and Initiative Scale; Robitschek, 1998), self-esteem (assessed with the Rosenberg Self-Esteem Scale; Rosenberg, 1979), and psychological well-being (assessed with the Psychological Well-being Scales; Ryff & Keyes, 1995) are all assessed via self-report items measuring typical behaviour. Given the content of these questionnaires, we strongly recommend to consider these measures as capturing aspects of personality that are decently represented in established models. We try to illustrate this jingle jangle issue

(Kelley, 1927) by comparing example items of these scales with closely related NEO-PI-R and IPIP items; see Table 1.

The Positive and Negative Affect Scale (Watson, Clark, & Tellegen, 1988) with the standard instruction (which we must assume was used in Study 3; as neither NTC, nor Neff, Rude, & Kirkpatrick, 2007, report details on the instruction) assesses current affective *states*. They might be considered meaningful outcomes of short-term interventions, but not stable long-term outcomes. If trait instructions are used, the adjectives can be considered adjective-based assessments of personality [specifically emotionality or neuroticism (e.g. anxiety and depression) and extraversion (e.g. positive emotions)], as done in the lexical approach underlying the Big Five (Goldberg, 1990).

Overall, among all outcome measures listed by NTC, only the assessments of satisfaction with life (Study 1) and happiness (Study 3) seem to be theoretically sound outcomes based on decent measurement instruments. Other outcomes selected by NTC in Studies 2 and 3 suffer from predictor-criterion contamination, where a measure is labelled a criterion, when it is in fact a befuddled measure of personality and thus, for the proposed models, a poor criterion. Consequently, in the following reanalysis, we will restrict ourselves to satisfaction with life and happiness when evaluating the incremental validity of the SCS over and above personality.

Furthermore, NTC's argument that a 'large and an ever-growing body of research indicates that self-compassion training increases compassionate and reduces uncompassionate behavior toward the self' is 'another important reason to retain the negative items of the SCS [as] they are crucial for measuring what changes when individuals learn to be more self-compassionate' (p. 33) is another case of predictor-criterion contamination. The aforementioned statement implies that the *outcome variable* (i.e. reduced uncompassionate responding) altered by the independent variable (i.e. self-compassion training) should be included in the general measurement of the independent variable (i.e. trait self-compassion). Applying this logic, one could similarly argue that life satisfaction and stress should be included in the measurement of self-compassion, as self-compassion training also increases life satisfaction and decreases stress (Neff & Germer, 2013).

Finally, as with discriminant validity, when conducting regression analyses, NTC did not take measurement error into account and tested only single predictors in the first step of their regressions. Consequently, the only valid test of incremental validity, using satisfaction with life or happiness as criteria, underestimates the predictive power of more established constructs and thus overestimates the incremental validity of self-compassion. We therefore conclude that the analysis presented by NTC is questionable as a test to overcome the third hurdle.

By reconsidering the three hurdles, we conclude that NTC's analyses circumvent these with questionable measurement models and unnecessarily restricted regression analyses, instead of using adequate methods that test whether the SCS is actually able to overcome the hurdles. Their models and analysis contain (systematic) measurement error due to distorted factors and the use of factor or total scores,

which deteriorates effects (e.g. the relations between facets of neuroticism and the SCS). Additionally, NTC only used single, never multiple, constructs in tests of discriminant and incremental predictive validity. Finally, most outcome measures introduced by NTC in Studies 2 and 3 do not meet proper standards of meaningful outcomes. Therefore, their conclusions regarding the extent of specific variance and predictive quality of self-compassion are arguably false positive decisions. Yet, their decisions to use factor scores and single predictors only were not necessary, because the newly presented samples by NTC offer the opportunity to evaluate self-compassion on an appropriate level. In the following section, we will present a reanalysis of all three studies based on the previously introduced criteria.

A REANALYSIS OF NTC'S DATA[3]

All reanalyses were conducted using R 3.4.3. (R Core Team, 2018) and the package lavaan (version 0.5.23; Rosseel, 2012). Mplus syntax was produced using the package MplusAutomation (Hallquist & Wiley, 2018) and run with Mplus 7.1 (Muthén & Muthén, 1998–2012). All analysis syntax is available on Open Science Framework (https://osf.io/hk53j/). Latent factor models with comparative fit index (CFI) ≥ 0.90 and root mean square error of approximation (RMSEA) < 0.08 are considered acceptable (Bentler, 1990; Steiger, 1990). We used the samples uploaded by NTC to the Open Science Framework. Significance was evaluated on an $\alpha = 0.05$ threshold. As presented earlier, in NTC's Studies 1 and 2, the two general factors solutions were superior to a single general factor solution, which is why we will only focus on the two general factors solutions.

Reanalysis of the first hurdle: Model structure

The only major analysis step not discussed so far is the use of bifactor models used by NTC, instead of higher order models used by Pfattheicher et al. (2017). Both models are very similar. There are two main differences between these models. First, in the higher order model, overarching factors account for variance in first-order factors as opposed to manifest variables for the bifactor model. Second, the unique variance of the lower order factors is represented in the residual terms of the first-order factors for the higher order model as opposed to be reflected in the nested factors themselves in case of the bifactor model. Overall, in the current case of the SCS, the higher-order model is somewhat more restricted and thus

more parsimonious, but typically less well fitting (Murray & Johnson, 2013). Yung, Thissen, and McLeod (1999) demonstrated that the models are nested and can thus be compared via the χ^2-difference test.

In all studies, a test comparing the two-factor higher-order versus the two-factor bifactor model favoured the less restricted bifactor model. Please note that the parsimonious comparative fit index (PCFI; Arbuckle, 2010) was in favour of the higher order model. A summary of fit indices and model comparisons of both models in all three studies are reported in Table 2.

The type of model that should be preferred remains up for debate. In the case of the SCS, for example, the correlation of the compassionate self-responding (CS) and reduced uncompassionate self-responding (RUS) factors, which is of major interest in the debate on higher order or general factors of self-compassion, shows almost no difference between both model types. Murray and Johnson (2013) argue that higher order models are less prone to parameter distortions due to arbitrary modelling issues, such as varying amounts of manifest variables per first-order factor or error-correlations that are not modelled. Usually, such distortions can be identified easily, and they were not an issue in any of the bifactor models presented here.

The very high second-order loadings reported by Pfattheicher et al. (2017) indicate that there is little specific variance in the first-order factors, with an exception for the factor 'Common Humanity' (CH; $R^2 = 0.624$–0.655 of variance explained by the respective higher order factor in all three studies). Also, all loadings of this specific factor were significant in the bifactor models. The other specific factors had much higher percentages of explained variance, ranging from $R^2 = 0.796$ to $R^2 = 0.955$. When modelled in bifactor models, in at least one of the three studies at a minimum of one loading of the other specific factors was non-significant, indicating definition problems and questioning the interpretation. As we repeatedly found a stable specific CH factor, we decided that this factor might also be tested in structural models. The other factors will not be modelled in further analysis, as it is questionable whether they have unique variance. In order to extend the results by NTC, we use a bifactor model with the correlated RUS and CS factors and an orthogonal CH factor in structural models. The final model is depicted in Figure 1.

The model had an acceptable fit in Study 1 [$\chi^2(294) = 1144$, $p < 0.001$; CFI = 0.915; RMSEA = 0.071; PCFI = 0.828], Study 2 [$\chi^2(294) = 1334$, $p < 0.001$; CFI = 0.902; RMSEA = 0.080; PCFI = 0.816), and Study 3 [$\chi^2(294) = 528$, $p < 0.001$; CFI = 0.851; RMSEA = 0.070; PCFI = 0.769) with an exception of the CFI in Study 3, which is below the CFI threshold. This exception is probably attributable to the small sample size of Study 3 ($N = 177$) and to the fact that the sample (undergraduate students) is less representative of the general population, which might result in variance restriction, thereby lowering correlations (Pearson, 1903) and model fit (Muthén, 1990). We also repeated the model comparison for one versus two general factors of self-compassion discussed in the previous chapter with only CH as specific bifactor. All χ^2-difference tests were significant with $\chi^2(1) = 227$–985 and all $ps < 0.001$—clearly

---

[3]We want to note that, in this chapter, we will only present results based on Maximum Likelihood (ML) estimation, whereas NTC presented results based on WLSMV estimation. Overall, the choice of ML versus WLSMV is highly debated, and studies vary largely in recommendations regarding appropriate sample sizes for WLSMV estimation (Bandalos & Leite, 2006; Liang & Yang, 2014; Moshagen & Musch, 2014; Muthén, du Toit, & Spisic, 1997; Nussbeck, Eid, & Lischetzke, 2006). However, a majority of the mentioned studies would agree that in Study 3, the sample is not sufficiently large to run stable WLSMV analyses. Furthermore, Studies 2 and 3 had missing data. In order to be able to include Study 3 and impute missing data via full information ML estimation, for our reanalysis, we decided for ML estimation. In Studies 1 and 2, WLSMV and ML estimations showed only marginal and negligible differences.

Table 2. Comparison of higher order and bifactor models with two general factors (CS and RUS) in Studies 1 to 3

Study	Model	χ^2-test	CFI	TLI	RMSEA	PCFI	χ^2-difference test	$r_{\text{CS*RUS}}$
1	Higher order	$\chi^2(292) = 959, p < 0.001$	0.933	0.926	0.063	0.838	$\chi^2(20) = 277, p < 0.001$	0.766
	Bifactor	$\chi^2(272) = 682, p < 0.001$	0.959	0.951	0.051	0.803		0.764
2	Higher order	$\chi^2(292) = 1209, p < 0.001$	0.917	0.907	0.074	0.824	$\chi^2(20) = 413, p < 0.001$	0.840
	Bifactor	$\chi^2(272) = 796, p < 0.001$	0.953	0.943	0.058	0.797		0.837
3	Higher order	$\chi^2(292) = 538, p < 0.001$	0.850	0.833	0.069	0.764	$\chi^2(20) = 86, p < 0.001$	0.607
	Bifactor	$\chi^2(272) = 452, p < 0.001$	0.890	0.869	0.061	0.745		0.600

Note: CS, compassionate self-responding; RUS, reduced uncompassionate self-responding; CFI, comparative fit index; PCFI, parsimonious comparative fit index; RMSEA, root mean square error of approximation; TLI, Tucker Lewis index.
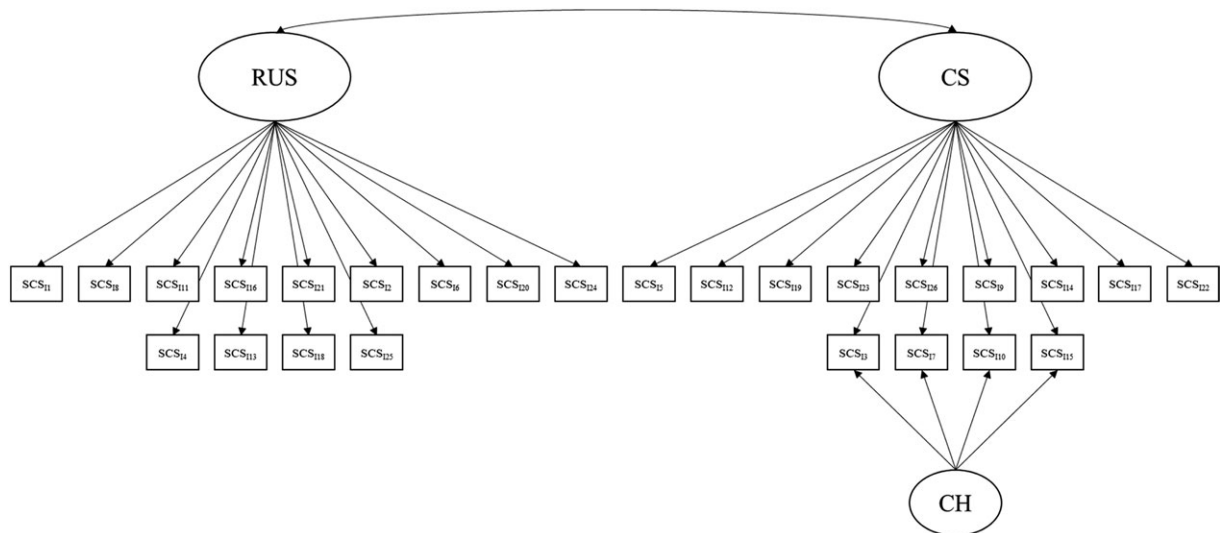

Figure 1. The measurement model of self-compassion used in tests of discriminant and incremental validity. RUS, reduced uncompassionate self-responding; CS, compassionate self-responding; SCS, self-compassion scale; CH, common humanity.

favouring the models with two general factors over the models with only one overarching general factor.

Reanalysis of the second hurdle: Discriminant validity

We used SEMs with multiple predictors to test both discriminant validity and incremental predictive validity. Facets of neuroticism were modelled based on selected items as described in Pfattheicher et al. (2017). They had reported issues of multicollinearity for Study 1. As such, we did not enter all facets of personality as predictors of the dimensions of the SCS in this reanalysis; rather, only anxiety and depression were used, as in Pfattheicher et al. (2017). Both facets, as well as the dimensions of the SCS, were modelled as latent factors and regressions were run simultaneously in a SEM. All dimensions of SCS were predicted by both facets of neuroticism. As in Pfattheicher et al. (2017), RUS was largely explained by the two facets of neuroticism with $R^2 = 0.805$. CS had more than half of its variance accounted for by both facets with $R^2 = 0.554$. Only for CH the variance accounted for was small ($R^2 = 0.040$).

We repeated the same model in Study 2 and found very similar results as in Study 1, namely, $R^2_{\text{RUS}} = 0.804$, $R^2_{\text{CS}} = 0.636$, and $R^2_{\text{CH}} = 0.042$. Overall, both reanalyses replicate the findings of Pfattheicher et al. (2017) in that RUS is

largely redundant with (facets of) neuroticism and that CS has a major overlap with neuroticism, too. New is the finding that the only consistent specific factor of the SCS, CH, is largely independent from neuroticism. Further exploration of the relation of the CH factor with other facets of personality might be advisable. A content analysis linking the CH items to items taken from the IPIP item pool (Goldberg, 1999) might be a fruitful way of starting the exploration of this specific variance. As Study 3 had only the short NEO-FFI, which is a measure of personality with only higher level dimensions and no facets, the tests of discriminant validity could not be conducted.

Reanalysis of the third hurdle: Incremental validity

When we criticized NTC's analyses methods in the previous chapter, we discussed that many outcomes introduced by NTC in Studies 2 and 3 were either unestablished constructs or not measured properly. Consequently, testing the predictive validity of any construct on these outcomes is pointless. As argued earlier, only satisfaction with life and happiness seem to be adequate criteria. As satisfaction with life has been shown to be predicted by both facets of neuroticism and facets of extraversion (Schimmack, Oishi, Furr, & Funder, 2004), we entered both dimensions (or in Study 1,

Table 3a. Evaluation of incremental validity of the SCS factors over facets of neuroticism and extraversion in predicting life satisfaction in Study 1

Step #	Predictors of life satisfaction	R^2	ΔR^2 to previous model
1	NEON1 + NEON3 + NEOE1 + NEOE3 + NEOE6	0.532	
2	NEON1 + NEON3 + NEOE1 + NEOE3 + NEOE6 + RUS	0.547	0.015
3	NEON1 + NEON3 + NEOE1 + NEOE3 + NEOE6 + RUS + CS	0.550	0.003
4	NEON1 + NEON3 + NEOE1 + NEOE3 + NEOE6 + RUS + CS + CH	0.552	0.002

Note: NEON1, anxiety; NEON3, depression; NEOE1, warmth; NEOE2, gregariousness; NEOE6, positive emotions; RUS, reduced uncompassionate self-responding; CS, compassionate self-responding; CH, common humanity; SCS, self-compassion scale.

Table 3b. Evaluation of incremental validity of the SCS factors over the factors neuroticism and extraversion in predicting happiness in Study 3

Step #	Predictors of happiness	R^2	ΔR^2 to previous model
1	NEON + NEOE	0.712	
2	NEON + NEOE + RUS	0.724	0.012
3	NEON + NEOE + RUS + CS	0.729	0.005
4	NEON + NEOE + RUS + CS + CH	0.729	0.000

Note: NEON, neuroticism; NEOE, extraversion; RUS, reduced uncompassionate self-responding; CS, compassionate self-responding; CH, common humanity; SCS, self-compassion scale.

the respective facets: anxiety, depression, warmth, gregariousness, and positive emotions; the choice of extraversion facets is based on Schimmack et al., 2004) in the first step. With satisfaction with life (Study 1) and happiness (Study 3) as the only acceptable outcomes, only these tests of incremental predictive validity will be calculated. The results are summarized in Tables 3a and 3b. The largest incremental predictive validity was of $\Delta R^2 = 0.015$ of RUS in Study 1 and $\Delta R^2 = 0.012$ of RUS in Study 3. However, all ΔR^2s can be considered negligible. Furthermore, we want to stress that in Study 3, we could only use personality dimension level predictors to happiness in the first step. Presumably,

with facet level predictors, the R^2 in step one would have been larger and the incremental validity of SCS dimensions even lower.

In sum, when using an empirically supported model of self-compassion, we replicated the findings of Pfattheicher et al. (2017). We found the negative dimensions of self-compassion (RUS) to be mostly redundant to facets of neuroticism or neuroticism in general and the positive dimension of self-compassion (CS) to be largely explained by those facets. With the available outcome measure, we found no relevant incremental predictive validity of self-compassion or the specific CH factor.
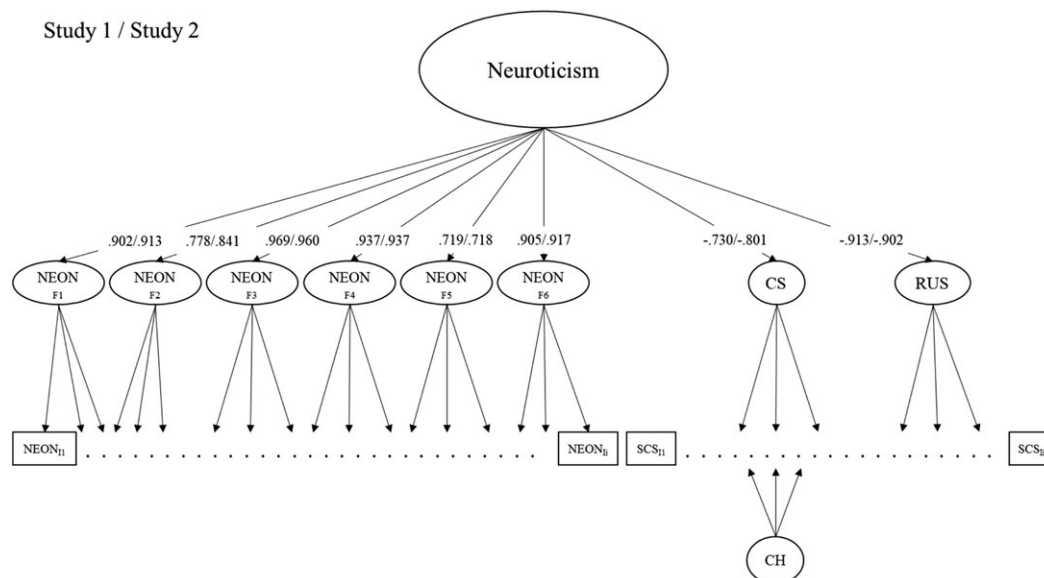

Figure 2. Schematic depiction of the self-compassion as a facet of neuroticism model. The circles NEONF1 to NEONF6 refer to the facets of neuroticism in the NEO-PI-R. Compassionate self-responding (CS), reduced uncompassionate self-responding (RUS), and common humanity (CH) refer to factors of the self-compassion scale (SCS). Numbers on arrows are standardized loadings.

ADDITIONAL EXPLORATORY ANALYSES

Given the findings replicating Pfattheicher et al. (2017), we decided to run additional analyses to explore whether RUS and CS could be subsumed as facets under neuroticism. To explore this possibility, we ran two higher order CFA models in Study 1 and Study 2. The models are schematically pictured in Figure 2, including second-level loadings. The model was specified as a higher order model of neuroticism over all first-order factors of neuroticism (with selected items as in Pfattheicher et al., 2017) and included RUS and CS as first-order factors, by adding loadings from the general neuroticism factor on both general SCS factors. CH was kept orthogonal to all factors. The SCS loadings in Study 1 ($|\lambda_{N \to RUS}| = 0.913$; $|\lambda_{N \to CS}| = 0.730$) and Study 2 ($|\lambda_{N \to RUS}| = 0.902$; $|\lambda_{N \to CS}| = 0.801$) were very high and within the range of the other loadings of neuroticism first-order facets (Study 1: $\lambda_{N \to \text{first-order factors}} = 0.719-0.969$; Study 2: $\lambda_{N \to \text{first-order factors}} = 0.718-0.960$). These results suggest that the general dimensions of the SCS are best considered facets of neuroticism.

DISCUSSION

In this paper, we repeated and elaborated on problems associated with the SCS and the underlying theory of self-compassion. The problems were ordered according to three hurdles that all new constructs should be able to overcome. With respect to the first hurdle—a theoretically and psychometrically sound construct—we had several concerns. Given the large body of literature on self-compassion using the SCS, proponents of self-compassion are tempted to deem self-compassion a sound construct. However, we illustrated that the problems with the SCS and the underlying theory are thorough and pervasive. The SCS items could be assigned to established personality constructs without much ado. The proposed ESEM measurement models are psychometrically not tenable. Conceptual terms seem to be used in a more belletristic or narrative sense. For example, NTC propose that self-compassion 'operates as a system' (NTC; p. 8), without providing an explanation what is meant by 'operates' and by 'system'. Similar problems arise with the definition of self-compassion as 'the *balance* between increased compassionate and reduced uncompassionate responding to personal struggle' (NTC; p. 3; emphasis added). Does this 'balance' suggest the joint activity of a positive and a negative self-compassion dimension? Why are there no such factors in the models specified by NTC, then? Unclear conceptualization is also an issue when focusing on the proposed dimensions of self-compassion. For example, the dimension 'mindfulness' is somewhat fuzzy in its definition (Nilsson & Kazemi, 2016), in its measurement and structure (Bergomi, Tschacher, & Kupper, 2013), and how and whether it can be trained (Creswell, 2017). For a comprehensive and recent critique of mindfulness, see Van Dam et al. (2018).

We recommend pursuing the measurement model we proposed because the fundamental issues brought forward against the model recommended by NTC do not apply to this model. Additionally, we recommend restraining from terminology such as self-compassion as operating system or a balance of positive and negative self-compassion that is seemingly inconsequential and infallible in the present context. Overall, we must conclude that the definition and measurement of self-compassion as proposed by NTC does not overcome the first hurdle.

With respect to the second hurdle, a similarly negative evaluation seems inevitable. NTC present results that indicate that self-compassion was discriminant from facets of neuroticism. However, these results are based on regression analyses that systematically accept superfluous restrictions, such as not controlling for measurement error and only using single predictors. With our reanalysis, we demonstrated that these restrictions are clearly unnecessary. Furthermore, we replicated the results from Pfattheicher et al. (2017) with the additional data, thus strengthening their conclusion that many aspects of self-compassion are redundant to neuroticism. Consequently, we conclude that self-compassion—at least the way currently captured with the SCS—also failed to overcome the second hurdle.

In higher order models, the inference for the second-level self-compassion factors is mimicked by issues with the specific factors. We found strong and significant relations of the higher order CS factor with the lower order mindfulness factor with $\lambda_{\text{Study1}} = 0.967$, $\lambda_{\text{Study2}} = 0.925$ and $\lambda_{\text{Study3}} = 0.905$ (or the low specific variance of the mindfulness factor in bifactor models). Given the weak discriminant validity of the SCS dimensions from neuroticism, it might be hypothesized that trait mindfulness, as defined and measured in the SCS, too, is not discriminant from neuroticism. Whether this hypothesis can be generalized to all mindfulness definitions and measurement instruments is an empirical question only partly addressed so far (Van Dam et al., 2018). Similar problems of discriminant validity arise with the other dimensions of self-compassion, with the exception of common humanity, which consistently had specific variance in all three studies. For the latter, we recommend further exploration, whether this dimension might be considered as distinct from personality constructs outside the neuroticism range.

Given that self-compassion still struggled to overcome the second hurdle after analysis with additional data, we decided to extend the Pfattheicher et al. (2017) analysis by testing a higher order measurement model of neuroticism that includes the general dimensions of the SCS as facets of neuroticism. With strong to very strong second-level loadings of the SCS factors that were clearly within the range of the other first-order factors, we concluded that self-compassion dimensions should—assuming one would ignore the theoretical problems discussed in regard to the first hurdle—at best be seen as facets of neuroticism. In this regard, it is up for discussion whether self-compassion dimensions are meaningful additional facets or are completely redundant with established neuroticism facets. Given the very high overlap of the negative dimension of SC with facets of neuroticism, we argue that this dimension is almost completely redundant with established neuroticism facets. However, based on our analyses, the positive dimension of the SCS captures some specific variance and might therefore be deemed a promising

candidate as future facet of neuroticism. Additionally, relations of the positive SCS domain to other facets of personality might be explored. Based on the findings of Pfattheicher et al. (2017), the most promising candidates here would be activity (extraversion), achievement striving (conscientiousness), self-discipline (conscientiousness), and trust (agreeableness). The variance of the nested common humanity factor requires further exploration, too. Two problems of such nested factors are their—ceteris paribus—limited variance and their contingency to the measurement of additional indicators. Please note that any research questions in such future studies exploring the specific variance in the SCS should be targeted, besides discriminant validity, towards incremental validity—that is whether or not the dimensions of the SCS improve prediction of relevant outcomes over and above established personality dimensions.

Finally, we were unable to find convincing evidence for incremental validity. Again, we argue that NTC circumvented the third hurdle with questionable decisions and problematic choices of outcome measures. Most tests of incremental validity presented by NTC had to be rejected, because their outcome measures suffered from predictor-criterion contamination. Furthermore, the results with an acceptable outcome (satisfaction with life) presented by NTC suffered from superfluous restrictions. We therefore decided to reanalyse them with more appropriate methods and, in contrast to NTC's analysis, found only negligible incremental validity of self-compassion over (facets of) neuroticism and extraversion. We therefore concluded that self-compassion could not overcome the third hurdle.

Conclusion

The main conclusion of this reply is that self-compassion was unable to overcome the hurdles required to be considered an established construct. Research on self-compassion that ignored its very close relation with neuroticism or did not succeed in providing clear evidence for incremental validity is suffering from the jangle fallacy (Kelley, 1927): the conclusion that self-compassion is a new construct when it is best thought of as a facet of neuroticism in need of evidence showing incremental validity. If future research continues to use self-compassion as a novel construct, it should be ensured that other facets of neuroticism are measured too, in order to allow for evaluating strong collinearity effects. This necessity also holds true for intervention studies of self-compassion, where *incremental* effects over neuroticism and its facets must be demonstrated. Otherwise, it will remain unclear whether any effect is general to neuroticism or specific to self-compassion.

Given the conceptual issues and the little specificity of SCS dimensions as facets of neuroticism, it might be advisable to reconsider the construct as a whole by redefining it and exploring the nature of any variance specific to it. This redefinition also offers the chance to get rid of conceptual ambiguities. A better understanding of it might help to further improve the measurement of neuroticism or other personality traits. However, given the negligible incremental predictive validity of these new facets and the specific CH

factor reported so far, we do not expect this endeavour to be fruitful.

Every psychological construct has its limitations. We want to make it clear that self-compassion is just a single example of the problematic practice of proposing new constructs while simultaneously making ambiguous construct definitions and ignoring psychometric standards ranging from item development to reasonable tests of different forms of validity (cf. Moshagen, Hilbig, & Zettler, in press). With our contributions, we highlighted problematic aspects of the definition and measurement of the construct self-compassion. We found clear empirical evidence against the position that this construct is distinct from neuroticism. This conclusion does not exclude the construct from the scientific field. Rather, it allows new perspectives on the construct such as a redefinition as a specific personality facet. This facet may be more malleable by the environment, situations, and interventions in contrast to more stable—and perhaps more genetically anchored—personality traits such as neuroticism (c.f. Kandler, Zimmermann, & McAdams, 2014).

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

REFERENCES

Arbuckle, J. L. (2010). *IBM SPSS Amos 19 user's guide*. Crawfordville, FL: Amos Development Corporation.

Ardelt, M. (2003). Empirical assessment of a three-dimensional wisdom scale. *Research on Aging*, *25*, 275–324. https://doi.org/10.1177/0164027503025003004.

Asparouhov, T., & Muthén, B. (2009). Exploratory structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *16*, 397–438. https://doi.org/10.1080/10705510903008204.

Bandalos, D. L., & Leite, W. (2006). The use of Monte Carlo studies in structural equation modeling research. *Structural equation modeling: A second course*, 385–426.

Barnard, L. K., & Curry, J. F. (2011). Self-compassion: Conceptualizations, correlates, & interventions. *Review of General Psychology*, *15*, 289–303. https://doi.org/10.1037/a0025754.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, *107*, 238–246. https://doi.org/10.1037/0033-2909.107.2.238.

Bergomi, C., Tschacher, W., & Kupper, Z. (2013). The assessment of mindfulness with self-report measures: Existing scales and open issues. *Mindfulness*, *4*, 191–202. https://doi.org/10.1007/s12671-012-0110-9.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review*, *111*, 1061–1071. https://doi.org/10.1037/0033-295X.111.4.1061.

Brenner, R. E., Heath, P. J., Vogel, D. L., & Credé, M. (2017). Two is more valid than one: Examining the factor structure of the self-compassion scale (SCS). *Journal of Counseling Psychology*, *64*, 696–707. https://doi.org/10.1037/cou0000211.

Costa, P. T. Jr., & McCrae, R. R. (1992). *NEO-PI-R professional manual*. Odessa, FL: Psychological Assessment Resources.

Creswell, J. D. (2017). Mindfulness interventions. *Annual Review of Psychology*, *68*, 491–516. https://doi.org/10.1146/annurev-psych-042716-051139.

Cronbach, L. J. (1949). *Essentials of psychological testing*. Oxford: Harper.

Diener, E. D., Emmons, R. A., Larsen, R. J., & Griffin, S. (1985). The satisfaction with life scale. *Journal of Personality Assessment*, 49, 71–75. https://doi.org/10.1207/s15327752jpa4901_13.

Eid, M., Geiser, C., Koch, T., & Heene, M. (2017). Anomalous results in G-factor models: Explanations and alternatives. *Psychological Methods*, 22, 541–562. https://doi.org/10.1037/met0000083.

Goldberg, L. R. (1990). An alternative "description of personality": The big-five factor structure. *Journal of Personality and Social Psychology*, 59, 1216–1229. https://doi.org/10.1037/0022-3514.59.6.1216.

Goldberg, L. R. (1999). A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models. *Personality Psychology in Europe*, 7, 7–28.

Gratz, K. L., & Roemer, L. (2004). Multidimensional assessment of emotion regulation and dysregulation: Development, factor structure, and initial validation of the difficulties in emotion regulation scale. *Journal of Psychopathology and Behavioral Assessment*, 26, 41–54. https://doi.org/10.1023/B:JOBA.0000007455.08539.94.

Hallquist, M. N., & Wiley, J. F. (2018). *MplusAutomation: An R package for facilitating large-scale latent variable analyses in Mplus structural equation modeling, 1–18.*

Jacobs, K. E., & Roodenburg, J. (2014). The development and validation of the self-report measure of cognitive abilities: A multitrait–multimethod study. *Intelligence*, 42, 5–21. https://doi.org/10.1016/j.intell.2013.09.004.

Kandler, C., Zimmermann, J., & McAdams, D. P. (2014). Core and surface characteristics for the description and theory of personality differences and development. *European Journal of Personality*, 28, 231–243. https://doi.org/10.1002/per.1952.

Kashdan, T. B., Rose, P., & Fincham, F. D. (2004). Curiosity and exploration: Facilitating positive subjective experiences and personal growth opportunities. *Journal of Personality Assessment*, 82, 291–305. https://doi.org/10.1207/s15327752jpa8203_05.

Kelley, E. L. (1927). *Interpretation of educational measurements*. Yonkers, NY: World Press.

Lee, K., & Ashton, M. C. (2004). Psychometric properties of the HEXACO personality inventory. *Multivariate Behavioral Research*, 39, 329–358. https://doi.org/10.1207/s15327906mbr3902_8.

Liang, X., & Yang, Y. (2014). An evaluation of WLSMV and Bayesian methods for confirmatory factor analysis with categorical indicators. *International Journal of Quantitative Research in Education*, 2, 17–38. https://doi.org/10.1504/IJQRE.2014.060972.

López, A., Sanderman, R., Smink, A., Zhang, Y., van Sonderen, E., Ranchor, A., & Schroevers, M. J. (2015). A reconsideration of the self-compassion scale's total score: Self-compassion versus self-criticism. *PLoS One*, 10, e0132940. https://doi.org/10.1371/journal.pone.0132940.

Lyubomirsky, S., & Lepper, H. S. (1999). A measure of subjective happiness: Preliminary reliability and construct validation. *Social Indicators Research*, 46, 137–155. https://doi.org/10.1023/A:1006824100041.

Moshagen, M., Hilbig, B. E., & Zettler, I. (in press). The dark core of personality. *Psychological Review*.

Moshagen, M., & Musch, J. (2014). Sample size requirements of the robust weighted least squares estimator. *Methodology*, 10, 60–70. https://doi.org/10.1027/1614-2241/a000068.

Muris, P. (2016). A protective factor against mental health problems in youths? A critical note on the assessment of self-compassion. *Journal of Child and Family Studies*, 25, 1461–1465. https://doi.org/10.1007/s10826-015-0315-3.

Muris, P., Meesters, C., Pierik, A., & de Kock, B. (2016). Good for the self: Self-compassion and other self-related constructs in relation to symptoms of anxiety and depression in non-clinical youths. *Journal of Child and Family Studies*, 25, 607–617. https://doi.org/10.1007/s10826-015-0235-2.

Muris, P., Otgaar, H., & Petrocchi, N. (2016). Protection as the mirror image of psychopathology: Further critical notes on the self-compassion scale. *Mindfulness*, 7, 787–790. https://doi.org/10.1007/s12671-016-0509-9.

Muris, P., & Petrocchi, N. (2017). Protection or vulnerability? A meta-analysis of the relations between the positive and negative components of self-compassion and psychopathology. *Clinical Psychology & Psychotherapy*, 24, 373–383. https://doi.org/10.1002/cpp.2005.

Muris, P., van den Broek, M., Otgaar, H., Oudenhoven, I., & Lennartz, J. (in press). Good and bad sides of self-compassion: A face validity check of the self-compassion scale and an investigation of its relations to coping and emotional symptoms in non-clinical adolescents. *Journal of Child and Family Studies*.

Murray, A. L., & Johnson, W. (2013). The limitations of model fit in comparing the bi-factor versus higher-order models of human cognitive ability structure. *Intelligence*, 41, 407–422. https://doi.org/10.1016/j.intell.2013.06.004.

Muthén, B. (1990). Moments of the censored and truncated bivariate normal distribution. *British Journal of Mathematical and Statistical Psychology*, 43, 131–143. https://doi.org/10.1111/j.2044-8317.1990.tb00930.x.

Muthén, B., du Toit, S.H.C., & Spisic, D. (1997). *Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes*. Conditionally accepted for publication in *Psychometrika*.

Muthén, L. K., & Muthén, B. O. (1998). *Mplus User's Guide* (Seventh ed.). Los Angeles, CA: Muthén & Muthén.

Neff, K. D. (2003). The development and validation of a scale to measure self-compassion. *Self and Identity*, 2, 223–250. https://doi.org/10.1080/15298860309027.

Neff, K. D. (2016). Does self-compassion entail reduced self-judgment, isolation, and over-identification? A response to Muris, Otgaar, and Petrocchi (2016). *Mindfulness*, 7, 791–797. https://doi.org/10.1007/s12671-016-0531-y.

Neff, K. D., & Dahm, K. A. (2015). Self-compassion: What it is, what it does, and how it relates to mindfulness. In *Handbook of mindfulness and self-regulation* (pp. 121–137). Springer.

Neff, K. D., & Germer, C. K. (2013). A pilot study and randomized controlled trial of the mindful self-compassion program. *Journal of Clinical Psychology*, 69, 28–44. https://doi.org/10.1002/jclp.21923.

Neff, K. D., Long, P., Knox, M., Davidson, O., Kuchar, A., Costigan, A., & Breines, J. (in press). The forest and the trees: Examining the association of self-compassion and its positive and negative components with psychological functioning. *Self and Identity*.

Neff, K. D., Rude, S. S., & Kirkpatrick, K. L. (2007). An examination of self-compassion in relation to positive psychological functioning and personality traits. *Journal of Research in Personality*, 41, 908–916. https://doi.org/10.1016/j.jrp.2006.08.002.

Neff, K. D., Tóth-Király, I., & Colosimo, K. (2018). Self-compassion is best measured as a global construct and is overlapping with but distinct from neuroticism: A response to Pfattheicher, Geiger, Hartung, Weiss, and Schindler (2017). *European Journal of Personality*. https://doi.org/10.1002/per.2148.

Neff, K. D., Whittaker, T. A., & Karl, A. (2017). Examining the factor structure of the self-compassion scale in four distinct populations: Is the use of a total scale score justified? *Journal of Personality Assessment*, 99, 596–607. https://doi.org/10.1080/00223891.2016.1269334.

Nilsson, H., & Kazemi, A. (2016). Reconciling and thematizing definitions of mindfulness: The big five of mindfulness. *Review of General Psychology*, 20, 183–193. https://doi.org/10.1037/gpr0000074.

Nussbeck, F. W., Eid, M., & Lischetzke, T. (2006). Analysing multitrait–multimethod data with structural equation models for ordinal variables applying the WLSMV estimator: What sample

size is needed for valid results? *British Journal of Mathematical and Statistical Psychology*, *59*, 195–213. https://doi.org/10.1348/000711005X67490.

Pearson, K. (1903). Mathematical contributions to the theory of evolution. XI. On the influence of natural selection on the variability and correlation of organs. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, *200*, 1–66.

Pfattheicher, S., Geiger, M., Hartung, J., Weiss, S., & Schindler, S. (2017). Old wine in new bottles? The case of self-compassion and neuroticism. *European Journal of Personality*, *31*, 160–169.

R Development Core Team (2018). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Reyes, D. (2012). Self-compassion: A concept analysis. *Journal of Holistic Nursing*, *30*, 81–89. https://doi.org/10.1177/0898010111423421.

Robitschek, C. (1998). Personal growth initiative: The construct and its measure. *Measurement and Evaluation in Counseling and Development*, *30*, 183–198.

Rosenberg, M. (1979). *Conceiving the self*. New York: Basic Books.

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*, 1–36.

Ryff, C. D., & Keyes, C. L. M. (1995). The structure of psychological well-being revisited. *Journal of Personality and Social Psychology*, *69*, 719–727. https://doi.org/10.1037/0022-3514.69.4.719.

Saucier, G., & Ostendorf, F. (1999). Hierarchical subcomponents of the big five personality factors: A cross-language replication. *Journal of Personality and Social Psychology*, *76*, 613–627. https://doi.org/10.1037/0022-3514.76.4.613.

Scheier, M. F., Carver, C. S., & Bridges, M. W. (1994). Distinguishing optimism from neuroticism (and trait anxiety, self-mastery, and self-esteem): A reevaluation of the life orientation test. *Journal of Personality and Social Psychology*, *67*, 1063–1078. https://doi.org/10.1037/0022-3514.67.6.1063.

Schimmack, U., Oishi, S., Furr, R. M., & Funder, D. C. (2004). Personality and life satisfaction: A facet-level analysis. *Personality and Social Psychology Bulletin*, *30*, 1062–1075. https://doi.org/10.1177/0146167204264292.

Steiger, J. H. (1990). Structural model evaluation and modification: An interval estimation approach. *Multivariate Behavioral Research*, *25*, 173–180. https://doi.org/10.1207/s15327906mbr2502_4.

Van Dam, N. T., van Vugt, M. K., Vago, D. R., Schmalzl, L., Saron, C. D., Olendzki, A., … Gorchov, J. (2018). Mind the hype: A critical evaluation and prescriptive agenda for research on mindfulness and meditation. *Perspectives on Psychological Science*, *13*, 36–61. https://doi.org/10.1177/1745691617709589.

Von Stumm, S., & Ackerman, P. L. (2013). Investment and intellect: A review and meta-analysis. *Psychological Bulletin*, *139*, 841–869. https://doi.org/10.1037/a0030746.

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology*, *54*, 1063–1070. https://doi.org/10.1037/0022-3514.54.6.1063.

Wilhelm, O. (2005). Measuring reasoning ability. In Wilhelm, O. & Engle, R. W. (Eds.), *Handbook of understanding and measuring intelligence*. (pp. 373–392). London: SAGE Publications. https://doi.org/10.4135/9781452233529.n21.

Wilhelm, O., Kaltwasser, L., & Hildebrandt, A. (2017). Will the real factors of prosociality please stand up? A comment on Böckler, Tusche, and Singer (2016). *Social Psychological and Personality Science*, Advance online publication.

Wilhelm, O., Witthöft, M., & Schipolowski, S. (2010). Self-reported cognitive failures: Competing measurement models and self-report correlates. *Journal of Individual Differences*, *31*, 1–14. https://doi.org/10.1027/1614-0001/a000001.

Yung, Y.-F., Thissen, D., & McLeod, L. D. (1999). On the relationship between the higher-order factor model and the hierarchical factor model. *Psychometrika*, *64*, 113–128. https://doi.org/10.1007/BF02294531.