# Assessing the temporal stability of psychological constructs

Lodder, Paul; Kupper, Nina; Mols, Floortje; Emons, Wilco H.M.; Wicherts, Jelte M.

*Published in:*
Journal of Research in Personality

*DOI:*
10.1016/j.jrp.2022.104299

*Publication date:*
2022

*Document Version*
Publisher's PDF, also known as Version of record

*Citation for published version (APA):*
Lodder, P., Kupper, N., Mols, F., Emons, W. H. M., & Wicherts, J. M. (2022). Assessing the temporal stability of psychological constructs: An illustration of Type D personality, anxiety and depression. *Journal of Research in Personality*, *101*, [104299]. https://doi.org/10.1016/j.jrp.2022.104299

Contents lists available at ScienceDirect

Journal of Research in Personality

journal homepage: www.elsevier.com/locate/jrp

# Assessing the temporal stability of psychological constructs: An illustration of Type D personality, anxiety and depression

Paul Lodder[a,b,*], Nina Kupper[b], Floortje Mols[b,c], Wilco H.M. Emons[a,d], Jelte M. Wicherts[a]

[a] Department of Methodology and Statistics, Tilburg University, the Netherlands
[b] Center of Research on Psychological and Somatic disorders (CoRPS), Department of Medical and Clinical Psychology, Tilburg University, the Netherlands
[c] Department of Research, Netherlands Comprehensive Cancer Organization (IKNL), Utrecht, the Netherlands
[d] Department of Research & Innovation, Cito, Arnhem, the Netherlands

ARTICLE INFO

*Keywords:*
Temporal stability
Type D personality
Latent variable models
Longitudinal measurement invariance
Individual change
Trait-State-Occasion Model
Latent Growth Curve Model
Depression
Anxiety

ABSTRACT

Various methods exist to assess the temporal stability of psychological constructs. In this paper we discuss common methods based on a review of the personality traits negative affectivity and social inhibition. Most methods ignore the non-normal distributions and measurement error in the questionnaire item scores. We illustrate how to handle these issues using three longitudinal latent variable models. We further highlight the importance of testing the often overlooked assumption of longitudinal measurement invariance. Lastly, we apply several longitudinal measurement invariance models, univariate and multivariate latent growth curves models, and latent trait-state occasions models to data from 2625 cancer survivors, to assess the temporal stability of negative affectivity, social inhibition, depression, anxiety, across a period of four years.

## 1. Introduction

Personality traits are considered relatively enduring sets of behaviors, feelings and thoughts that characterize individuals (Roberts & Mroczek, 2008). These behavioral, emotional, and cognitive patterns develop from an interplay between biological and environmental influences, and were typically thought to remain stable after reaching adulthood (McGraw & Costa, 1994). However, more recent evidence suggests that personality traits may continue to change throughout adulthood and even into old age (Mroczek & Spiro, 2003; Mroczek, Graham, Turiano, & Aro-Lambo, 2021). Such change in personality can either be normative or non-normative. Normative change is defined as the generalizable patterns of personality development typically seen in most individuals, whereas non-normative change reflects all individual deviation from the normative developmental trajectories (Roberts, Walton & Viechtbauer, 2006). An example of normative change is that people generally become more socially mature (calm, responsible, confident) as they grow older (Roberts & Wood, 2006). However, (*epi*) genetic influences or environmental factors, such as major life events and work experiences, may alter the course of such normative development in directions unique to how each individual interacts with his or her environment (Roberts, Wood & Caspi, 2008; Leszko, Elleman, Bastarache, Graham & Mroczek, 2016). Although personality can change into adulthood, several studies suggests that both the genetic and environmental influences on personality increase in stability with age (Briley & Tucker-Drob, 2014; Li-Gao et al., 2021).

Throughout the years, several statistical methods have been used to assess the temporal stability of personality traits, or rather the stability of psychological constructs in general (for an overview see De Fruyt et al., 2006). A first distinction can be made between methods assessing absolute change vs methods that focus on relative change (also known as differential change, Caspi, Roberts & Shiner, 2005). The absolute change perspective involves determining whether an individual or aggregate score on one time point differs from the score at one or more other time points. Absolute change can be assessed for separate individuals (*individual-level absolute stability*) or for groups of individuals such as the entire sample (*mean-level absolute stability*). From a relative change perspective, it is by definition not possible to assess the change of a single individual, because relative change is defined as change relative to others. Therefore, *relative stability* methods typically assess whether the ranking of people's scores on a measured construct changes over time. In Study 1, we focus on the literature on the temporal stability of

[*] Corresponding author at: Department of Methodology and Statistics, Tilburg School of Social and Behavioral Sciences (TSB), Tilburg University, PO Box 90153, 5000 LE Tilburg, the Netherlands.
*E-mail addresses:* p.lodder@tilburguniversity.edu, p.lodder@uvt.nl (P. Lodder).

https://doi.org/10.1016/j.jrp.2022.104299
Received 18 January 2022; Received in revised form 26 August 2022; Accepted 16 September 2022
Available online 23 September 2022

Type D personality to review current practices and to discuss methods designed to detect various types of temporal stability.

Personality and other psychological characteristics are often assessed using scores on multi-item questionnaires. Scores on such psychological questionnaires are known to contain measurement error, where someone's item score does not perfectly reflect this person's score on the latent psychological construct. As a result, measurement error may obscure the true association between constructs, leading to attenuated effect sizes, a phenomenon known as attenuation bias (Spearman, 1904). Moreover, as statistical models get more complex, ignoring measurement error may even result in overestimated associations between constructs (Cole & Preacher, 2014). This highlights the importance of using statistical models that can handle measurement error when assessing the stability of psychological constructs. In Study 2, we use state of the art psychometric (latent variable) modelling approaches to investigate the temporal stability of the two Type D personality traits in relation to depression and anxiety.

### 1.1. *Type D personality*

Type D personality is most prominently studied in the field of psychosomatic medicine, where it is seen as a risk factor of cardiac events in patients suffering from cardiovascular disease (Grande, Romppel & Barth, 2012; Piepoli et al., 2016; Kupper & Denollet, 2018). Research on the temporal stability of Type D personality illustrates many statistical and psychometric issues in the study of the temporal stability of psychological constructs. Type D is measured with a multi-item questionnaire (DS14; Denollet, 2005) involving ordinal and often skewed item scores. Individuals with a Type D (Distressed) personality are considered to score high on the two personality traits negative affectivity (NA) and social inhibition (SI). NA concerns the tendency of people to experience negative thoughts and emotions, while SI concerns the difficulty in expressing such thoughts and emotions, especially in social interactions.

These is a strong association between Type D's two 'distressed' personality traits and the negative emotional states depression and anxiety (Lodder et al., 2019). This association is especially pronounced for NA and depression, with correlations between scale scores ranging between 0.4 and 0.7 (Spindler, Kruse, Zwisler, & Pedersen, 2009; Ossola, De Panfilis, Tonna, Ardissino, & Marchesi, 2015). These correlations point to a significant statistical overlap between the *trait* NA and the more *episodic* depression, leading some scholars to question whether NA is really a personality trait, or whether depression has *trait*-like characteristics (Ossola et al., 2015).

A limitation of these studies is that they failed to take into account the measurement error in the questionnaire scores. In Study 2, we illustrate how to assess temporal stability of psychological constructs. Our illustration will focus on the personality traits NA and SI and how their temporal stability relates to that of depression and anxiety. We will use latent variable models that not only take into account measurement error in the questionnaire scores, but can also appropriately model the non-normally distributed ordinal item scores typically encountered in psychological research. Treating such ordinal scores as continuous and normally distributed may result in biased parameter estimates (Rhemtulla, Brosseau-Liard & Savalei, 2012; Lodder, Emons, Denollet & Wicherts, 2021) and may therefore result in misleading conclusions regarding the stability of psychological constructs.

### 1.2. *Study aims and overview*

The aims of this study are twofold. First, in Study 1, we systematically review the methods typically used by researchers to assess the stability of psychological constructs, and use Type D personality (Denollet, 2005) as an example. We discuss how these common methods risk incorrect conclusions regarding the temporal stability of the personality traits NA and SI by ignoring the presence of measurement error in the item scores and by not testing the often-ignored assumption of longitudinal measurement invariance. This review does not only shed light on the earlier research on this issue, but also provides an ideal opportunity to introduce the statistical methods applied researchers typically use to assess temporal stability. Second, in Study 2, we illustrate how to handle these issues using a series of three longitudinal latent variable models used to investigate and compare the temporal stability of Type D personality, anxiety and depression. We will discuss how to handle the often-overlooked problem that the questionnaire item scores are ordinal and non-normally distributed when building the latent variable models. We subsequently illustrate how to test the often overlooked—yet crucial—assumption underlying the longitudinal analysis of questionnaire data, namely that the properties of your instrument's measurement model (e.g. item factor loadings) are invariant across all measurement occasions (Liu et al., 2017). The relative temporal stability and autoregressive effects of the psychological constructs can also be inferred from these models. Next, we show how latent growth curve models (Hertzog, Lindenberger, Ghisletta & von Oertzen, 2006) can be used to investigate the mean and individual level absolute stability of psychological constructs, while taking into account measurement error in the item scores. Multivariate latent growth curve models also allow for estimating how intra-individual change in for instance depression correlates with intra-individual change in negative affectivity. Lastly, we illustrate the benefit of a latent trait-state-occasion model (Cole, Martin & Steiger, 2005) to estimate what part of a construct can be considered a stable trait and what part a changeable state.

We hypothesized that NA and SI both show absolute and relative temporal stability over time, and that both constructs correspond more to a stable trait than to a changeable state. In line with earlier research (Ossola et al., 2015), we further hypothesized that any individual changes in the personality trait NA would correlate with individual changes in depression and anxiety, but that SI would not show this association.

## 2. Study 1: Systematic review

The goal of this systematic review was to review the temporal stability studies conducted in the context of research on Type D personality, to document common practices used to analyze stability in this literature, and to discuss the limitations of these methods. Our review included all studies that assessed Type D personality using the DS14 questionnaire on at least two measurement occasions and then used a statistical analysis to determine whether Type D personality, NA, or SI showed temporal stability.

### 2.1. *Method*

On November 4th 2019, the electronic databases Pubmed and Psycinfo were used to search the full text of empirical articles for the terms '("Type D personality" OR "negative affectivity" OR "social inhibition") AND ("stability" OR "test-retest")'. The search resulted in 142 unique studies. After screening the full texts, 24 studies met the inclusion criteria of our review. In total, those 24 studies reported 75 tests for the temporal stability of either Type D personality or its subcomponents NA or SI. An updated literature search on August 3rd 2022 resulted in 23 eligible full texts. After screening we included 1 additional study, resulting in a total inclusion of 25 studies, reporting 76 temporal stability tests. We subsequently divided these studies across the statistical approach(es) taken to analyze the stability of NA and SI.

### 2.2. *Results*

For each of the 76 tests included in the review, Table 1 reports the sample characteristics, the statistical method used to assess stability, the personality construct studied, the longest follow-up time, and the results of the stability assessment. The following sections will discuss the

**Table 1**

For all 24 studies included in the systematic review, the type of investigated stability, the sample characteristics, the statistical method used, the personality construct studied, the longest follow-up time in months, and the results of the stability assessment.

| Study | Sample | Statistical method | Construct | FU (months) | Results |
|---|---|---|---|---|---|
| *Relative stability* | | | | | |
| Dannemann et al. (2010) | 126 German cardiac patients | Test-retest correlation | NA | 6 | Rxx’ = 0.61 |
| Romppel et al. (2012) | 679 German cardiac patients | Test-retest reliability | NA | 72 | Rxx’ = 0.61 |
| Gremigni & Sommaruga (2005) | 30 Italian cardiac patients | Test-retest correlation | NA | 1 | Rxx’ = 0.62 |
| Bunevicius et al. (2013) | 49 Lithuanian CHD patients | Test-retest reliability | NA | 0.5 | Rxx’ = 0.69 |
| Denollet (2005) | 121 Cardiac rehabilitation patients) | Test-retest correlation | NA | 3 | Rxx’ = 0.72 |
| Zohar (2016) | 285 Israeli community volunteers | Test-retest correlation | NA | 72 | Rxx’ = 0.72 |
| Aluja et al. (2019) | 65 Spanish university students | Test-retest reliability | NA | 2 | Rxx’ = 0.77 |
| Denollet (1998) | 60 Belgian CHD patients | Test-retest reliability | NA | 3 | Rxx’ = 0.78 |
| Spindler et al. (2009) | 117 Danish cardiac patients | Test-retest reliability | NA | 3 | Rxx’ = 0.78 |
| Alçelik et al. (2012) | 100 Turkish hemodialysis patients | Test-retest reliability | NA | 1 | Rxx’ = 0.84 |
| Pedersen et al. (2009) | 57 Healthy Ukrainians | Test-retest reliability | NA | 1 | Rxx’ = 0.85 |
| Bagherian & Ehsan (2011) | 71 Iranians (MI + healthy) | Test-retest correlation | NA | 2 | Rxx’ = 0.86 |
| Montero et al. (2017) | 253 Spaniards (MI + cancer + healthy) | Test-retest reliability | NA | 6 | Rxx’ = 0.88 |
| Dannemann et al. (2010) | 126 German cardiac patients | Test-retest correlation | SI | 6 | Rxx’ = 0.59 |
| Romppel et al. (2012) | 679 German cardiac patients | Test-retest reliability | SI | 72 | Rxx’ = 0.60 |
| Pedersen et al. (2009) | 57 Healthy Ukrainians | Test-retest reliability | SI | 1 | Rxx’ = 0.63 |
| Bagherian & Ehsan (2011) | 71 Iranians (MI + healthy) | Test-retest correlation | SI | 2 | Rxx’ = 0.77 |
| Alçelik et al. (2012) | 100 Turkish hemodialysis patients | Test-retest reliability | SI | 1 | Rxx’ = 0.78 |
| Spindler et al. (2009) | 117 Danish cardiac patients | Test-retest reliability | SI | 3 | Rxx’ = 0.79 |
| Gremigni & Sommaruga (2005) | 30 Italian cardiac patients | Test-retest correlation | SI | 1 | Rxx’ = 0.81 |
| Bunevicius et al. (2013) | 49 Lithuanian CAD patients | Test-retest reliability | SI | 0.5 | Rxx’ = 0.81 |
| Aluja et al. (2019) | 65 Spanish university students | Test-retest reliability | SI | 2 | Rxx’ = 0.82 |
| Denollet (2005) | 121 Cardiac rehabilitation patients) | Test-retest correlation | SI | 3 | Rxx’ = 0.82 |
| Zohar (2016) | 285 Israeli community volunteers | Test-retest correlation | SI | 72 | Rxx’ = 0.82 |
| Denollet (1998) | 60 Belgian CHD patients | Test-retest reliability | SI | 3 | Rxx’ = 0.87 |
| Montero et al. (2017) | 253 Spaniards (MI + cancer + healthy) | Test-retest reliability | SI | 6 | Rxx’ = 0.89 |
| Zohar (2016) | 285 Israeli community volunteers | Test-retest reliability | Type D (NA*SI) | 72 | Rxx’ = 0.78 |
| Ossola et al. (2015) | 304 Italian CHD patients | ICC | NA | 12 | ICC = 0.48 |
| Nefs et al. (2012) | 1012 Dutch primary care patients | ICC (2-way; consistency; average) | NA | 12 | ICC = 0.64 (men) & 0.63 (women) |
| Conden et al. (2014) | 313 Swedish acute MI patients | ICC (2-way) | NA | 12 | ICC = 0.71 |
| Bouwens et al. (2019) | 294 Dutch vascular surgery patients | ICC | NA | 12 | ICC = 0.71 |
| Loosman et al. (2018) | 249 Dutch dialysis patients | ICC | NA | 6 | ICC = 0.72 |
| Lim et al. (2011) | 111 Korean CHD patients | ICC | NA | 2 | ICC = 0.76 |
| Yu et al. (2010) | 100 Chinese CHD patients | ICC | NA | 3 | ICC = 0.76 |
| Kupper et al. (2011) | 730 Dutch twins | ICC | NA | 108 | ICC = 0.78 (twin A) & 0.72 (twin B) |
| Spindler et al. (2009) | 117 Danish cardiac patients | ICC (2-way; consistency; average) | NA | 3 | ICC = 0.87 |
| Loosman et al. (2018) | 249 Dutch dialysis patients | ICC | SI | 6 | ICC = 0.69 |
| Ossola et al. (2015) | 304 Italian CHD patients | ICC | SI | 12 | ICC = 0.70 |
| Nefs et al. (2012) | 1012 Dutch primary care patients | ICC (2-way; consistency; average) | SI | 12 | ICC = 0.73 (men) & 0.65 (women) |
| Yu et al. (2010) | 100 Chinese CHD patients | ICC | SI | 3 | ICC = 0.74 |
| Lim et al. (2011) | 111 Korean CHD patients | ICC | SI | 2 | ICC = 0.77 |
| Conden et al. (2014) | 313 Swedish acute MI patients | ICC (2-way) | SI | 12 | ICC = 0.80 |
| Bouwens et al. (2019) | 294 Dutch vascular surgery patients | ICC | SI | 12 | ICC = 0.80 |
| Kupper et al. (2011) | 730 Dutch twins | ICC | SI | 108 | ICC = 0.83 (twin A) & 0.82 (twin B) |
| Spindler et al. (2009) | 117 Danish cardiac patients | ICC (2-way; consistency; average) | SI | 3 | ICC = 0.88 |
| Kupper et al. (2011) | 730 Dutch twins | ICC | Type D (2-groups) | 108 | ICC = 0.62 (twin A) & 0.58 (twin B) |
| Ossola et al. (2015) | 304 Italian CHD patients | ICC | Type D (NA*SI) | 12 | ICC = 0.52 |
| Bouwens et al. (2019) | 294 Dutch vascular surgery patients | ICC | Type D (NA*SI) | 12 | ICC = 0.72 |
| Conden et al. (2014) | 313 Swedish acute MI patients | ICC | Type D (NA*SI) | 12 | ICC = 0.76 |
| *Mean absolute stability* | | | | | |
| Pedersen et al. (2009) | 57 Healthy Ukrainians | Paired *t*-test | NA | 1 | *d* = 0.004, NS * |
| Pedersen et al. (2009) | 57 Healthy Ukrainians | Paired *t*-test | SI | 1 | *d* = 0.10, NS * |
| Romppel et al. (2012) | 679 German cardiac patients | Cohen’s d | NA | 72 | *d* = 0.08, NS |
| Romppel et al. (2012) | 679 German cardiac patients | Cohen’s d | SI | 72 | *d* = 0.01, NS |
| Dannemann et al. (2010) | 126 German cardiac patients | RM ANOVA | NA | 6 | *d* = 0.04, NS * |
| Dannemann et al. (2010) | 126 German cardiac patients | RM ANOVA | SI | 6 | *d* = 0.12, p <.05 * |
| *Individual absolute stability* | | | | | |

(continued on next page)

**Table 1** (*continued*)

| Study | Sample | Statistical method | Construct | FU (months) | Results |
|---|---|---|---|---|---|
| Romppel et al. (2012) | 679 German cardiac patients | RCI | NA | 72 | Significant change: 26.4 % |
| Romppel et al. (2012) | 679 German cardiac patients | RCI | SI | 72 | Significant change: 22.7 % |
| *Ipsative stability* | | | | | |
| Pelle et al. (2008) | 386 Dutch CAD patients | % caseness | Type D (2-groups) | 3 | Stable classification: 81 % |
| Zohar (2016) | 285 Israeli community volunteers | % caseness | Type D (2-groups) | 72 | Stable classification: 82 % |
| Nefs et al. (2012) | 1012 Dutch primary care patients | % caseness | Type D (2-groups) | 12 | Stable classification: 85 % |
| Aguayo-Carreras et al. (2020) | 106 Brazilian psoriaris patients | % caseness | Type D (2-groups) | 48 | Stable classification: 47.5 % |
| Martens et al. (2007) | 475 Dutch acute MI patients | Logistic regression | Type D (2-groups) | 18 | $\chi^2$ (2) = 1.6, $p$ = 0.45 |
| Zohar (2016) | 285 Israeli community volunteers | Chi-square test | Type D (2-groups) | 72 | κ = 0.50* |
| Bouwens et al. (2019) | 294 Dutch vascular surgery patients | Cohen's Kappa | Type D (2-groups) | 12 | κ = 0.32 |
| Conden et al. (2014) | 313 Swedish acute MI patients | Cohen's Kappa | Type D (2-groups) | 12 | κ = 0.40 |
| Romppel et al. (2012) | 679 German cardiac patients | Cohen's Kappa | Type D (2-groups) | 72 | κ = 0.42 |
| Ossola et al. (2015) | 304 Italian CHD patients | Cohen's Kappa | Type D (2-groups) | 12 | κ = 0.49 |
| Loosman et al. (2018) | 249 Dutch dialysis patients | Cohen's Kappa | Type D (2-groups) | 6 | κ = 0.52 |
| Dannemann et al. (2010) | 126 German cardiac patients | RM ANOVA | Type D (2-groups) | 6 | κ = 0.26 |
| Conden et al. (2014) | 313 Swedish acute MI patients | Cohen's Kappa | NA (dichotomized) | 12 | κ = 0.48 |
| Bouwens et al. (2019) | 294 Dutch vascular surgery patients | Cohen's Kappa | NA (dichotomized) | 12 | κ = 0.49 |
| Conden et al. (2014) | 313 Swedish acute MI patients | Cohen's Kappa | SI (dichotomized) | 12 | κ = 0.53 |
| Bouwens et al. (2019) | 294 Dutch vascular surgery patients | Cohen's Kappa | SI (dichotomized) | 12 | κ = 0.54 |
| *Genetic stability* | | | | | |
| Kupper et al. (2011) | 730 Dutch twins | ACE model | NA | 108 | Non-genetic variance 55–60 % |
| Kupper et al. (2011) | 730 Dutch twins | ACE model | SI | 108 | Non-genetic variance 51–58 % |
| Kupper et al. (2011) | 730 Dutch twins | ACE model | Type D (2-groups) | 108 | Non-genetic variance 51–66 % |
| *Longitudinal measurement invariance* | | | | | |
| Romppel et al. (2012) | 679 German cardiac patients | SEM | NA & SI | 72 | Stable factor loadings |
| Conden et al. (2014) | 313 Swedish acute MI patients | SEM | NA & SI | 12 | Stable factor loadings |

ACE = statistical model used to analyze the results of twin studies; CHD = coronary heart disease; FU = follow-up; ICC = intraclass correlation; MI = myocardial infarction; NA = negative affectivity; RCI = reliable change index; SEM = structural equation modeling; SI = social inhibition.
*effect size calculated based on statistics reported in the published study.

findings of these stability tests separately for each of the investigated stability types.

### 2.2.1. *Relative stability*

Of all 25 studies included in the review, 22 (88.0 %) investigated relative stability, making this the most popular approach to study temporal stability. Relative stability measures assess whether the relative ranking of individual scores remains stable over time. In their basic form, statistical models assessing relative stability estimate whether scores on $T_X$ covary with scores on $T_Y$, where X and Y denote two distinct measurement occasions.

*2.2.1.1. Test-retest correlation.* In the reviewed literature, the test–retest correlation (in some included studies referred to as test–retest reliability) was the most popular method to study the relative stability of the Type D personality traits, arguably because it simply involves computing the Pearson correlation coefficient between the $T_X$ and $T_Y$ scores. The Pearson correlation coefficient can be used when the association between two measurements is linear, while Spearman's rho is useful for estimating non-linear but monotonically increasing associations. The Pearson correlation coefficients reported in Table 1 ranged from 0.61 to 0.88 (median $r$ = 0.77) for NA and from 0.59 to 0.89 (median $r$ = 0.81) for SI. One way to operationalize Type D personality is by multiplying the scores of the NA and SI variables (Lodder, 2020a; 2020b; Ferguson et al., 2009). One study (Zohar, 2016) reported a correlation of 0.78 between the repeated measurements of this NA*SI product score. Taken together, these findings suggest that both Type D as well as NA and SI generally showed acceptable relative stability based on the Pearson correlation coefficients.

Drawbacks of the Pearson correlation coefficient are that it is limited to correlations between two measurements and that it ignores the measurement error if no proper correction for attenuation (Muchinsky, 1996) is applied. We instead recommend using a latent variable model to directly estimate the correlation between the latent variable scores at two time points (see Study 2).

*2.2.1.2. Intraclass correlation.* A popular method to assess relative stability for two or more repeated measurements is the intraclass correlation coefficient (ICC; Bartko, 1976; McGraw & Wong, 1996; Weir, 2005). Historically it was developed as an index of reliability (e.g. test–retest or interrater reliability), but several other purposes exist. There exist various types of intraclass correlation models, but when the goal is to assess temporal stability of repeated measurements, then the 2-way mixed-effects model is the method of choice (Koo & Li, 2016). Researchers also need to decide whether the ICC is calculated based on single item scores or on an average (or sum) of multiple item scores. Use of average measurements is the preferred option when psychological constructs are measured using multi-item questionnaires. Lastly, researchers should decide whether they are interested in consistency or absolute agreement. Similar to the correlation coefficient, the consistency ICC is sensitive to the relative ranking of individuals, but a key difference between them is that the correlation expresses the degree to which variables Y and X are associated through a linear transformation (Y = aX + b), while the consistency ICC measures the extent to which Y and X are associated by adding a constant (Y = X + b). The model used to estimate ICCs assumes equal variance at the repeated measurements. Violating this assumption often results in attenuated ICC estimates relative to the correlation coefficient (McGraw & Wong, 1996). If all individuals change to the same degree, then the consistency ICC will be equal to one. The absolute agreement ICC, on the other hand, also considers absolute changes over time and will therefore be lower than one if there are individual differences in the intra-individual change.

Table 1 shows that for NA, the eleven included ICCs ranged from 0.48 to 0.87 (median ICC = 0.72), while for SI the eleven ICCs ranged from 0.65 to 0.88 (median ICC = 0.77). For Type D personality (the NA*SI product score), the three included ICCs ranged from 0.52 to 0.76 (median ICC = 0.72). For most ICCs included in our systematic review researchers did not specify the investigated type of ICC. Given that decisions regarding ICC type (single vs average ratings & consistency vs absolute agreement) may considerably influence the estimated ICC, it is difficult to determine whether Type D, NA and SI are temporally stable based on the ICCs reported in these studies.

Neither of the two studies that reported the chosen ICC method used an absolute agreement definition. The ICC estimated using a consistency definition is always equal to or larger than the ICC estimated according to the absolute agreement definition. Consequently, the results of studies that have investigated temporal stability using the consistency ICC (Nefs et al., 2012; Spindler et al., 2009), may appear to be more temporally stable than they really are than if they would have also taken into account absolute stability.

### 2.2.2. *Mean-level absolute stability*

Of all 24 studies included in the review, three (12.5 %) investigated mean-level absolute stability, each using a different statistical method, including the paired *t*-test, the repeated measures ANOVA and the standardized mean difference.

*2.2.2.1. Paired t-test.* The paired (or dependent) *t*-test assesses absolute difference in the mean scores of two repeated or dependent measurements. Commonly, the null hypothesis of a paired *t*-test is that the mean scores on the two measurements are equal. This null hypothesis is rejected when the difference becomes large enough in relation to its standard error to be statistically significant, with the standard error being a function of the sample size, standard deviation and correlation between the two repeated measurements. When assessing absolute stability, researchers typically conclude absolute stability when the difference between two measurements is *not* statistically significant, which entails a statistically invalid conclusion. Table 1 shows that one study (Pedersen et al., 2009) included in the review assessed the absolute stability of NA and SI using a paired *t*-test. Absolute stability was concluded based on both tests because they were not statistically significant with *t*-values of 0.064 (NA) and 0.7 (SI).

*2.2.2.2. Standardized mean difference.* Absolute stability can also be assessed by computing the standardized mean difference of the scores on two measurements using the Cohen's *d* effect size for repeated measures. ***Appendix A*** shows the formula used to compute Cohen's *d* for paired data. This method assumes homogeneous variances across repeated measurements (Cohen, 1988). It determines the mean-level absolute stability and can therefore not assess individual-level absolute stability. Table 1 shows that one included study (Romppel, Herrmann-Lingen, Vesper & Grande, 2012) assessed the mean-level absolute stability of NA and SI using the standardized mean difference. This study concluded absolute stability for both personality traits based on non-significant Cohen's *d* estimates of 0.08 (NA) and 0.01 (SI).

*2.2.2.3. Repeated measures ANOVA.* The mean-level absolute stability of two or more repeated measurements can also be assessed using a repeated measures (RM) ANOVA. When there are only two repeated measurements, the RM ANOVA is equivalent to a paired *t*-test. Typically, researchers conclude absolute stability when the within-subjects effect (e.g. Time or Measurement) does not reach statistical significance. Table 1 indicates that one study (Dannemann et al., 2010) used an RM ANOVA to test the absolute stability of NA and SI based on the within subjects Time effect. It turned out that absolute stability was concluded for NA, but not for SI. Note that this non-significant mean difference does not necessarily indicate the absence of temporal stability in the population because the p-value of this test is also influenced by the sample size.

### 2.2.3. *Individual-level absolute stability*

Whereas the absolute stability methods discussed in the previous section all assessed stability at the group level (e.g., the full sample), the reliable change index (RCI) is a method developed to determine for each individual separately whether there is significant absolute change over time (Jacobson & Truax, 1992). Whether this individual change is statistically significant depends in part on the amount of measurement error in the questionnaire scores. As many psychological questionnaires are not perfectly reliable, the RCI can be used to assess whether the observed individual change is larger than the change that may occur due to measurement error. Although change scores have often been criticized for having low reliability, recent psychometric advanced suggest that this is not necessarily the case when modeling change scores from a multilevel perspective, distinguishing individual change on the within-subjects level from group change on the between-subjects level (Gu, Emons, & Sijtsma, 2018). ***Appendix A*** present the mathematical details behind computing the RCI. Note that when calculating the RCI from reliability estimates within a classical test theory perceptive, then researchers make the strong assumption that the variance of the two measurements are equal, as well as the error variances (implying equal reliability coefficients across measurements; Maassen, Bossema & Brand, 2009).

In our review, one included study used the RCI to assess individual stability of NA and SI (Romppel et al., 2012). Although this study did not find absolute change averaged across all participants, significant *individual* change was observed for 26.4 % of the participants on NA for 22.7 % on SI. Of these changes, the proportion of significant change involving either increased or decreased scores was approximately equal.

### 2.2.4. *Ipsative stability*

Of all 25 studies included in the review, ten (40 %) investigated the ipsative stability of Type D personality. Ipsative stability refers to the continuity or temporal stability of a trait pattern within individuals (De Fruyt et al., 2006). This trait pattern typically involves two or more traits, but in some instances, researchers assess the continuity of having high scores on a single trait. The temporal stability of trait patterns can be assessed using latent variable models such as latent transition analysis or repeated measures latent class analysis (Collins & Lanza, 2009). In the Type D literature, researchers have assessed ipsative stability by investigating temporal changes in the classification of individuals in personality groups. The classification in personality groups is based on whether or not individuals score above a cutoff on NA and/or SI. This either results in two (Type D & no Type D) or four (Type D, NA + SI-, NA-SI+, NA-SI-) personality groups and researchers subsequently calculate the percentage of individuals that change group membership across time. Some studies have assessed the ipsative stability of NA and SI separately by classifying participants in High vs Low NA groups, and High vs Low SI groups. A major disadvantage of this approach is that the initial classification in personality groups ignores valuable information on individual differences in these personality traits. For NA and SI, ipsative stability methods cannot detect changes happening within each of the 0–9 or 10–28 ranges, because changes do not affect classification. We therefore argue that ipsative stability methods should only be used when the main interest is the stability in the classification, rather than stability in the underlying personality traits.

The studies included in our review utilized several statistical methods to assess stability in classification, including descriptive statistics, logistic regression, chi-square tests and Cohen's Kappa. Table 1 indicates that four studies (Aguayo-Carreras et al., 2021; Pelle et al. 2008; Nefs, et al., 2012; Zohar, 2016) used descriptive statistics to assess the ipsative stability of Type D personality. Note that for one of these studies (Pelle et al., 2008), the goal was not to assess temporal stability, but rather to investigate whether patients receiving cardiac rehabilitation would change in their Type D classification. According to three of those studies, between 81 % and 85 % of the participants did not change in their Type D or no Type D classification over time. However, the fourth study found that 47 % of the participants remained stable in their Type D classification. These authors studied a relatively small patient sample across a 4-year follow-up. As these analyses did not involve inferential statistics it is not possible to generalize these findings beyond the studied samples. To solve that problem, another study (Martens et al. 2007) used a logistic regression to show that measurement occasion did not predict the Type D classification, suggesting that it is stable over time. Another study (Zohar, 2016) used a chi-square test to reject the null hypothesis that the classifications at two measurement occasions are independent. However, the null hypothesis of no dependency is unrealistic, as at least some dependency in classification is expected when the same participants are measured twice. Lastly, five studies used Cohen's Kappa to assess the agreement in classification at two measurement occasions. The Kappa estimates ranged from 0.32 to 0.52 (median = 0.42), suggesting fair to moderate agreement between in Type D classifications over time. Two of these studies also used Cohen's Kappa to study the ipsative stability separately for NA and SI, indicating moderate agreement with Kappa estimates of 0.48 and 0.49 for NA, and 0.53 and 0.54 for SI.

### *2.2.5. Genetic stability*

In behavior genetics, ACE models are frequently applied to twin data to estimate for various psychological traits the proportion of variance explained by either additive genetic (A), shared (C), or non-shared (E) environmental influences (Rijsdijk & Sham, 2002). Such studies typically use data from both identical and fraternal twins to determine the relative contribution of these latent genetic and environmental components in explaining variation in a psychological trait of interest. When longitudinal data is available, then researchers can investigate whether the relative importance of these components remains stable over time.

Traditional genetic stability models do not assess the mean structure of the psychological construct and can therefore not assess absolute stability (neither on the individual level nor on the sample level; for exceptions see McArdle, 1986; Neale & McArdle, 2000; Nivard et al., 2015). Whereas relative and absolute stability methods determine *whether* and *to what extent* the construct shows temporal stability, genetic stability methods intend to elucidate *why* individual differences on a construct show temporal stability or not (Figueredo, de Baca, & Black, 2014). Genetic stability methods estimate the proportion of variance in traits that is attributable to either genetic or environmental influences and determines whether this variance decomposition is stable across time by assessing whether later time points contain genetic or environmental influences not shared with the first point. As this provides valuable information regarding the etiology of the psychological traits, genetic stability methods could complement relative and absolute stability methods when assessing the temporal stability of psychological traits.

Table 1 indicates that one study included in our review used an ACE model to assess the genetic stability of both Type D personality and its subcomponents NA and SI (Kupper et al., 2011). This study used structural equation modeling to fit longitudinal ACE model on the aggregate NA and SI scores. The results showed that across nine years, the heritability of NA was stable and varied only slightly (between 40 and 45 %). Similar genetic stability over time was found for SI, with heritability estimates varying between 42 and 49 %.

A limitation of this study is that modeling the aggregate NA and SI scores rather than raw item scores fail to consider measurement error in the item scores. It also inhibits testing the assumption of longitudinal measurement invariance (Liu et al., 2015). Simulation studies have indicated that modeling aggregated rather than raw item scores result in underestimated heritability estimates and component correlations across time, which may bias conclusions regarding the stability of the genetic and environmental components (van den Berg, Glas & Boomsma, 2007; Schwabe, Gu, Tijmstra, Hatemi & Pohl, 2019). Future research could prevent this problem by specifying a measurement model for the latent NA and SI constructs when testing the genetic stability of these traits with an ACE model.

### *2.2.6. A small simulation study*

The methods reviewed above differ in the types of temporal stability they can detect. In the section we illustrate based on simulated data whether various statistical methods can detect relative stability, mean-level absolute stability or individual-level absolute stability. We simulated data for two repeated measurements of a particular construct and in four scenarios varied the presence or absence of relative stability, mean-level absolute stability and individual absolute stability. The upper row of Fig. 1 reports the estimated temporal stability in terms of Pearson correlation coefficient, Cohen's *d* for paired data, consistency ICC, absolute agreement ICC, and the reliable change index (assuming a test reliability of 0.9). The middle row visualizes the individual and mean scores on two repeated measurements. The bottom row shows the individual and mean growth curves.

In the *first* scenario, all individual scores remain constant across time. As expected, all methods suggest temporal stability because there is no intra-individual change and no interindividual differences in this intraindividual change. The *second* scenario also does not involve interindividual differences in change, because the intraindividual change of each individual is similarly positive. Consequently, both the mean-level (Cohen's d) and individual-level (RCI) absolute stability methods indicate that there is significant change across time, while the relative stability methods (Pearson correlation and consistency ICC) suggest perfect relative stability because the ranking of individual scores does not change. As opposed to the consistency ICC, the absolute agreement ICC is sensitive to deviations from absolute stability and therefore does not suggest temporal stability.

In the *third* scenario, the simulated scores on both time points are completely unrelated yet similar on average. As expected, Cohen's d suggests mean-level absolute stability. Pearson correlation and both ICCs indicate no temporal stability because the ranking of individuals completely changes across time. Interestingly, the RCI suggests poor individual-level absolute stability because significant change across time is concluded for more than half (51 %) of the individuals.

Lastly, in the *fourth* scenario the change across time depends on the score at the first timepoint. The highest baseline scores increase across time; the lowest baseline scores decrease across time; average baseline scores remain stable across time. Cohen's *d* indicates mean-level absolute stability. The RCI suggests no poor individual-level absolute stability because more than half of the individuals (52 %) show significant change across time. The ICC's and Pearson correlation coefficient all indicate acceptable temporal stability, yet the ICC's estimates are slightly smaller than the Pearson correlation coefficient, likely because ICC's assumption of equal variances is violated (McGraw & Wong, 1996).

Based on this simulated example we can infer several relations between the various types of temporal stability. First, the presence of perfect individual-level absolute stability implies both relative stability and mean-level absolute stability (column 1). Second, the presence of perfect mean-level absolute stability does not necessarily imply relative stability or individual-level absolute stability (column 2). Third, the presence of perfect relative stability does not necessarily imply mean-level absolute stability (column 3) or individual-level absolute stability (column 4).

### *2.2.7. Synthesis*

The temporal stability of a construct is arguably not a dichotomous property (i.e., stable or not), but rather a dimensional property with varying degrees of stability. Although for some methods researchers can conclude temporal stability based on a threshold (e.g., a test–retest correlation of at least 0.7), a dimensional conceptualization of temporal

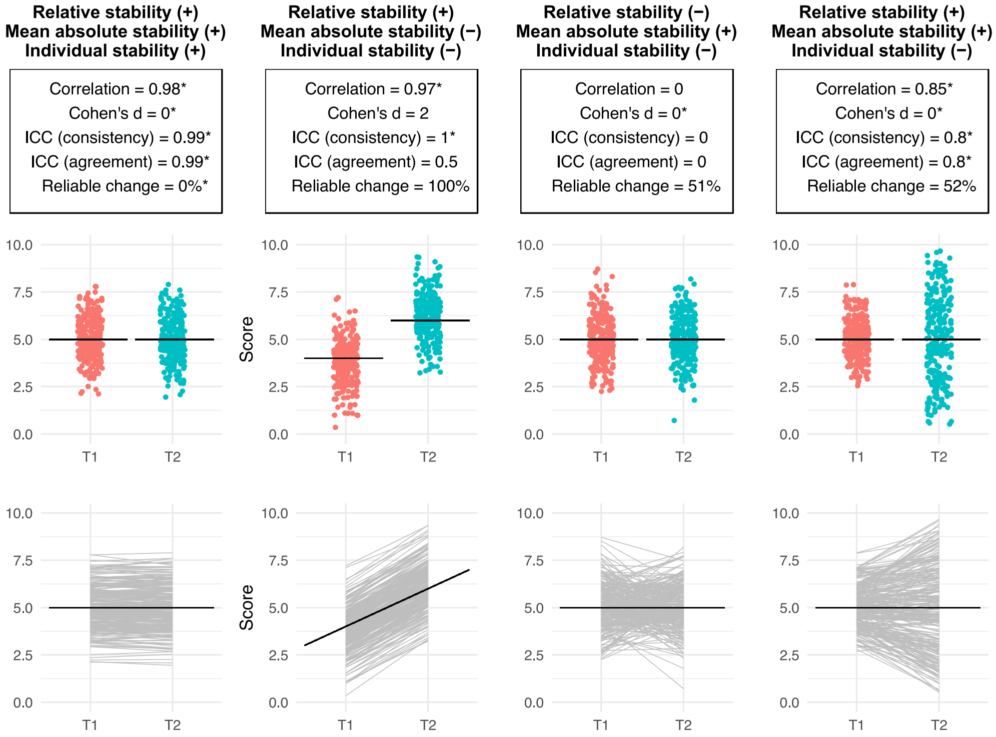

**Fig. 1.** Simulated data on two time points, varying in the presence (+) or absence (-) of relative stability, mean-level absolute stability and individual absolute stability. The upper row shows estimated stability statistics. The middle row shows the individual- and mean scores at each time point. The bottom row shows the individual- and mean growth curves. ICC = intraclass correlation. * Temporal stability concluded based on this statistic.

stability implies a focus on differences in degree of temporal stability. For instance, when two constructs show test–retest correlations of 0.69 and 0.71 in the same population and across the same follow-up time, then researchers may conclude based on a cutoff of 0.70 that one construct is stable and the other not, yet arguably both have a very similar temporal stability.

Several conclusions can be drawn regarding the reviewed studies on the temporal stability of Type D, NA and SI. First, our simulated data example illustrates why researchers should first explicitly define the type of temporal stability they want to assess and subsequently select one or more models that can adequately detect such stability. Table 2 summarizes for each reviewed method the types of temporal stability that can be detected. If researchers want to know whether individuals do not change in their personality across time, then only using a relative stability method is not sufficient and may better be complemented by an absolute stability method.

When answering the question whether NA and SI are stable personality traits, we should therefore focus on studies that have reported separate analyses of relative and absolute stability, or on studies that have used analysis that are sensitive to both types of stability, such as the absolute agreement ICC. Three of the reviewed studies assessed both relative and absolute stability on the same sample (Dannemann et al., 2010; Pedersen et al., 2009; Romppel et al., 2012). Mean-level absolute stability was concluded for NA in all studies and for SI in two of the three studies. Regarding the relative stability of these personality traits, the three studies generally showed less than adequate test–retest

**Table 2**
Characteristics of methods used to assess temporal stability in the reviewed studies.

| Statistical focus | Statistical method | Handles measurement error in<br>item scores | Detects relative stability | Detects mean absolute stability | Detects individual absolute stability | Detects ipsative stability | Detects genetic stability |
|---|---|---|---|---|---|---|---|
| Association | Pearson correlation | – | + | – | – | – | – |
| | ICC (consistency) | – | + | – | – | – | – |
| | ICC (agreement) | – | + | – | + | – | – |
| Mean difference | Paired *t*-test | – | – | + | – | – | – |
| | Cohen’s d | – | – | + | – | – | – |
| | RM ANOVA | – | – | + | – | – | – |
| | Reliable change index | + | – | – | + | – | – |
| Classification | % caseness | – | – | – | – | + | – |
| | Chi-square test | – | – | – | – | + | – |
| | Logistic regression | – | – | – | – | + | – |
| | Cohen’s Kappa | – | – | – | – | + | – |
| Genetic | ACE model | + | – | – | – | – | + |

ACE = statistical model used to assess genetic and environmental influences in twin studies; ICC = intraclass correlation; LGCM = latent growth curve model; LMI = Longitudinal measurement invariance; RM ANOVA = repeated measures analysis of variance.

correlations (*NA*: 0.61, 0.61, 0.85; *SI*: 0.59, 0.60, 0.63). These results suggest that although on average participants did not change in their NA and SI scores over time, the relative ranking of participants showed less temporal stability. Still, the Pearson correlations used to assess this relative ranking likely suffered from attenuation bias because these correlations were not estimated while taking into account the measurement error in the NA and SI item scores. In sum, based on these studies the temporal stability of both NA and SI appears suboptimal.

However, a limitation of the reviewed studies is that the commonly used methods do not consider the properties of the Type D measurement instrument when assessing temporal stability. The DS14 item scores are not equivalent, generally not normally distributed, and more informative in the higher than in the lower NA and SI score ranges (Emons, Meijer & Denollet, 2007). Such characteristics may contribute to a violation of certain model assumptions (e.g., linearity and homoscedasticity) discussed in this study. Another major limitation is that none appropriately tested for longitudinal measurement invariance, though two studies partly investigated this assumption (Romppel et al., 2012; Condén et al., 2014; see Study 2). Longitudinal measurement invariance is often overlooked in longitudinal research and when this assumption is violated then changes on the observed scores do not necessarily reflect changes in the latent constructs of interest.

We therefore argue that the starting point of any temporal stability assessment should be a test for longitudinal measurement invariance. After establishing measurement invariance, relative and absolute stability should ideally be investigated using approaches that adjust for measurement error, such as latent variable modeling. Study 2 illustrates three such longitudinal latent variables models that can be used to investigate the temporal stability of constructs measured with ordinal items that often show skewed score distributions.

## 3. Study 2: Longitudinal latent variable models

A major disadvantage of the temporal stability methods discussed in Study 1 is that they were applied to observed scores and therefore assess change in aggregated item scores (e.g., sum or mean scores) and not in latent construct scores. These observed score methods implicitly assumes that these aggregate scores do not contain measurement error and thus are perfectly reliable measures of the underlying construct at all measurement occasions. However, questionnaire scores generally are imperfect measures of the underlying latent construct. Modern test theory assumes that each questionnaire has measurement properties (e.g., loadings/intercepts/thresholds/residuals in factor models, or discrimination/difficulty parameters in item response models) that relate observed item scores to latent construct(s). Scores on questionnaire items are therefore not exclusively caused by the variation in the latent construct, but also by other factors unique to each particular item.

### *3.1. Longitudinal measurement invariance*

An assumption underlying most absolute stability methods is that these measurement properties do not change over time, a requirement called longitudinal measurement invariance (Pentz & Chou, 1994; Liu et al., 2017). The fact that factor loadings and other measurement properties *can* change over time, implies that absolute changes in item scores (and therefore also in aggregated scores) are not necessarily the result of changes on the construct level, but may also result from changes in measurement properties. When assessing the temporal stability of constructs, researchers should first test for longitudinal measurement invariance to disentangle changes in the measurement properties from changes on construct level.

As an example, consider a sample of participants completing a seven-item negative affectivity questionnaire in both summer and winter. Further suppose that researchers concluded no absolute stability for negative affectivity, because a paired *t*-test indicated significantly lower negative affectivity sum scores in winter than in summer. Lastly, suppose that people did not change in their true negative affectivity scores over time. How is it possible that the paired *t*-test suggested significant change while there was no true change in negative affectivity over time? One reason could be that the intercept of the negative affectivity item ’I often take a gloomy view of things’, was lower in winter than in summer due to people having a lower mood in winter, a phenomenon called the winter blues (Rosenthal, 2012). Such temporary changes in mood involve a different psychological process than scoring high on the personality trait negative affectivity. Using latent variable models to distinguish the latent construct of interest (e.g., negative affectivity) and the individual items measuring it, allows for detecting changes in the item scores due to other influences while the latent negative affectivity scores remain constant. Differences in item intercepts also violate longitudinal measurement invariance (Oort, 2005), further complicating the comparisons of scores over time. Violating this assumption at intercept level would imply that participants scored higher on that particular item in winter than in summer, regardless of their latent negative affectivity score.

Researchers typically test for longitudinal measurement invariance using a series of increasingly restricted structural equation models (SEM; Bollen, 2005). According to these models, each psychological construct is a latent (unobserved) variable and one or more observed item scores reflect the scores on this latent variable. However, each item generally is an imperfect representation of the construct of interest. The variance in an item score not explained by the latent variable is called measurement error or unique variance. By distinguishing the item variance explained by the latent construct from the variance explained by measurement error, latent variable models allow for estimating the association between latent constructs themselves (rather than aggregated observed scores), resulting in estimates that are unaffected by measurement error.

When testing longitudinal measurement invariance using SEM, researchers typically fit a series of nested models to the data. First, a *configural invariance* model is tested to determine whether the factor structure is similar across time. In each step, an additional type of measurement model parameters is constrained to be equal across time (Millsap & Cham, 2012). In the second step, the *weak invariance* model constrains the factor loadings to be equal at each measurement occasion. Next, a *strong invariance* model adds the constraint that either the intercepts (for continuous data) or the thresholds (for ordered categorical data) are equal across time. Lastly, the *strict invariance* model constrains the residual variances for each item to be invariant across time. These four models are nested because each additional constraint builds upon the already existing constraints of a previous model. In structural equation modeling, such nested models can be compared using for instance a chi-square difference test or likelihood ratio test. Such tests indicate whether the additionally imposed constraints cause a significantly worsening in model fit. If so, then longitudinal measurement invariance is violated for the newly constrained parameter type. Follow-up tests for partial measurement invariance can be performed to investigate whether measurement invariance is still plausible for a subset of the parameters that were not invariant across time.

Regarding the temporal stability of the Type D personality traits, Table 1 in Study 1 shows that longitudinal measurement invariance of NA and SI has been investigated in two studies (Romppel et al., 2012; Conden et al., 2014). However, these studies merely showed that the NA and SI factor loadings were invariant across time and did not test for longitudinal invariance of the intercepts/thresholds and residuals. These tests are essential when assessing temporal stability of psychological constructs, because if researchers want to interpret absolute changes over time as resulting from changes at the latent construct level, then at least strong invariance has to be established for continuous scores and strict invariance for ordinal scores (Liu et al., 2019). Given that the two studies included in our review only investigated a weak invariance model, it is not clear whether the observed changes (or stability) in the NA and SI scores are attributable to changes in the NA and SI constructs, or to changes in the unstudied measurement properties (i.e., intercepts, thresholds, residual variances). Moreover, these studies failed to consider the ordered categorical measurement level of the NA and SI items. Treating ordered categorical data as continuous and normally distributed might cause biased parameter estimates in the structural equation model when there are fewer than five answer categories or when the item scores are not normally distributed (Rhemtulla, Brosseau-Liard, & Savalei, 2012). To solve these problems, study 2 illustrates a test for the longitudinal measurement invariance of NA and SI that can adequately handle the skewed ordinal nature of these item scores.

In this second study, we assess the temporal stability of the Type D personality traits NA and SI, and depression and anxiety using various latent variable models. These models do not only allow us to determine the relative and absolute stability of the Type D traits, but also assess their relation to the temporal stability of the related psychological states depression and anxiety. Before estimating the longitudinal latent variable models necessary to answer this research question, we will test the assumption of longitudinal measurement invariance separately for each of these four constructs. The next section introduces each of those three latent variable models.

## 3.2. *Longitudinal latent variable models*

### 3.2.1. *Estimation with ordinal items*

When testing the longitudinal measurement invariance of a measurement instrument (and when fitting latent variable models in general), researchers should first evaluate whether the item scores conform to an ordinal or continuous measurement level. Psychological questionnaires often involve Likert-type data with item options ranging between two and nine response categories. Whether these item scores can be considered ordinal or approximately continuous depends not only on the number of response categories, but also on whether the response are approximately normally distributed. Simulation studies have indicated that normally distributed Likert scale data with five or more response categories can be analyzed as continuous variables (Dolan, 1994). Item scores should be analyzed as ordinal variables if they are not (approximately) normally distributed, or if they result from Likert scales with fewer than five response categories. Ignoring the ordinal nature of item scores and treating them as continuous in subsequent analyses results in biased factor loadings and standard errors, especially when the number of response categories is low or the item score distribution is skewed (Rhemtulla, Brosseau-Liard, & Savalei, 2012).

The structural equation models used to test longitudinal measurement invariance on ordinal item scores involve different parameter types than models based continuous items. For continuous item scores, the strong invariance model evaluates the longitudinal invariance of the item *intercepts* (expected item score when the score on the latent construct is zero), while for ordinal item scores, it tests the longitudinal invariance of the item *threshold* parameters. For an ordinal item with X response categories, there are X-1 estimated threshold parameters that connect the observed ordinal response pattern to an assumed underlying continuous item score. The strong invariance model tests whether the threshold parameters of each ordinal item are invariant across time. Regardless of whether the data are continuous or ordinal, the weak invariance model tests whether the factor loadings are invariant across time and the strict invariance model tests whether the residual variances are invariant across time.

In practice, models based on continuous or ordinal item scores often differ in the method used to estimate the parameters of the structural equation models. For continuous item scores, the model parameters are typically estimated using a full information method such as maximum likelihood (ML) estimation, while the parameters of ordinal item score models are often estimated using limited information methods such as weighted least squares (WLS), diagonally weighted least squares (DWLS) or unweighted least squares (ULS). See Liu and colleagues (2017) for an excellent review of the differences between longitudinal measurement invariance tests based on ordinal or continuous item scores.

### 3.2.2. *Latent growth curve models*

After testing longitudinal measurement invariance, we will use latent growth curve (LGC) models to determine the absolute stability of the NA, SI, depression and anxiety, and to investigate how changes over time in one of these constructs relate to changes in the other constructs. LGC models are a special kind of latent variable model where a longitudinal growth curve is estimated for each individual (Hertzog, Lindenberger, Ghisletta & von Oertzen, 2006). These models use latent variables to express the individual differences in the growth curve parameters (i.e., the intercepts and slopes). Different types of growth curve models have been proposed in the literature. A first distinction is between first- and second-order LGC models. First-order models estimate the latent growth curves directly from the observed data (typically repeated measurements of sum scores). Second-order models estimate the latent growth curves based on other latent variables, where each latent variable has a measurement model that connects it to the observed item scores. A second distinction is between univariate and multivariate LGC models. Univariate models concern the longitudinal change in a single latent construct, whereas multivariate models assess change in two or more constructs simultaneously.

LGC models are comparable to longitudinal multilevel models (Curran, Obeidat & Losardo, 2010). Indeed, the least complex LGC model, the univariate single-order model, is mathematically identical to a multilevel model that allows the intercept and slope (effect of time) parameters to vary across individuals (random regression coefficients), instead of assuming they are similar for each individual (fixed regression coefficients). These random regression parameters can be seen as latent variables because they are unobservable, vary across individuals and are estimated from the observed data. We did not use standard multilevel models to assess the temporal stability of psychological constructs because these do not explicitly model measurement error in the item scores.

Both issues are handled by multivariate second-order latent growth models. Being multivariate, these models allow for studying correlated change (Allemand & Martin, 2017) by testing the association between the growth parameters of multiple constructs. In the context of the present study, it would for instance be interesting to assess how individual change over time (the slope parameter) in the latent constructs depression or anxiety relates to individual change in NA. Earlier longitudinal research suggests that NA fluctuates together with the severity of depressive symptoms, indicating that the NA construct may not be temporally stable due to its sensitivity to changes in mood. However, this study (Marchesi et al., 2014) ignored the presence of measurement error in the item scores and did not to test the longitudinal measurement invariance assumption. Consequently, it could be possible that changes in NA over time were not caused by changes in the NA construct, but rather by changes in the measurement model (e.g., factor loadings).

Although the two Type D personality traits are the primary focus of our study, in this study we compare their temporal stability to that of the related constructs depression and anxiety. We first used four univariate second-order latent growth curve models to determine the absolute temporal stability of NA, SI, depression and anxiety. The test whether the average latent slope parameter differs from zero indicates whether a construct shows absolute stability across all participants. The test whether the variance of the latent slope differs from zero indicates whether there are individual differences between individual in the change in the construct over time. Subsequently, we used six multivariate second-order latent growth models to investigate how the change over time in each latent construct relates to change in the other constructs. Fig. 2 visualizes the multivariate growth model of NA and depression. In this model, the correlation between the latent slopes of two constructs indicates to what extent individuals show similar change over time on both constructs.

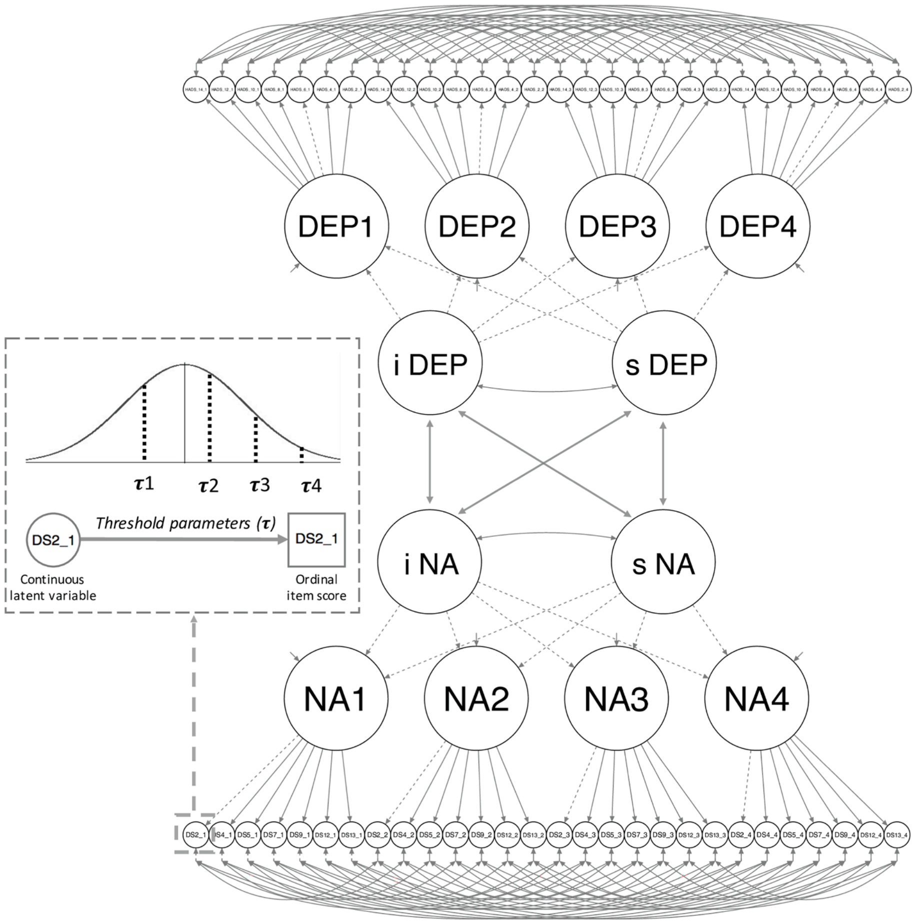

**Fig. 2.** Multivariate second-order latent growth curve model for negative affectivity and depression. Ordinal item scores are modeled using threshold parameters (τ), mapping for each item the observed ordinal response pattern to an assumed continuous normally distributed latent variable. Change in the latent variables scores across the four time points is modeled using higher-order intercept (i) and slope (s) latent variables. The residuals of the same item at different time points are allowed to correlate. Dotted lines represent fixed parameters and solid lines represent freely estimated parameters.

### 3.2.3. *Latent trait-state-occasion models*

Latent trait-state-occasion (LTSO) models are latent variable models used to estimate what proportion of the variance in longitudinal scores can be seen as a stable trait and what proportion as a changeable state. LTSO models are related to various other latent variable models with a similar purpose (e.g., the trait-state-error model or the state-trait model with autoregression), but simulation studies found LTSO models to outperform these other models in decomposing the trait-state variance when the constructs are highly correlated across time (Cole, Martin & Steiger, 2005). LTSO models assume that a latent state is at each repeated measurement occasion explained by two sources of variance: (1) a shared latent trait variable similar across all measurements; (2) a time-specific latent occasion variable at each measurement. Like LGC models, there exist single-order and second-order LTSO models. We will use a second-order LTSO model to the repeated measurements of NA, SI, depression and anxiety, to determine for each of these constructs the proportion of variance that can be considered trait or state, while dealing with the skewed ordinal nature of the item scores. Fig. 3 visualizes an example of such a second-order LTSO model fitted on four repeated measurements of a latent construct.

## 3.3. *Method*

### 3.3.1. *Participants*

Data were used from a study conducted using the PROFILES registry (van de Poll-Franse et al., 2011). This population-based longitudinal cohort study assessed patient reported outcomes of colorectal cancer survivors. Eligible participants included all colorectal cancer patients (Stage I to IV) admitted to hospitals in the southern part of the Netherlands between 2000 and 2009. Full details on the inclusion and exclusion criteria and the data collection can be found online (https://www.dataarchive.profilesregistry.nl/study_units/view/22). The data collection was approved by the ethics committee of the Maxima Medical Centre in Veldhoven, the Netherlands (approval number 0822). An informed consent statement was signed by all participants. In the present study, we included all participants who completed the psychological questionnaires on at least one time point. Measurements were performed yearly starting in 2010 and ending in 2013. At baseline, the 2625 colorectal cancer survivors were on average 69.4 years old (SD = 9.5, range = 29 to 86). A larger percentage of survivors identified as male (55.1 %) than as female (44.9 %). On average, they participated 5.2 years since diagnosis (SD = 2.8, range = 1 to 11 years). Part of the sample was lost to follow-up, with 75.8 %, 55.5 % and 45.3 % of the participants responding at the second, third and fourth measurement occasion, respectively. Earlier research on this dataset has indicated that dropouts were significantly more likely to be female, have older age, a lower education and socio-economic status, and show more depressive symptoms than full responders (Ramsey et al., 2019).

### 3.3.2. *Measures*

*3.3.2.1. Type D personality.* The DS14 questionnaire (Denollet, 2005) was used to measure NA and SI, the two traits underlying Type D personality. Each trait was measured with seven items on a five-point Likert scale, ranging from “false” (0) to “true” (4). The DS14 has been validated in several populations, including the general population (Denollet, 2005) and a breast cancer population (Batselé et al., 2017). Item scores should be considered as having an ordered categorical measurement level, as they are often slightly positively skewed. In the current sample, the estimated McDonald’s Omega (total) based on the polychoric

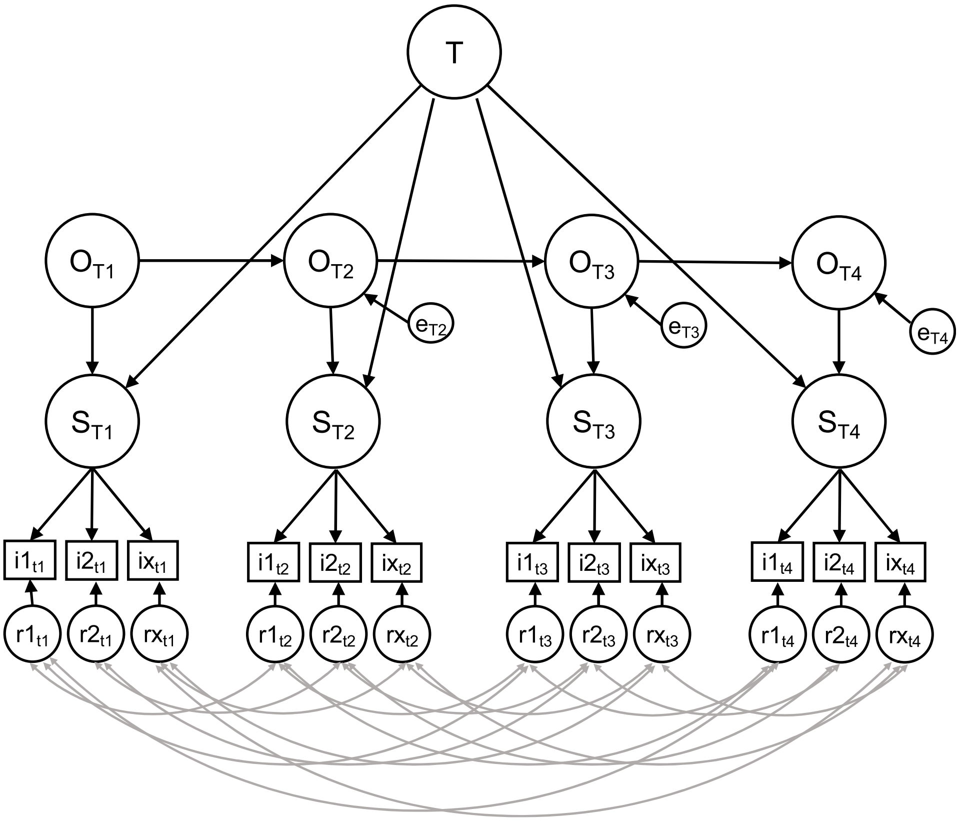
T
O$_{T1}$
O$_{T2}$
O$_{T3}$
O$_{T4}$
e$_{T2}$
e$_{T3}$
e$_{T4}$
S$_{T1}$
S$_{T2}$
S$_{T3}$
S$_{T4}$
i1$_{t1}$ i2$_{t1}$ ix$_{t1}$
i1$_{t2}$ i2$_{t2}$ ix$_{t2}$
i1$_{t3}$ i2$_{t3}$ ix$_{t3}$
i1$_{t4}$ i2$_{t4}$ ix$_{t4}$
r1$_{t1}$ r2$_{t1}$ rx$_{t1}$
r1$_{t2}$ r2$_{t2}$ rx$_{t2}$
r1$_{t3}$ r2$_{t3}$ rx$_{t3}$
r1$_{t4}$ r2$_{t4}$ rx$_{t4}$

**Fig. 3.** Second-order latent trait-state-occasion model fitted on four repeated measurements of a latent construct. T = Trait; O = Occasion; S = State; T1-T4 = repeated measurements; i = observed item score; e = prediction error variance; r = measurement error variance; The curved arrows indicate that the measurement errors of the same item are allowed to correlate over time.

correlation matrices of the NA and SI item scores suggested adequate reliability estimates at each of the four measurement occasions (NA = [0.92, 0.93, 0.93, 0.93]; SI = [0.91, 0.91, 0.91, 0.90]).

*3.3.2.2. Symptoms of depression and anxiety.* The Hospital Anxiety and Depression questionnaire (HADS; Zigmond & Snaith, 1983) was used to measure symptoms of anxiety and depression. Each construct was measured with seven items on a four-point Likert scale, ranging from 0 to 3. These items should be considered ordered categorical scores as they are generally positively skewed and have only four response categories. The HADS questionnaire has been validated in several populations, including in the general population (Spinhoven et al., 1997) and several cancer populations (Bjelland, Dahl, Haug, & Neckelmann, 2002). In the current sample, the estimated McDonald's Omega (total) based on the polychoric correlation matrices of the depression and anxiety item scores suggested adequate reliability estimates at each of the four measurement occasions (Depression = [0.89, 0.88, 0.88, 0.89]; Anxiety = [0.89, 0.88, 0.88, 0.89]).

### *3.3.3. Statistical analyses*

We used to the R-package lavaan for all latent variable models (Version 0.6–4; Rosseel, 2012). The questionnaires used to measure these constructs involve Likert scale items with either 3 or 4 response categories and typically result in positively skewed item scores (see ***Appendix B***). Therefore, in each latent variable model these ordinal item scores will be modeled using threshold parameters and the models will be estimated using DWLS estimation. P-values smaller than 0.05 were considered statistically significant.

*3.3.3.1. Longitudinal measurement invariance.* The R-scripts used to test for longitudinal measurement invariance were based on the scripts reported by Liu and colleagues (2017). Testing for measurement invariance involves fitting a series of nested models, starting with a baseline model. In the present study, this baseline model is a correlated four-factor model, where each factors represent the latent construct at one of the repeated measurements. In each of the three subsequent models a new type of structural equation model parameter is constrained to be invariant across time. This invariance constraint applies to the factor loadings in the weak invariance model, to both the factor loadings and thresholds in the strong invariance model, and to the factor loadings, thresholds and residual variances in the strict invariance model. The chi-square difference (likelihood ratio) test was used to determine whether the more constrained model more poorly fitted the data than the lesser constrained model. Given that the study's large sample size will make even the smallest deviations from measurement invariance statistically significant, we also evaluated the change in the fit indices RMSEA, SRMR and CFI. When the decrease in model fit of the more constrained model was$<$0.015 (RMSEA), 0.030 (SRMR) or 0.002 (CFI), the newly induced longitudinal invariance constraints were considered to show an acceptable fit to the data (Chen, 2007; Meade, Johnson & Braddy, 2008). Fit of the baseline models was evaluated based on the RMSEA ($<$0.07), SRMR ($<$0.10) and CFI (greater than0.95). Missing data on the repeated measurements were handled using available case analysis (pairwise deletion).

*3.3.3.2. Relative stability.* After investigating longitudinal measurement invariance, relative stability was investigated by inspecting the correlations between the repeated measures of each latent variable. Another model was fitted to assess relative stability from a different perspective by regressing the latent variable scores at the T2, T3, and T4 on the latent variables score at the preceding time point. The standardized regression coefficients of these autoregressive effects indicate the extent

**Table 3**
Fit statistics for models testing the longitudinal measurement invariance of the DS14 (negative affectivity & social inhibition) and HADS (depression & anxiety).

| Model | df | Δdf | $\chi^2$ | $\Delta\chi^2$ | RMSEA (95 %CI) | ΔRMSEA | SRMR | ΔSRMR | CFI | ΔCFI |
|---|---|---|---|---|---|---|---|---|---|---|
| *Negative affectivity* | | | | | | | | | | |
| 1 | 320 | – | 998.5 | – | 0.035 [0.033 0.037] | – | 0.034 | – | 0.998 | – |
| 2 | 338 | 18 | 1016.4 | 22.5 | 0.033 [0.031 0.035] | 0.002 | 0.034 | <0.001 | 0.998 | <0.001 |
| 3 | 380 | 42 | 1064.0 | 74.3* | 0.031 [0.029 0.033] | 0.002 | 0.034 | <0.001 | 0.998 | <0.001 |
| 4 | 401 | 21 | 1186.2 | 49.1* | 0.028 [0.026 0.029] | 0.003 | 0.035 | 0.001 | 0.998 | <0.001 |
| *Social inhibition* | | | | | | | | | | |
| 1 | 320 | – | 3338.8 | – | 0.068 [0.066 0.070] | – | 0.052 | – | 0.992 | – |
| 2 | 338 | 18 | 3352.6 | 18.7 | 0.065 [0.063 0.067] | 0.003 | 0.052 | <0.001 | 0.992 | <0.001 |
| 3 | 380 | 42 | 3412.1 | 97.3* | 0.061 [0.059 0.063] | 0.004 | 0.052 | <0.001 | 0.992 | <0.001 |
| 4 | 401 | 21 | 3783.2 | 120.6* | 0.053 [0.051 0.054] | 0.008 | 0.054 | 0.002 | 0.991 | 0.001 |
| *Depression* | | | | | | | | | | |
| 1 | 299 | – | 382.0 | – | 0.018 [0.016 0.021] | – | 0.028 | – | 0.999 | – |
| 2 | 317 | 18 | 400.3 | 22.2 | 0.018 [0.015 0.020] | <0.001 | 0.028 | <0.001 | 0.999 | <0.001 |
| 3 | 359 | 42 | 433.9 | 41.9 | 0.016 [0.014 0.018] | 0.002 | 0.029 | 0.001 | 0.999 | <0.001 |
| 4 | 380 | 21 | 497.6 | 28.4 | 0.014 [0.012 0.017] | 0.002 | 0.030 | 0.001 | 0.999 | <0.001 |
| *Anxiety* | | | | | | | | | | |
| 1 | 299 | – | 759.4 | – | 0.030 [0.028 0.033] | – | 0.037 | – | 0.997 | – |
| 2 | 317 | 18 | 767.5 | 11.8 | 0.029 [0.027 0.031] | 0.001 | 0.037 | <0.001 | 0.997 | <0.001 |
| 3 | 359 | 42 | 820.9 | 65.7* | 0.027 [0.025 0.029] | 0.002 | 0.037 | <0.001 | 0.997 | <0.001 |
| 4 | 380 | 21 | 899.6 | 31.8 | 0.023 [0.021 0.025] | 0.004 | 0.039 | 0.002 | 0.996 | 0.001 |

Model 1 (Configural invariance): Baseline model.
Model 2 (Weak invariance): Invariant factor loadings.
Model 3 (Strong invariance): Invariant factor loadings & thresholds.
Model 4 (Strict invariance): Invariant factor loadings, thresholds & residuals.
[*] $p < .05$ on the scaled chi-squared difference test (Satorra, 2000). Note that these scaled differences are larger than the raw chi-square differences.

to which the latent variables scores at a certain time point can be predicted from the score on a preceding time point (Borghuis et al., 2017).

*3.3.3.3. Latent growth curve models.* We fitted a univariate growth model for each latent construct and six multivariate growth models for the six pairs of two constructs. The first longitudinal measurement invariance model was used as a baseline model. Assuming conditional independence, the correlations between each construct's repeated latent measurements were fixed to zero and latent intercept and linear slope variables were added to model individual differences in change over time. The factor loadings of these latent growth parameters on the four repeated measurements of each latent construct were fixed to 1 for the latent intercept and to 1, 2, 3, and 4 for the latent slope. In the multivariate models, the correlations between the latent intercepts and slopes of both constructs were freely estimated. Identifying the multivariate growth model required several parameter constraints. First, for each construct the variances of the latent variables were constrained to be equal across the four time points.

(homogeneity of variances). Second, the correlation between the two latent constructs was constrained to be equal at each time point. Given that second-order multivariate latent growth models require estimation of many parameters, a robustness test was performed by determining the correlation between the latent slopes of two constructs in multivariate *first*-order growth models based on the observed scores of both constructs (i.e. the sum of the item scores at each time point).

*3.3.3.4. Latent Trait-State-Occasion models.* A separate LTSO model was fitted for each of the four constructs. As previously, the first longitudinal measurement invariance model was used as a baseline model. For each construct, the correlations between the repeated latent measurements were fixed to zero. To identify the LTSO model, for each latent construct, the latent variable correlations and the state variance at each measurement occasion was fixed to zero (Cole, Martin & Steiger, 2005). Next, for each latent state the factor loadings on both the shared latent trait and the time-specific latent occasion were fixed to one. Autoregressive effects between the four time-specific latent occasion variables were freely estimated (Gana & Broc, 2019). The decomposition into trait and state variance was calculated by dividing respectively the latent state or latent occasion variance at baseline by the sum of these two variances.

### *3.3.4. Transparency and openness*

PROFILES registry data is freely available according to the FAIR (Findable, Accessible, Interoperable, Reusable) data principles for non-commercial (international) scientific research, subject only to privacy and confidentiality restrictions. Data is made available through Questacy (DDI 3.x XML) and can be accessed at (https://www.profilesregistry.nl). The R-scripts for all analyses reported in this article can be found on the Open Science Framework (https://osf.io/e7ajr/?view_only=41e3de13bb2e4a2eaa4168f0e124fdcc). This study's design and analyses were not preregistered.

## *3.4. Results*

### *3.4.1. Longitudinal measurement invariance*

Of all 2625 participants, data from 2597 (99.4 %) participants were available for the NA models, 2604 (99.7 %) participants for the SI models, 2603 (99.7 %) for the Depression models and 2602 (99.6 %) for the Anxiety models. Table 3 presents the fit statistics for the models used to test the longitudinal measurement invariance of the DS14 (negative affectivity & social inhibition). For comparison we also included results of the HADS (depression & anxiety). For all constructs, the chi-square statistic of the baseline model was statistically significant, suggesting a strict mismatch between the observed and model implied covariance matrices. However, as the chi-square test is sensitive to large sample sizes, model fit was also evaluated using the RMSEA, SRMR and CFI. Based on these fit indices all baseline models showed adequate fit to the data.

The next step in the longitudinal measurement invariance procedure is to introduce the constraint that the factor loadings of each latent construct are invariant across time (*weak invariance model*). For all constructs, the chi-square difference test was not significant, indicating that this constraint did not lead to deterioration in model fit. This result was corroborated by the small changes in RMSEA, SRMR and CFI compared to the baseline model.

The third step is to constrain the thresholds of each item to be invariant across time (*strong invariance model*). Based on the chi-square difference test, only the depression.

**Table 4**
Across four measurements of negative affectivity, social inhibition, depression and anxiety, the relative stability, autoregression coefficients, and latent variable mean differences relative to the first time point, including 95% confidence intervals.

| Latent construct | Negative affectivity | Social inhibition | Depression | Anxiety |
|---|---|---|---|---|
| *Latent variable correlations* | | | | |
| $r_{(T1,\ T2)}$ | 0.780 (0.748, 0.812) | 0.803 (0.778, 0.828) | 0.821 (0.785, 0.857) | 0.825 (0.795, 0.855) |
| $r_{(T1,\ T3)}$ | 0.736 (0.696, 0.776) | 0.819 (0.792, 0.847) | 0.805 (0.760, 0.851) | 0.820 (0.783, 0.857) |
| $r_{(T1,\ T4)}$ | 0.733 (0.678, 0.779) | 0.823 (0.795, 0.850) | 0.767 (0.712, 0.822) | 0.791 (0.749, 0.834) |
| *Autoregression coefficients* | | | | |
| $\beta_{(T1,\ T2)}$ | 0.845 (0.822, 0.867) | 0.894 (0.875, 0.912) | 0.873 (0.849, 0.897) | 0.885 (0.860, 0.910) |
| $\beta_{(T2,\ T3)}$ | 0.879 (0.857, 0.900) | 0.929 (0.915, 0.944) | 0.927 (0.907, 0.948) | 0.932 (0.911, 0.952) |
| $\beta_{(T3,\ T4)}$ | 0.893 (0.868, 0.919) | 0.928 (0.911, 0.946) | 0.904 (0.877, 0.931) | 0.915 (0.891, 0.939) |

**Table 5**
Fit indices and individual change in negative affectivity, social inhibition, depression and anxiety in terms of the mean and variance of the latent intercept and slope.

| | Negative affectivity | Social inhibition | Depression | Anxiety |
|---|---|---|---|---|
| *Model fit* | N = 2597 | N = 2604 | N = 2602 | N = 2601 |
| Free parameters | 211 | 211 | 183 | 183 |
| $\chi^2$ | 1262.3* | 4016.8* | 565.5* | 1020.0* |
| RMSEA (95 %CI) | 0.035 [0.033, 0.037] | 0.069 [0.067, 0.071] | 0.018 [0.016, 0.021] | 0.030 [0.028, 0.032] |
| SRMR | 0.034 | 0.052 | 0.029 | 0.037 |
| CFI | 0.987 | 0.962 | 0.995 | 0.987 |
| *Latent growth parameters* | | | | |
| Mean Intercept | −0.253* | −0.291* | −0.527* | −0.401* |
| Variance Intercept | 0.455* | 1.111* | 1.626* | 1.798* |
| Mean Slope | −0.004 | 0.027 | 0.059* | 0.035 |
| Variance Slope | 0.011* | 0.009 | 0.030* | 0.029* |

* p <.05.

thresholds appeared to be invariant, yet this test is very sensitive with large sample sizes. The effect size of this invariance violation can be evaluated by computing for each item the change in the estimated response probabilities when freely estimating the thresholds across time, compared to constraining them to be equal. These changes in response probability never exceeded 0.05, suggesting that different item responses than those observed are expected only for a small percentage of participants. Based on guidelines by Liu and colleagues (2017), a change in response probability smaller than 0.05 should be no reason for concern. ***Appendix C*** shows that none of these estimated probabilities exceeded 0.05, indicating that the effect sizes of the invariance violations were very small. Moreover, for all constructs the change in RMSEA, SRMR and CFI relative to the weak invariance model was very small (i.e., < 0.005), suggesting that these thresholds can be considered invariant across time.

The last step is to constrain the residual variance of the items to be invariant across time (*strict invariance model*). Based on the chi-square difference test this constraint adequately fitted the data of the depression and anxiety models, but did not fit the data for the SI and NA models. However, again the changes in RMSEA, SRMR and CFI are small compared to the strong invariance model, suggesting that the residual variances can also be considered invariant across time.

The current findings indicate that it is safe to assume that the properties of the instruments used to measure NA, SI, depression and anxiety are invariant across time. The evidence in favor of longitudinal measurement invariance renders it unlikely that longitudinal changes in the observed scores resulted from changes in the properties of the measurement instrument. Consequently, any reported change (or absence of change) can be interpreted as change in the latent construct.

The first rows of Table 4 report for NA, SI, depression and anxiety, the correlations between the baseline estimate of each construct and the estimates at the three later time points. As these estimates concern the correlation between the latent constructs, they are, opposed the correlations between total questionnaire scores, uncontaminated by the presence of measurement error in the item scores. The correlations suggest that the relative stability of these constructs was moderate to high (Shrout, 1998). Interestingly, as the time interval between the repeated measurements increased from one to two and three years, the relative stability of NA, depression and anxiety decreased, while it slightly increased for SI. Lastly, Table 4 also reports the autoregression coefficients, indicating high rank order stability in the latent variable scores at each time point when regressed on scores at a preceding time point.

### 3.4.2. Latent growth curve models

When assessing absolute temporal stability using latent growth curve models, the mean of the latent slope indicates whether there is absolute stability averaged across all participants, whereas the variance of the latent slope indicates whether this absolute stability is identical for all individuals. Concluding perfect absolute temporal stability would require that both the mean and variance of the estimated latent slope are equal to zero.

Table 5 shows the fit indices for each of the four univariate growth models, as well as results of the Wald tests indicating whether the mean and variance of the latent intercepts and slopes differed significantly from zero. Though each of the four univariate growth models showed misfit based on the significant chi-square test (which is sensitive to misfit in large sample sizes), the RMSEA, SRMR and CFI all suggested good model fit. Fig. 4 shows the estimated individual growth curves for NA, SI, depression and anxiety. The red and blue curves represent decreasing and increasing individual trends respectively. The black line indicates the mean estimated latent growth curve across all participants. These plots suggest that on average there were no large changes in these latent constructs over time. Indeed, Table 5 shows that only the average latent slope of depression differed significantly from zero, with a slight increase in depression over time. Based on these findings, the two Type D personality traits NA and SI showed absolute stability when change over time was averaged across all participants. Mean-level stability was also found for anxiety, but not for depression.Fig. D1..

The estimated variance of the latent intercept and slope indicate individual deviations from the average intercept and slope. For all four constructs, the variance of the latent intercept was significantly larger than zero, showing that there were significant individual differences in the baseline scores of the four constructs. For NA, depression and anxiety, the estimated variance of the latent slope was small, yet significantly larger than zero, suggesting that for these constructs the estimated individual growth curves deviated from the mean latent slope. Although participants did on average not change in NA and anxiety over time, the significant estimated slope variance indicated that a considerable number of individuals deviated from this pattern showing either positive or negative change over time.

Table 6 reports the estimated correlations between the latent slopes of the four constructs according to both first- and second-order

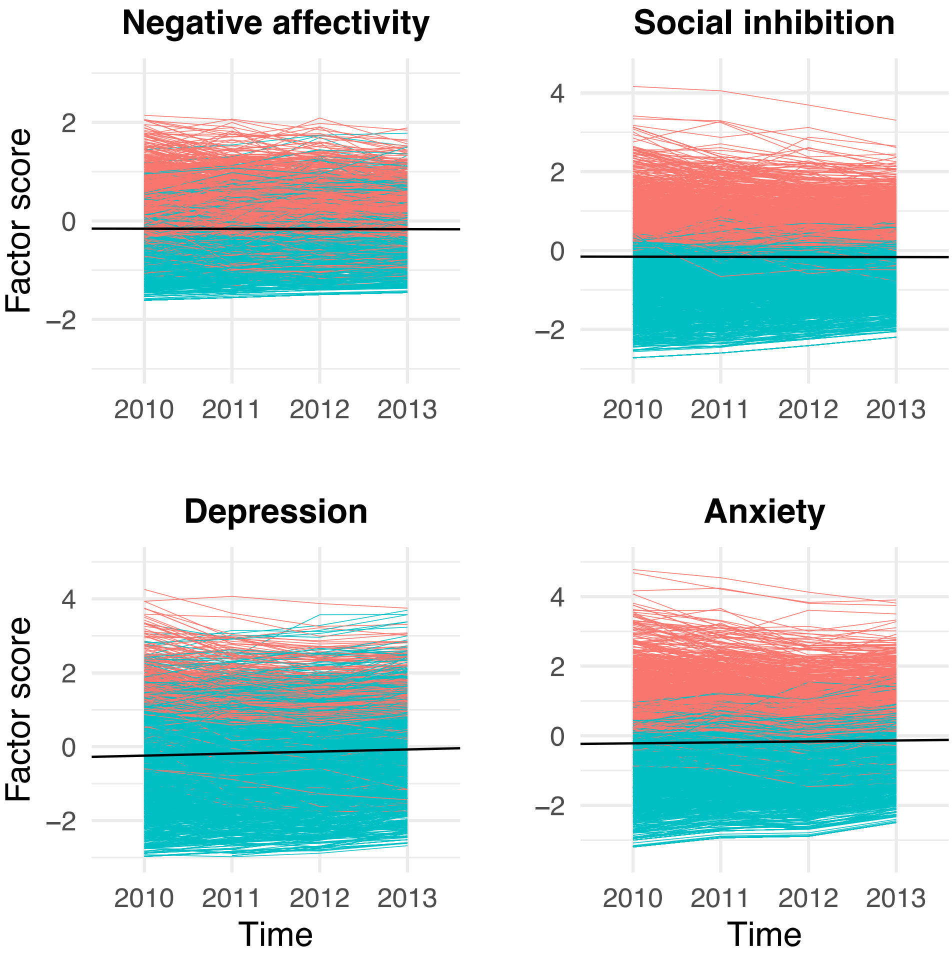

**Fig. 4.** Estimated individual growth curves for negative affectivity, social inhibition, depression and anxiety. Red and blue curves represent decreasing and increasing individual trends respectively. The black line indicates the mean latent slope across all participants.

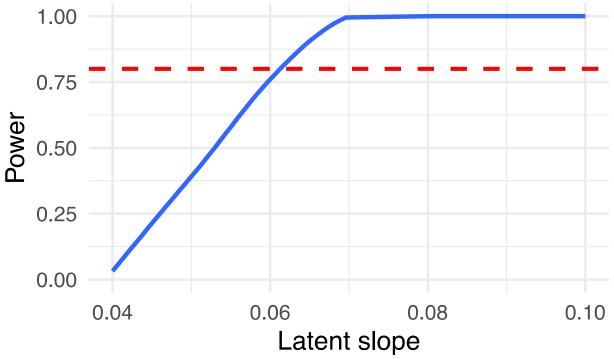

**Fig. D1.** The statistical power to detect a latent slope of particular size, given a significance level of 0.05 and a sample size of 2600 participants.

**Table 6**
Estimated correlation (95% confidence interval) between the latent slopes of negative affectivity, social inhibition, depression and anxiety, according to both first- and second-order multivariate latent growth models.

| Latent slope correlation | First-order growth model | Second-order growth model |
|---|---|---|
| NA & SI | 0.38 (-0.01, 0.78) | 0.30 (-0.10, 0.70) |
| NA & Depression | 0.55 (0.31, 0.79)* | 0.49 (0.14, 0.85)* |
| NA & Anxiety | 0.69 (0.45, 0.92)* | 0.51 (0.22, 0.80)* |
| SI & Depression | 0.56 (0.20, 0.92)* | 0.57 (-0.34, 1.00) |
| SI & Anxiety | 0.51 (0.14, 0.89)* | 0.56 (-0.38, 1.00) |
| Depression & Anxiety | 0.66 (0.50, 0.82)* | 0.71 (0.36, 1.00)* |

[*] p <.05.

multivariate latent growth models. First-order growth models do not handle measurement error in the item scores, but served as a robustness test. In general, both models resulted in similar estimates of the correlation between the latent slopes. The weakest slope correlation was found for NA and SI ($r = 0.30$, $p = .142$) and the strongest for depression and anxiety ($r = 0.71$, $p < .001$). Interestingly, the NA slopes correlated substantially with the slopes of both depression ($r = 0.49$, $p = .007$) and anxiety ($r = 0.51$, $p = .001$). Similar estimates were found for the slope correlation of SI with depression ($r = 0.57$, $p = .221$) and anxiety ($r = 0.56$, $p = .243$), though these estimates failed to reach significance due to the very large standard errors. All growth models involving SI resulted in very broad confidence intervals for the correlation between the latent slopes. Although we are not entirely sure how to explain this, it may have been caused by the low variability in the latent SI slopes, as shown by the result of the univariate growth models reported in Table 5.

In sum, the latent growth curves models showed that averaged across all participants, the two Type D personality traits showed absolute stability, as was the case for anxiety. The mean estimated slope of depression differed significantly from zero, yet indicated only a slight increase in depression over time. The variability estimates of the latent slopes suggested that for NA, depression and anxiety the absolute stability did not apply equally to every individual. Lastly, the multivariate growth models revealed that change in these constructs over time was correlated. These results suggest that although NA and SI are personality traits, especially NA appears to covary with changes in depression and

**Table 7**
Fit indices and trait/state variance proportions (95 %CI) for the second-order latent trait-state-occasion models of negative affectivity, social inhibition, depression and anxiety.

| | Negative affectivity | Social inhibition | Depression | Anxiety |
|---|---|---|---|---|
| *Model fit* | | | | |
| N | 2597 | 2604 | 2602 | 2601 |
| Free parameters | 186 | 186 | 158 | 158 |
| $\chi^2$ | 1059.1* | 3533.5* | 565.5* | 868.4* |
| RMSEA (95 %CI) | 0.028 [0.026, 0.030] | 0.059 [0.058, 0.061] | 0.018 [0.016, 0.021] | 0.024 [0.022, 0.026] |
| SRMR | 0.036 | 0.061 | 0.029 | 0.039 |
| CFI | 0.991 | 0.967 | 0.995 | 0.991 |
| *Proportion explained variance* | | | | |
| Latent trait | 0.74 [0.68, 0.80] | 0.83 [0.79, 0.87] | 0.76 [0.69, 0.83] | 0.78 [0.72, 0.83] |
| Latent occasion | 0.26 [0.20, 0.32] | 0.17 [0.13, 0.21] | 0.24 [0.17, 0.31] | 0.22 [0.17, 0.27] |

* p <.05.

anxiety. It seems that part of the NA construct is a stable personality trait, while another part behaves more state-like and is susceptible to internal and external influences. However, the models presented so far do not speak to the extent to which the constructs are traits or states. This last part of this article will investigate this using a Latent trait-state-occasion model.

### 3.4.3. Latent Trait-State-Occasion models

Table 7 presents for the LTSO models of NA, SI, depression and anxiety the fit indices and variance estimates of the latent state and trait variables. Similar to the LGC models, each of the four LTSO models showed misfit based on the significant chi-square test (which is very sensitive to misfit under large sample sizes), while the RMSEA, SRMR and CFI all suggested good model fit. For each of the four constructs, after partialling out the measurement error variance, the estimated variance proportions corresponded more to a stable trait than to a changeable state. SI turned out to be most trait-like (83 %), followed by anxiety (78 %), depression (76 %), an NA (74 %). However, because the confidence intervals of these percentages showed overlap, the differences between these constructs are likely not statistically significant.

### 3.4.4. Models assuming continuous item scores

As a sensitivity analysis, ***Appendix E*** reports the fit indices and estimates of the longitudinal measurement invariance analysis, the LGC model and the LTSO model, when treating the ordinal item scores as continuous variables. In general, the fit of these continuous models is worse than that of the ordinal models. Longitudinal measurement invariance is still established, but the growth models for continuous item scores did not fit the data very well. Lastly, the estimated proportion of trait variance was slightly larger for depression and anxiety when those ordinal item scores were assumed to be continuous.

## 4. General discussion

In this study, we reviewed methods commonly used to assess the temporal stability of psychological constructs and focused on Type D as an example. Furthermore, we illustrated how to test the assumption of longitudinal measurement invariance and how to assess temporal stability using various latent variable models that handle skewed and ordinal nature of item scores and measurement error. Based on simulated data we illustrated what types of temporal stability can be detected by several commonly used statistical models. We recommend researchers to explicitly report the type of temporal stability they are interested in and then select a statistical model that can detect such temporal stability. If the researcher is not interested in a specific type of temporal stability, then we recommend them to use multiple stability methods to comprehensively assess individual differences in the change on a construct across time.

In Study 1, our review of temporal stability methods used in the literature on Type D personality covered 25 studies that jointly reported 76 tests for the temporal stability of either Type D personality or its underlying personality traits NA or SI. The review concluded that the temporal stability of both NA and SI was less than optimal based on studies investigating both the relative and absolute stability. The stability methods encountered in the Type D literature failed to account for measurement error when estimating the relative and absolute stability, thereby risking attenuated stability estimates. Furthermore, the reviewed studies did not test the assumption of longitudinal measurement invariance. When this assumption is violated (or not investigated) researchers cannot exclude the possibility that any observed changes over time were merely caused by changes in the measurement instrument, rather than by change in the psychological construct.

In Study 2, we showed how to handle these issues using various kinds of latent variable models. We assessed both the relative and absolute stability of the personality traits NA and SI and the psychological states depression and anxiety over a period of four yearly measurements. First, we illustrated how latent variable models can take into account the often skewed and ordinal nature of the item scores measuring these constructs. Next, we showed that the assumption of longitudinal measurement invariance was met for all constructs of interest in the current sample of colorectal cancer survivors. Because this assumption was met, any observed change (or stability) in questionnaire scores could be interpreted as being caused by the construct, rather than by the measurement instrument.

Based on the latent variable models, we concluded moderate to good relative stability for NA, SI, depression and anxiety, based on guidelines by Shrout (1998). The four-year relative stability estimates were lowest for NA and highest for SI. This finding is in line with the relative stability estimates discussed in our review, with NA showing a slightly lower median relative stability than SI. Our estimates were often higher than those seen in our review. This may be explained by the fact that the reviewed studies did not adjust for measurement error, thereby risking an underestimation of the true relative stability.

The univariate LGC models indicated absolute stability for NA, SI, and anxiety. Absolute stability could not be concluded for depression, yet the significant increase in depression over time was small. Earlier research on the current dataset has shown that dropouts were more likely to have high depressive symptoms than full responders (Ramsey et al., 2019). This suggests that the depressive symptoms at later measurement occasions may be overestimated. Indeed, our findings indicate that of all four psychological constructs, only depression shows a significantly positive latent slope, suggesting that on average these participants increased in their depression during the four-year follow-up. However, this estimate should be interpreted with care, as it is possible that without attrition this latent slope estimate for depression would have been closer to zero, similar to the slopes of anxiety, NA and SI.

SI was the only construct with both absolute stability and no significant individual differences in this stability over time. Although NA and anxiety were on average stable over time, these constructs showed significant individual deviations from this absolute stability on group level. The multivariate LGC models revealed that these significant individual differences in change over time correlated between the constructs. In line with our expectations, individual changes in NA moderately correlated with changes in anxiety and depression. This suggests that NA is not entirely a stable personality trait, but may also in part be susceptible to changes in an individual's life, such as an increase or decrease in depression or anxiety. These findings resonate with earlier research by Ossola et al. (2015) who used a repeated measures ANOVA to show that the observed DS14 (NA) sum scores covaried over time with the HADS-D (depression) sum scores. That study also used an exploratory factor analysis on the DS14 and HADS items, revealing that the NA and depression items loaded on the same factor, while the SI and anxiety items all loaded on separate factors.

The LGC models suggested that the NA construct may in part reflect the episodic or transient distress also reflected in psychological states such as anxiety and depression.

The LTSO models highlighted the extent to which each of those constructs can be considered a stable trait or a changeable state. All constructs were more trait than state, with SI being most trait like (83 %), followed by anxiety (78 %), depression (76 %) and NA (74 %). The finding that SI corresponds more to a stable trait than NA is in line with our findings regarding the absolute and relative stability of these constructs. Interestingly, together with NA, depression and anxiety also turned out to be less trait-like than SI. According to Baltes (1987) both stable and unstable processes underlie most psychological constructs. Personality traits are known to become less stable as the time between the measurements increases (Roberts & DelVecchio, 2000). Other studies have indicated that both trait and state processes underlie constructs such as depression (Hartlage, Arduino, & Alloy, 1998) and anxiety (Kantor et al., 2001). Our finding is partly in line with an earlier study showing that anxiety and the personality traits behavioral inhibition and neuroticism are more trait-like than depression, which was found to be more episodic in nature (Prenoveau et al., 2011).

One explanation for this unexpected finding is methodological. In the LTSO model, the stable latent trait variable captures the individual differences in the construct that remain unchanged across time, whereas the time-specific latent occasion variables capture the individual differences in the construct that are unique at each time point. In theory, when none of the participants in a dataset change over time, then the LTSO model would indicate that the construct is a completely stable trait. This could also happen in case of floor effects, when there is so little anxiety or depression in the population that most participants show very low scores on these measurements. This implies that the conclusions resulting from LTSO models should always be interpreted in light of the characteristics of the dataset. If the participants in the current dataset would have shown more changes in depression and anxiety over time, then the LTSO models would likely have estimated a smaller proportion of trait variance. It would therefore be interesting for future research to apply the LTSO models to a dataset where individuals change in their depression or anxiety over time.

## 4.1. Strengths and limitations

One strength of the present study is that it provides a comprehensive review of commonly used methods to assess temporal stability. Another strength is that we use a large sample of 2625 cancer survivors to illustrate how several longitudinal latent variable models can be used to study the temporal stability of psychological constructs that are measured with skewed and ordinal item scores.

A limitation of the present study is that the Type D personality traits, depression, and anxiety were measured only once during four consecutive years. Although personality changes happen across longer timespans, they are often quite small and tend to peak in stability after the age of 50 (Roberts & Nickel, 2017). Therefore, it would be interesting to replicate our findings using data stretching over lengthier time periods (e.g., decades). The long interval between consecutive measurement in the current study limits conclusions regarding the temporal stability of state-like constructs, as these tend to vary over much shorter time frames (e.g., hours, days). Therefore, future research could also investigate whether our results hold when these psychological constructs are more frequently measured than once a year. In recent years, there has been a surge in researchers focusing on time-intensive longitudinal data (e.g., daily depression measurements) and latent variable models can certainly be used in that context (e.g. Vogelsmeier, Vermunt, van Roekel, & de Roover, 2019).

A limitation of the univariate LGC model (and all other reviewed methods testing for the absolute stability) is that a non-significant average latent linear slope does not necessarily imply that there is absolute mean-level stability, because the non-significant finding may also be caused by insufficient statistical power. The present study involved over 2000 participants and was sufficiently powered to detect small changes in the constructs over time (see Figure D1 in ***Appendix D***). We recommend researchers to evaluate the statistical power when testing for absolute stability. Alternatively, researchers can use Bayesian statistics to directly estimate the evidence in favor of the null hypothesis of no difference between the measurements (Kruschke, 2014). Researchers wishing to stay within a frequentist statistics framework are advised to use equivalence tests (Lakens, 2017) to test the plausibility of absolute stability.

## 4.2. Recommendations

Longitudinal latent variable models can become quite complex and fitting them can be a daunting task, often with unexpected complications (Geiser, Keller & Lockhart, 2013). When fitting LGC models, the number of measurement occasions determines which growth curve shapes can be modeled. Linear growth curves can be estimated with as little as two measurement occasions, Quadratic and cubic growth curves require at least 4 and 5 measurement occasions respectively, though in practice often more measurements are required to properly estimate the growth model without convergence problems, especially when also simultaneously a measurement model for each of the repeated measurements of the latent constructs. Indeed, in the current study-four timepoints were not sufficient to model quadratic growth curves.

With respect to the sample size required to adequately fit a LGC model, several aspects need to be considered (Preacher, 2010). First, the sample size needs to be sufficiently large to produce a stable estimation of all model parameters. Second-order growth models that include a measurement model generally require more participants then first-order growth models, because a larger number of parameters need to be estimated. Second, to adequately determine the fit of the model to the data, researchers should conduct a power analysis to determine the sample size required to detect a poorly fitting model based on the chi-square test (or another model fit measure) with a power of for instance 0.80. Lastly, when researchers have specific hypotheses regarding specific model parameters, such as the latent slope, then they should conduct a power analysis (such as the one reported in ***Appendix D***) to determine the sample size required to detect a latent slope of a particular size with sufficient power.

For an elaborate discussion of other common analysis choices when fitting longitudinal latent variable models to ordinal item scores, we refer the reader to excellent tutorials specifically focused on longitudinal measurement invariance analyses (Liu et al., 2018) and second order LGC models (Masyn, Petras & Liu, 2014; Zheng, Yang & Harring, 2022).

## 4.3. Conclusion

In this study, we provided a comprehensive review of the methods traditionally used to assess the temporal stability of psychological constructs. We noted how most of the reviewed methods do not handle the measurement error in the questionnaire item scores. At least in the literature on Type D personality, we observed that the crucial assumption of longitudinal measurement invariance is typically not tested. We illustrated how these issues can be handled using several longitudinal latent variable models. As we focused on commonly used latent variables, other latent variables model such as continuous time models (e.g., Haehner, Kritzler, Fassbender & Numann, 2021) fell beyond the scope of the current study. Nevertheless, our work illustrates the general benefit of latent variable modeling when assessing the temporal stability of psychological constructs.

**Open Science Framework page**:

https://osf.io/e7ajr/?view_only=41e3de13bb2e4a2eaa4168f0e124fdcc.

## Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Data availability

Data will be made available on request.

## Acknowledgements

In this paper use is made of data of PROFILES (Patient Reported Outcomes Following Initial treatment and Long-term Survivorship)" and refers to the European Journal of Cancer publication : van de Poll-Franse LV, Horevoorts N, van Eenbergen M, Denollet J, Roukema JA, Aaronson NK, Vingerhoets A, Coebergh JW, de Vries J, Essink-Bot ML, Mols F; Profiles Registry Group. The Patient Reported Outcomes Following Initial treatment and Long term Evaluation of Survivorship registry: scope, rationale and design of an infrastructure for the study of physical and psychosocial outcomes in cancer survivorship cohorts. Eur J Cancer. 2011 Sep;47(14):2188- 94. Epub 2011 May 27

## Appendix A:. Mathematical details behind several temporal stability methods

### Cohen's d (for paired data).

Cohen's d can be calculated by standardizing the raw difference in mean scores between the two measurements ($M_2 - M_1$) by a pooled standard deviation for paired data (formula 1; Borenstein, Hedges, Higgins & Rothstein, 2011, page 29), where $SD_{T1}$ and $SD_{T2}$ are the standard deviations of the scores at each measurement and where $r_{(T1,T2)}$ denotes the correlation between the two measurements. Because the numerator is a single difference between average scores, this method only assesses mean-level absolute stability and does therefore not necessarily detect individual-level absolute stability.

$$d = \frac{M_2 - M_1}{\left(\frac{\sqrt{\left(SD^2_{T_1} + SD^2_{T_2} - 2{*}r_{(T1,T2)}{*}SD_{T_1}{*}SD_{T_2}\right)}}{\sqrt{2{*}(1-r_{(T1,T2)})}}\right)} \tag{1}$$

### Reliable change index (RCI).

The RCI can be computed using Formula 2 (Christensen, 1986), where $D_i$ denotes the difference between scores on two measurements for individual $i$ ($D_i = X_{(T2)i} - X_{(T1)i}$) and $S_E$ the standard error of measurement, which can be calculated using formula 3, where $S_{(T1)}$ indicates the standard deviation of scores at T1 and $r_{(T1,T2)}$ the test-rest reliability.

$$RCI = \frac{D_i}{\sqrt{(2({S_E}^2)}} \tag{2}$$

$$S_E = S_{(T1)}{*}\sqrt{\left(1 - r_{(T1,T2)}\right)} \tag{3}$$

The last step involves an interpretation of the RCI's computed for each individual. An RCI larger than 1.96 suggests significant increase, an RCI smaller than −1.96 suggests significant decrease, and an RCI between −1.96 and 1.96 indicates no significant change at a significance level of 0.05.

## Appendix B:. Item baseline characteristics for the DS14 (negative affectivity & social inhibition) and HADS (depression and anxiety) questionnaires

| Construct | Item | n | mean | sd | median | min | max | skew | kurtosis |
|---|---|---|---|---|---|---|---|---|---|
| NA | DS_2 | 2542 | 1.51 | 1.34 | 2 | 0 | 4 | 0.37 | −1.01 |
| | DS_4 | 2555 | 0.84 | 1.13 | 0 | 0 | 4 | 1.14 | 0.34 |
| | DS_5 | 2553 | 1.12 | 1.19 | 1 | 0 | 4 | 0.68 | −0.59 |
| | DS_7 | 2555 | 0.92 | 1.17 | 0 | 0 | 4 | 1.04 | 0.06 |
| | DS_9 | 2562 | 0.65 | 0.95 | 0 | 0 | 4 | 1.39 | 1.28 |
| | DS_12 | 2562 | 1.37 | 1.31 | 1 | 0 | 4 | 0.5 | −0.92 |
| | DS_13 | 2560 | 0.76 | 1.07 | 0 | 0 | 4 | 1.31 | 0.86 |
| SI | DS_1 | 2579 | 0.99 | 1.09 | 1 | 0 | 4 | 0.72 | −0.33 |
| | DS_3 | 2554 | 1.67 | 1.29 | 2 | 0 | 4 | 0.26 | −0.86 |
| | DS_6 | 2563 | 0.97 | 1.15 | 1 | 0 | 4 | 0.92 | −0.15 |
| | DS_8 | 2560 | 1 | 1.18 | 1 | 0 | 4 | 0.9 | −0.24 |
| | DS_10 | 2562 | 1.14 | 1.26 | 1 | 0 | 4 | 0.73 | −0.64 |
| | DS_11 | 2561 | 1.15 | 1.18 | 1 | 0 | 4 | 0.66 | −0.54 |
| | DS_14 | 2564 | 0.93 | 1.13 | 0 | 0 | 4 | 0.96 | −0.01 |
| DEP | HADS_2 | 2572 | 0.63 | 0.75 | 0 | 0 | 3 | 1.02 | 0.52 |
| | HADS_4 | 2575 | 0.43 | 0.65 | 0 | 0 | 3 | 1.41 | 1.59 |
| | HADS_6 | 2572 | 0.39 | 0.7 | 0 | 0 | 3 | 1.82 | 2.74 |
| | HADS_8 | 2572 | 1.05 | 0.82 | 1 | 0 | 3 | 0.63 | 0.06 |
| | HADS_10 | 2573 | 0.56 | 0.74 | 0 | 0 | 3 | 1.3 | 1.37 |
| | HADS_12 | 2575 | 0.63 | 0.79 | 0 | 0 | 3 | 1.11 | 0.6 |
| | HADS_14 | 2582 | 0.72 | 0.9 | 0 | 0 | 3 | 0.96 | −0.22 |

(continued on next page)

(*continued*)

| Construct | Item | n | mean | sd | median | min | max | skew | kurtosis |
|---|---|---|---|---|---|---|---|---|---|
| ANX | HADS_1 | 2571 | 0.8 | 0.73 | 1 | 0 | 3 | 0.82 | 0.83 |
| | HADS_3 | 2572 | 0.71 | 0.85 | 0 | 0 | 3 | 0.98 | 0.08 |
| | HADS_5 | 2556 | 0.73 | 0.8 | 1 | 0 | 3 | 0.93 | 0.31 |
| | HADS_7 | 2574 | 0.62 | 0.78 | 0 | 0 | 3 | 1 | 0.09 |
| | HADS_9 | 2570 | 0.45 | 0.64 | 0 | 0 | 3 | 1.37 | 1.7 |
| | HADS_11 | 2567 | 0.9 | 0.98 | 1 | 0 | 3 | 0.78 | −0.52 |
| | HADS_13 | 2578 | 0.43 | 0.61 | 0 | 0 | 3 | 1.41 | 2.26 |

## Appendix C:. Difference in item response probabilities between the models with and without the constraint that item threshold parameters are invariant over time. These probabilities indicate the change on an item response probability when freely estimating the threshold parameters. Absolute changes larger than 0.05 are considered significant.

| Time | Response | Item | | | | | | |
|---|---|---|---|---|---|---|---|---|
| *Negative affectivity* | | DS_2 | DS_4 | DS_5 | DS_7 | DS_9 | DS_12 | DS_13 |
| T1 | False | 0.00 | −0.01 | 0.00 | 0.00 | −0.01 | 0.01 | 0.00 |
| | Rather false | 0.00 | 0.01 | 0.00 | 0.00 | 0.00 | −0.01 | 0.00 |
| | Neutral | −0.01 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| | Rather true | 0.01 | 0.00 | 0.01 | 0.02 | 0.00 | 0.01 | 0.01 |
| | True | −0.01 | 0.00 | 0.00 | −0.01 | 0.00 | −0.01 | −0.01 |
| T2 | False | −0.01 | 0.00 | 0.00 | 0.01 | 0.00 | 0.00 | 0.00 |
| | Rather false | 0.00 | 0.00 | 0.00 | 0.00 | 0.01 | 0.00 | 0.01 |
| | Neutral | 0.01 | 0.01 | 0.01 | 0.01 | 0.00 | 0.01 | 0.00 |
| | Rather true | −0.01 | 0.00 | −0.01 | −0.01 | 0.00 | 0.00 | 0.00 |
| | True | 0.01 | −0.01 | 0.00 | 0.00 | −0.01 | 0.00 | −0.01 |
| T3 | False | 0.01 | 0.01 | 0.00 | 0.00 | 0.01 | −0.01 | 0.00 |
| | Rather false | 0.00 | 0.00 | 0.01 | 0.00 | 0.00 | 0.01 | 0.01 |
| | Neutral | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| | Rather true | 0.01 | 0.00 | 0.01 | −0.01 | 0.00 | −0.01 | −0.01 |
| | True | −0.01 | −0.01 | −0.01 | 0.01 | 0.00 | 0.01 | 0.01 |
| T4 | False | 0.00 | 0.01 | −0.01 | 0.00 | 0.01 | −0.01 | 0.01 |
| | Rather false | 0.00 | 0.00 | 0.01 | 0.01 | −0.01 | 0.01 | −0.01 |
| | Neutral | 0.01 | 0.01 | 0.01 | 0.01 | 0.00 | 0.01 | 0.00 |
| | Rather true | −0.01 | −0.01 | −0.01 | −0.02 | −0.01 | 0.00 | −0.01 |
| | True | 0.00 | 0.00 | −0.01 | 0.01 | 0.00 | −0.01 | 0.01 |
| *Social inhibition* | | DS_1 | DS_3 | DS_6 | DS_8 | DS_10 | DS_11 | DS_14 |
| T1 | False | −0.01 | −0.01 | 0.01 | 0.01 | 0.00 | 0.01 | 0.00 |
| | Rather false | 0.01 | 0.01 | −0.01 | −0.01 | 0.00 | −0.01 | 0.00 |
| | Neutral | −0.01 | −0.01 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| | Rather true | 0.00 | 0.01 | 0.01 | 0.01 | 0.00 | 0.01 | 0.01 |
| | True | 0.00 | −0.01 | −0.01 | −0.01 | 0.01 | −0.01 | 0.00 |
| T2 | False | 0.00 | −0.01 | 0.00 | −0.01 | 0.00 | 0.00 | 0.00 |
| | Rather false | 0.00 | 0.01 | 0.00 | 0.01 | 0.00 | 0.00 | 0.00 |
| | Neutral | 0.01 | 0.01 | 0.01 | 0.00 | 0.00 | 0.01 | 0.00 |
| | Rather true | 0.00 | 0.00 | 0.00 | −0.01 | 0.01 | −0.01 | −0.02 |
| | True | −0.01 | 0.00 | −0.01 | 0.00 | −0.01 | 0.01 | 0.01 |
| T3 | False | 0.01 | 0.00 | 0.00 | 0.00 | −0.01 | 0.00 | 0.00 |
| | Rather false | −0.01 | 0.00 | 0.00 | 0.00 | 0.01 | 0.01 | 0.00 |
| | Neutral | 0.00 | 0.01 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| | Rather true | 0.00 | −0.01 | −0.01 | 0.00 | 0.01 | 0.00 | 0.01 |
| | True | −0.01 | 0.01 | 0.01 | 0.00 | −0.01 | 0.00 | −0.01 |
| T4 | False | 0.01 | 0.01 | −0.01 | 0.00 | −0.01 | −0.02 | 0.00 |
| | Rather false | −0.01 | −0.02 | 0.01 | 0.00 | 0.01 | 0.02 | 0.00 |
| | Neutral | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 |
| | Rather true | −0.01 | −0.01 | −0.03 | −0.01 | −0.01 | 0.00 | −0.01 |
| | True | 0.00 | 0.00 | 0.01 | 0.00 | 0.00 | −0.01 | −0.01 |
| *Depression* | | HADS_2 | HADS_4 | HADS_6 | HADS_8 | HADS_10 | HADS_12 | HADS_14 |
| T1 | Not at all | 0.00 | 0.00 | 0.01 | 0.00 | −0.01 | 0.00 | 0.01 |
| | Sometimes | 0.00 | 0.00 | −0.01 | 0.00 | 0.00 | 0.00 | −0.01 |
| | A lot of the time | 0.00 | 0.01 | 0.00 | 0.00 | 0.01 | 0.00 | 0.00 |
| | Most of the time | 0.00 | 0.00 | 0.00 | −0.01 | 0.00 | 0.00 | 0.00 |
| T2 | Not at all | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| | Sometimes | 0.00 | −0.01 | 0.00 | 0.00 | −0.01 | 0.00 | −0.01 |
| | A lot of the time | 0.00 | 0.00 | 0.00 | 0.00 | 0.01 | −0.01 | 0.01 |
| | Most of the time | 0.01 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | −0.01 |
| T3 | Not at all | −0.01 | 0.01 | 0.01 | −0.01 | 0.01 | 0.01 | 0.00 |
| | Sometimes | 0.02 | 0.01 | 0.01 | 0.02 | 0.02 | 0.00 | 0.02 |

(*continued on next page*)

(*continued*)

| Time | Response | Item | | | | | | |
|---|---|---|---|---|---|---|---|---|
| *Negative affectivity* | | DS_2 | DS_4 | DS_5 | DS_7 | DS_9 | DS_12 | DS_13 |
| | A lot of the time | 0.01 | −0.01 | −0.01 | 0.00 | −0.01 | 0.01 | 0.00 |
| | Most of the time | −0.03 | −0.01 | −0.01 | 0.00 | −0.02 | −0.01 | −0.01 |
| T4 | Not at all | 0.00 | 0.01 | 0.01 | 0.00 | 0.01 | −0.01 | −0.01 |
| | Sometimes | 0.03 | 0.04 | 0.03 | 0.01 | 0.04 | 0.03 | 0.04 |
| | A lot of the time | 0.01 | −0.02 | −0.02 | −0.01 | −0.01 | −0.01 | 0.01 |
| | Most of the time | −0.04 | −0.02 | −0.02 | 0.00 | −0.04 | −0.01 | −0.04 |
| | | | | | | | | |
| *Anxiety* | | HADS_1 | HADS_3 | HADS_5 | HADS_7 | HADS_9 | HADS_11 | HADS_13 |
| T1 | Not at all | 0.01 | 0.01 | −0.01 | −0.01 | 0.00 | −0.01 | 0.00 |
| | Sometimes | −0.01 | 0.00 | 0.01 | 0.00 | −0.01 | 0.01 | −0.01 |
| | A lot of the time | 0.00 | −0.01 | 0.00 | 0.01 | 0.00 | 0.00 | 0.00 |
| | Most of the time | 0.00 | 0.00 | −0.01 | 0.00 | 0.00 | 0.00 | 0.00 |
| T2 | Not at all | −0.01 | 0.00 | 0.01 | 0.00 | −0.02 | 0.00 | 0.01 |
| | Sometimes | 0.02 | 0.02 | 0.00 | 0.01 | 0.02 | 0.01 | 0.01 |
| | A lot of the time | 0.00 | 0.00 | 0.00 | −0.01 | −0.01 | 0.00 | −0.01 |
| | Most of the time | −0.01 | −0.01 | −0.01 | 0.00 | 0.00 | −0.01 | −0.01 |
| T3 | Not at all | −0.02 | 0.00 | 0.00 | 0.02 | 0.02 | −0.01 | 0.01 |
| | Sometimes | 0.04 | 0.02 | 0.02 | 0.01 | 0.00 | 0.02 | 0.03 |
| | A lot of the time | 0.01 | 0.02 | −0.01 | −0.02 | −0.01 | 0.01 | −0.01 |
| | Most of the time | −0.04 | −0.04 | 0.00 | −0.01 | −0.01 | −0.02 | −0.03 |
| T4 | Not at all | −0.01 | 0.00 | −0.01 | 0.01 | 0.02 | 0.01 | 0.01 |
| | Sometimes | 0.04 | 0.02 | 0.03 | 0.02 | 0.01 | 0.01 | 0.04 |
| | A lot of the time | 0.00 | 0.01 | −0.01 | 0.01 | −0.02 | 0.00 | 0.00 |
| | Most of the time | −0.03 | −0.04 | −0.01 | −0.03 | −0.01 | −0.01 | −0.05 |

## Appendix D:. Power analysis for the univariate latent growth curve model

We conducted a power analysis for the univariate LGC model to test the null hypothesis that the average latent slope parameter is equal to zero. We used the R-package simsem to conduct the power analysis for our LGC model fitted in Study 2. We assumed a significance level of 0.05 and investigated the minimal size of the latent slope that could be detected with sufficient power given the current sample size of 2600 participants and a significance level of 0.05. Figure D1 below indicates that this number of participants is enough to detect a latent slope of 0.06 or larger with a power of 0.80. A latent linear slope of 0.06 implies that on average participants change 0.18 on the scale of the latent variable over a four-year period. The estimated variance of the latent NA factor varies across repeated measurements between 0.17 and 0.22, corresponding to standard deviations of 0.40 and 0.47. The most conservative estimate is that our study was sufficiently powered to detect a latent slope corresponding to a change in NA of 0.18 / 0.40 = 0.45 standard deviations.

## Appendix E:. Analyses assuming the ordinal items to be continuous

This appendix shows the fit and estimates for models that assume the ordinal item scores to be continuous by fitting a standard linear structural equation model to the Pearson correlation matrix. Table E1 presents the model fit for each of the continuous longitudinal measurement invariance models fitted to the repeated measures of NA, SI, depression and anxiety. The results indicate that the continuous invariance model also adequately fitted the data, though the fit of the ordinal model reported in the main text was slightly better on almost all fit indices. Based on the small changes in RMSEA, CFI and SRMR across the tested measurement invariance models, this continuous model similarly supports the presence of longitudinal measurement invariance of NA, SI, depression, and anxiety.

Table E2 presents the model fit indices and latent growth curve estimates of the univariate latent growth curve models, assuming continuous NA, SI, depression and anxiety item scores. The continuous models fitted the data less adequately than the ordinal models reported in the main text. Especially the SRMR and CFI showed poor model fit. The estimates and statistical significance of the latent slope parameters were very similar to the estimates produced by the ordinal model. However, the average latent intercept differed considerably from the estimate in the ordinal model. One explanation could be that the mean structure of skewed ordinal item scores can be more adequately modeled using multiple threshold parameters than with a single intercept parameter.

Table E3 presents the model fit indices and trait/state variance estimates of the trait-state-occasion model fitted to the NA, SI, depression and anxiety item scores, assuming that these item scores are continuous. The results again indicate slightly worse model fit than for the ordinal models reported in the main text. Nevertheless, the estimated proportions of trait and state variance were very similar to those produced by the ordinal models, except for depression and anxiety, for which the continuous model suggested more trait variance.

**Table E1**
Fit statistics for models testing the longitudinal measurement invariance of the DS14 (negative affectivity & social inhibition) and HADS (depression & anxiety).

| Model | df | Δdf | $\chi^2$ | $\Delta\chi^2$ | RMSEA (95 %CI) | ΔRMSEA | SRMR | ΔSRMR | CFI | ΔCFI |
|---|---|---|---|---|---|---|---|---|---|---|
| *Negative affectivity* | | | | | | | | | | |
| 1 | 302 | – | 989.1 | – | 0.049 [0.046 0.052] | – | 0.043 | – | 0.963 | – |
| 2 | 320 | 18 | 1006.8 | 17.6 | 0.048 [0.044 0.051] | 0.001 | 0.046 | 0.003 | 0.963 | <0.001 |
| 3 | 338 | 18 | 1047.3 | 40.5* | 0.047 [0.044 0.050] | 0.001 | 0.046 | <0.001 | 0.962 | 0.001 |
| 4 | 359 | 21 | 1100.2 | 52.9* | 0.047 [0.044 0.050] | <0.001 | 0.046 | <0.001 | 0.960 | 0.002 |
| *Social inhibition* | | | | | | | | | | |
| 1 | 302 | – | 1642.9 | – | 0.067 [0.063 0.070] | – | 0.079 | – | 0.930 | – |
| 2 | 320 | 18 | 1664.3 | 21.3 | 0.065 [0.062 0.068] | 0.002 | 0.080 | 0.001 | 0.930 | <0.001 |
| 3 | 338 | 18 | 1712.7 | 48.4* | 0.064 [0.061 0.067] | 0.001 | 0.081 | 0.001 | 0.928 | 0.002 |
| 4 | 359 | 21 | 1791.2 | 78.5* | 0.063 [0.060 0.066] | 0.001 | 0.081 | <0.001 | 0.925 | 0.003 |
| *Depression* | | | | | | | | | | |
| 1 | 302 | – | 477.7 | – | 0.024 [0.020 0.028] | – | 0.028 | – | 0.989 | – |
| 2 | 320 | 18 | 501.4 | 23.6 | 0.023 [0.019 0.027] | 0.001 | 0.032 | 0.004 | 0.988 | 0.001 |
| 3 | 338 | 18 | 535.9 | 34.6* | 0.024 [0.020 0.028] | -0.001 | 0.032 | <0.001 | 0.987 | 0.001 |
| 4 | 359 | 21 | 581.3 | 45.4* | 0.024 [0.021 0.028] | <0.001 | 0.033 | 0.001 | 0.986 | 0.001 |
| *Anxiety* | | | | | | | | | | |
| 1 | 302 | – | 660.9 | – | 0.034 [0.030 0.038] | – | 0.038 | – | 0.976 | – |
| 2 | 320 | 18 | 683.4 | 22.5 | 0.033 [0.030 0.037] | 0.001 | 0.040 | 0.002 | 0.976 | <0.001 |
| 3 | 338 | 18 | 712.8 | 29.4* | 0.033 [0.029 0.036] | <0.001 | 0.041 | 0.001 | 0.975 | 0.001 |
| 4 | 359 | 21 | 754.6 | 41.8* | 0.033 [0.029 0.036] | <0.001 | 0.042 | 0.001 | 0.974 | 0.001 |

Model 1 (Configural invariance): Baseline model.
Model 2 (Weak invariance): Invariant factor loadings.
Model 3 (Strong invariance): Invariant factor loadings & intercepts.
Model 4 (Strict invariance): Invariant factor loadings, intercepts & residuals.
* p <.05.

**Table E2**
Fit indices and individual change in negative affectivity, social inhibition, depression and anxiety in terms of the mean and variance of the latent intercept and slope, estimated using a latent growth curve model assuming continuous item scores.

| | Negative affectivity | Social inhibition | Depression | Anxiety |
|---|---|---|---|---|
| *Model fit* | N = 2597 | N = 2604 | N = 2602 | N = 2601 |
| Parameters | 99 | 99 | 99 | 99 |
| $\chi^2$ | 5168.3* | 5116.4* | 8767.9* | 8633.2* |
| RMSEA (95 %CI) | 0.075 [0.073, 0.076] | 0.074 [0.072, 0.076] | 0.098 [0.097, 0.100] | 0.098 [0.096, 0.099] |
| SRMR | 0.112 | 0.124 | 0.216 | 0.200 |
| CFI | 0.840 | 0.838 | 0.650 | 0.645 |
| *Latent growth parameters* | | | | |
| Mean Intercept | 1.299* | 0.991* | 0.461* | 0.760* |
| Variance Intercept | 0.795* | 0.452* | 0.090 | 0.221* |
| Mean Slope | −0.056 | 0.016 | 0.022 | −0.004 |
| Variance Slope | 0.015* | 0.005 | 0.002 | 0.005* |

* p <.05.

**Table E3**
Fit indices and trait/state variance proportions (95 %CI) for the second-order latent trait-state-occasion models of negative affectivity, social inhibition, depression and anxiety, assuming the ordinal item scores to be continuous variables.

| | Negative affectivity | Social inhibition | Depression | Anxiety |
|---|---|---|---|---|
| *Model fit* | | | | |
| N | 2597 | 2604 | 2602 | 2601 |
| Parameters | 130 | 130 | 130 | 130 |
| $\chi^2$ | 1491.0* | 2872.1* | 751.3* | 970.6* |
| RMSEA (95 %CI) | 0.037 [0.035, 0.039] | 0.055 [0.053, 0.057] | 0.023 [0.021, 0.025] | 0.028 [0.026, 0.030] |
| SRMR | 0.042 | 0.076 | 0.030 | 0.041 |
| CFI | 0.961 | 0.914 | 0.982 | 0.972 |
| *Proportion explained variance* | | | | |
| Latent trait | 0.75 [0.71, 0.78] | 0.82 [0.78, 0.85] | 0.80 [0.77, 0.83] | 0.81 [0.78, 0.84] |
| Latent occasion | 0.25 [0.22, 0.28] | 0.18 [0.15, 0.21] | 0.20 [0.17, 0.23] | 0.18 [0.15, 0.21] |

* p <.05.

## References

Aguayo-Carreras, P., Ruiz-Carrascosa, J. C., Ruiz-Villaverde, R., & Molina-Leyva, A. (2021). Four years stability of type D personality in patients with moderate to severe psoriasis and its implications for psychological impairment. *Anais Brasileiros de Dermatologia, 96*, 558–564. https://doi.org/10.1016/j.abd.2021.02.005

Alçelik, A., Yildirim, O., Canan, F., Eroglu, M., Aktas, G., & Savli, H. (2012). A preliminary psychometric evaluation of the type D personality construct in Turkish hemodialysis patients. *Journal of Mood Disorders, 2*(1), 1. https://doi.org/10.5455/jmood.20120307062608.

Allemand, M., & Martin, M. (2017). On correlated change in personality. *European Psychologist*. https://doi.org/10.1027/1016-9040/a000256

Aluja, A., Malas, O., Lucas, I., Worner, F., & Bascompte, R. (2019). Assessment of the Type D personality distress in coronary heart disease patients and healthy subjects in Spain. *Personality and Individual Differences, 142*, 301–309. https://doi.org/10.1016/j.paid.2018.08.011

Bagherian, R., & Ehsan, H. B. (2011). Psychometric properties of the Persian version of type D personality scale (DS14). *Iranian journal of psychiatry and behavioral sciences, 5* (2), 12.

Baltes, P. B. (1987). Theoretical propositions of life-span developmental psychology: On the dynamics between growth and decline. *Developmental psychology, 23*(5), 611. https://doi.org/10.1037/0012-1649.23.5.611

Bartko, J. J. (1976). On various intraclass correlation reliability coefficients. *Psychological bulletin, 83*(5), 762. https://doi.org/10.1037/0033-2909.83.5.762

Batselé, E., Denollet, J., Lussier, A., Loas, G., Vanden Eynde, S., Van de Borne, P., & Fantini-Hauwel, C. (2017). Type D personality: Application of DS14 French version in general and clinical populations. *Journal of health psychology, 22*(8), 1075–1083. https://doi.org/10.1177/1359105315624499

Bjelland, I., Dahl, A. A., Haug, T. T., & Neckelmann, D. (2002). The validity of the Hospital Anxiety and Depression Scale: An updated literature review. *Journal of psychosomatic research, 52*(2), 69–77. https://doi.org/10.1016/S0022-3999(01)00296-3

Bollen, K. A. (2005). Structural equation models. *Encyclopedia of biostatistics, 7*. https://doi.org/10.1002/0470011815.b2a13089

Borghuis, J., Denissen, J. J., Oberski, D., Sijtsma, K., Meeus, W. H., Branje, S., … Bleidorn, W. (2017). Big Five personality stability, change, and codevelopment across adolescence and early adulthood. *Journal of Personality and Social Psychology, 113*(4), 641. https://doi.org/10.1037/pspp0000138

Bouwens, E., van Lier, F., Rouwet, E. V., Verhagen, H. J., Stolker, R. J., & Hoeks, S. E. (2019). Type D Personality and Health-Related Quality of Life in Vascular Surgery Patients. *International journal of behavioral medicine, 26*(4), 343–351. https://doi.org/10.1007/s12529-018-09762-3

Briley, D., & Tucker-Drob, E. (2014). Genetic and environmental continuity in personality development: a meta-analysis. *Psychological bulletin, 140 5*, 1303-1331. https://doi.org/10.1037/a0037091.

Bunevicius, A., Staniute, M., Brozaitiene, J., Stropute, D., Bunevicius, R., & Denollet, J. (2013). Type D (distressed) personality and its assessment with the DS14 in Lithuanian patients with coronary artery disease. *Journal of health psychology, 18*(9), 1242–1251. https://doi.org/10.1177/1359105312459098

Caspi, A., Roberts, B. W., & Shiner, R. L. (2005). Personality development: Stability and change. *Annual Review of Psychology, 56*, 453–484. https://doi.org/10.1146/annurev.psych.55.090902.141913

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 14*(3), 464–504. https://doi.org/10.1080/10705510701301834

Christensen, L. (1986). A method of assessing change in a single subject: An alteration of the RC index. *Behav Ther, 17*, 305–308. https://doi.org/10.1016/S0005-7894(86)80060-0

Cole, D. A., Martin, N. C., & Steiger, J. H. (2005). Empirical and conceptual problems with longitudinal trait-state models: Introducing a trait-state-occasion model. *Psychological methods, 10*(1), 3. https://doi.org/10.1037/1082-989X.10.1.3

Cole, D. A., & Preacher, K. J. (2014). Manifest variable path analysis: Potentially serious and misleading consequences due to uncorrected measurement error. *Psychological Methods, 19*(2), 300. https://doi.org/10.1037/a0033805

Collins, L. M., & Lanza, S. T. (2009). *Latent class and latent transition analysis: With applications in the social, behavioral, and health sciences* (Vol. 718). John Wiley & Sons.

Condén, E., Rosenblad, A., Ekselius, L., & Åslund, C. (2014). Prevalence of Type D personality and factorial and temporal stability of the DS 14 after myocardial infarction in a Swedish population. *Scandinavian Journal of Psychology, 55*(6), 601–610. https://doi.org/10.1111/sjop.12162

Curran, P. J., Obeidat, K., & Losardo, D. (2010). Twelve frequently asked questions about growth curve modeling. *Journal of cognition and development, 11*(2), 121–136. https://doi.org/10.1080/15248371003699969

Dannemann, S., Matschke, K., Einsle, F., Smucker, M. R., Zimmermann, K., Joraschky, P., … Köllner, V. (2010). Is type-D a stable construct? An examination of type-D personality in patients before and after cardiac surgery. *Journal of psychosomatic research, 69*(2), 101–109. https://doi.org/10.1016/j.jpsychores.2010.02.008

De Fruyt, F., Bartels, M., Van Leeuwen, K. G., De Clercq, B., Decuyper, M., & Mervielde, I. (2006). Five types of personality continuity in childhood and adolescence. *Journal of Personality and Social Psychology, 91*(3), 538. https://doi.org/10.1037/0022-3514.91.3.538

Denollet, J. (1998). Personality and coronary heart disease: The type-D scale-16 (DS16). *Annals of Behavioral Medicine, 20*(3), 209–215. https://doi.org/10.1007/BF02884962

Denollet, J. (2005). DS14: Standard assessment of negative affectivity, social inhibition, and Type D personality. *Psychosomatic medicine, 67*(1), 89–97. https://doi.org/10.1097/01.psy.0000149256.81953.49

Dolan, C. V. (1994). Factor analysis of variables with 2, 3, 5 and 7 response categories: A comparison of categorical variable estimators using simulated data. *British Journal of Mathematical and Statistical Psychology, 47*(2), 309–326. https://doi.org/10.1111/j.2044-8317.1994.tb01039.x

Ferguson, E., Williams, L., O'Connor, R. C., Howard, S., Hughes, B. M., Johnston, D. W., … O'Carroll, R. E. (2009). A taxometric analysis of type-D personality. *Psychosomatic medicine, 71*(9), 981–986. https://doi.org/10.1097/PSY.0b013e3181bd888b

Figueredo, A. J., de Baca, T. C., & Black, C. (2014). No matter where you go, there you are: The genetic foundations of temporal stability. *Journal of Methods and Measurement in the Social Sciences, 5*(2), 76–106. https://doi.org/10.2458/v5i2.18477

Gana, K., & Broc, G. (2019). *Structural equation modeling with lavaan*. John Wiley & Sons.

Grande, G., Romppel, M., & Barth, J. (2012). Association between type D personality and prognosis in patients with cardiovascular diseases: A systematic review and meta-analysis. *Annals of behavioral medicine, 43*(3), 299–310. https://doi.org/10.1007/s12160-011-9339-0

Gremigni, P., & Sommaruga, M. (2005). Type D personality, a relevant construct in cardiology. Preliminary study of validation of the Italian questionnaire. *Cognitive and Behavioral. Psychotherapy, 11*(1), 7–18.

Gu, Z., Emons, W. H., & Sijtsma, K. (2018). Review of issues about classical change scores: A multilevel modeling perspective on some enduring beliefs. *Psychometrika, 83*(3), 674–695. https://doi.org/10.1007/s11336-018-9611-3

Haehner, P., Kritzler, S., Fassbender, I., & Luhmann, M. (2021). Stability and Change of Perceived Characteristics of Major Life Events. *PsyArxIv*. https://psyarxiv.com/2yzcs/download.

Hartlage, S., Arduino, K., & Alloy, L. B. (1998). Depressive personality characteristics: State dependent concomitants of depressive disorder and traits independent of current depression. *Journal of Abnormal Psychology, 107*(2), 349. https://doi.org/10.1037/0021-843X.107.2.349

Hertzog, C., Lindenberger, U., Ghisletta, P., & von Oertzen, T. (2006). On the power of multivariate latent growth curve models to detect correlated change. *Psychological methods, 11*(3), 244. https://doi.org/10.1037/1082-989X.11.3.244

Jacobson, N. S., & Truax, P. (1992). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of consulting and clinical psychology, 59*(1), 12–19. https://doi.org/10.1037/0022-006X.59.1.12

Kantor, L., Endler, N. S., Heslegrave, R. J., & Kocovski, N. L. (2001). Validating self-report measures of state and trait anxiety against a physiological measure. *Current Psychology, 20*(3), 207–215. https://doi.org/10.1007/s12144-001-1007-2

Kruschke, J. (2014). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan*. London: Academic Press.

Kupper, N., Boomsma, D. I., de Geus, E. J., Denollet, J., & Willemsen, G. (2011). Nine-year stability of type D personality: Contributions of genes and environment. *Psychosomatic Medicine, 73*(1), 75–82. https://doi.org/10.1097/PSY.0b013e3181fdce54

Kupper, N., & Denollet, J. (2018). Type D personality as a risk factor in coronary heart disease: A review of current evidence. *Current cardiology reports, 20*(11), 104. https://doi.org/10.1007/s11886-018-1048-x

Lakens, D. (2017). Equivalence tests: A practical primer for t tests, correlations, and meta- analyses. *Social psychological and personality science, 8*(4), 355–362. https://doi.org/10.1177/1948550617697177

Leszko, M., Elleman, L. G., Bastarache, E. D., Graham, E. K., & Mroczek, D. K. (2016). Future directions in the study of personality in adulthood and older age. *Gerontology, 62*(2), 210–215. https://doi.org/10.1159/000434720

Li-Gao, R., Boomsma, D., Dolan, C. V., de Geus, E. J. C., Denollet, J., & Kupper, N. (2021). Genetic and environmental contributions to stability and change in social inhibition across the adolescent and adult lifespan. *Preprint received through personal.* communication.

Lim, H. E., Lee, M. S., Ko, Y. H., Park, Y. M., Joe, S. H., Kim, Y. K., … Denollet, J. (2011). Assessment of the type D personality construct in the Korean population: A validation study of the Korean DS14. *Journal of Korean medical science, 26*(1), 116–123. https://doi.org/10.3346/jkms.2011.26.1.116

Liu, Y., Millsap, R. E., West, S. G., Tein, J. Y., Tanaka, R., & Grimm, K. J. (2017). Testing measurement invariance in longitudinal data with ordered-categorical measures. *Psychological Methods, 22*(3), 486. https://doi.org/10.1037/met0000075

Lodder, P. (2020a). Modeling synergy: How to assess a Type D personality effect. *Journal of Psychosomatic Research, 109990*. https://doi.org/10.1016/j.jpsychores.2020.109990

Lodder, P. (2020b). A re-evaluation of the Type D personality effect. *Personality and Individual Differences, 167*, Article 110254. https://doi.org/10.1016/j.paid.2020.110254

Lodder, P., Denollet, J., Emons, W. H., Nefs, G., Pouwer, F., Speight, J., & Wicherts, J. M. (2019). Modeling interactions between latent variables in research on Type D personality: A Monte Carlo simulation and clinical study of depression and anxiety. *Multivariate behavioral research, 54*(5), 637–665. https://doi.org/10.1080/00273171.2018.1562863

Lodder, P., Emons, W. H., Denollet, J., & Wicherts, J. M. (2021). Latent logistic interaction modeling: A simulation and empirical illustration of Type D personality. *Structural Equation Modeling: A Multidisciplinary Journal, 28*(3), 440–462. https://doi.org/10.1080/10705511.2020.1838905

Loosman, W. L., de Jong, R. W., Haverkamp, G. L., van den Beukel, T. O., Dekker, F. W., Siegert, C. E., & Honig, A. (2018). The stability of type D personality in dialysis patients. *International journal of behavioral medicine, 25*(1), 85–92. https://doi.org/10.1007/s12529-017-9667-y

Martens, E. J., Kupper, N., Pedersen, S. S., Aquarius, A. E., & Denollet, J. (2007). Type-D personality is a stable taxonomy in post-MI patients over an 18-month period. *Journal of psychosomatic research, 63*(5), 545–550. https://doi.org/10.1016/j.jpsychores.2007.06.005

Masyn, K. E., Petras, H., & Liu, W. (2014). Growth curve models with categorical outcomes. *Encyclopedia of criminology and criminal justice, 2013–2025*. https://doi.org/10.1007/978-1-4614-5690-2_404

McArdle, J. J. (1986). Latent variable growth within behavior genetic models. *Behavior Genetics, 16*(1), 163–200. https://doi.org/10.1007/BF01065485

McGraw, R. R., & Costa, P. T., Jr (1994). The stability of personality: Observations and evaluations. *Current Directions in Psychological Science, 3*(6), 173–175. https://doi.org/10.1111/1467-8721.ep10770693

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological methods, 1*(1), 30. https://doi.org/10.1037/1082-989X.1.1.30

Meade, A. W., Johnson, E. C., & Braddy, P. W. (2008). Power and sensitivity of alternative fit indices in tests of measurement invariance. *Journal of Applied Psychology, 93*(3), 568. https://doi.org/10.1037/0021-9010.93.3.568

Millsap, R. E., & Cham, H. (2012). Investigating factorial invariance in longitudinal data. In B. Laursen, T. D. Little, & N. A. Card (Eds.), *Handbook of developmental research methods* (pp. 109–127). New York, NY: Guilford Press.

Montero, P., Bermúdez, J., & Rueda, B. (2017). Adaptation to Spanish of the DS-14 Scale («Type D Scale-14») for the measurement of type D personality. *Journal of Psychopathology and Clinical Psychology, 22*(1), 55–67. https://doi.org/10.5944/rppc.vol.22.num.1.2017.16585

Mroczek, D. K., Graham, E. K., Turiano, N. A., & Aro-Lambo, M. O. (2021). Personality development in adulthood and later life. In O. P. John, & R. W. Robins (Eds.), *Handbook of personality: Theory and research* (pp. 336–351). The Guilford Press.

Mroczek, D. K., & Spiro, A., III (2003). Modeling intraindividual change in personality traits: Findings from the Normative Aging Study. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences, 58*(3), P153–P165. https://doi.org/10.1093/geronb/58.3.P153

Muchinsky, P. M. (1996). The correction for attenuation. *Educational and psychological measurement, 56*(1), 63–75. https://doi.org/10.1177/0013164496056001004

Neale, M. C., & McArdle, J. J. (2000). Structured latent growth curves for twin data. *Twin Research and Human Genetics, 3*(3), 165–177. https://doi.org/10.1375/twin.3.3.165

Nefs, G., Pouwer, F., Pop, V., & Denollet, J. (2012). Type D (distressed) personality in primary care patients with type 2 diabetes: Validation and clinical correlates of the DS14 assessment. *Journal of psychosomatic research, 72*(4), 251–257. https://doi.org/10.1016/j.jpsychores.2012.01.006

Nivard, M. G., Dolan, C. V., Kendler, K. S., Kan, K. J., Willemsen, G., Van Beijsterveldt, C. E. M., … Boomsma, D. I. (2015). Stability in symptoms of anxiety and depression as a function of genotype and environment: A longitudinal twin study from ages 3 to 63 years. *Psychological medicine, 45*(5), 1039–1049. https://doi.org/10.1017/S003329171400213X

Oort, F. J. (2005). Using structural equation modeling to detect response shifts and true change. *Quality of Life Research, 14*(3), 587–598. https://doi.org/10.1007/s11136-004-0830-y

Ossola, P., De Panfilis, C., Tonna, M., Ardissino, D., & Marchesi, C. (2015). DS14 is more likely to measure depression rather than a personality disposition in patients with acute coronary syndrome. *Scandinavian Journal of Psychology, 56*(6), 685–692. https://doi.org/10.1111/sjop.12244

Pedersen, S. S., Yagensky, A., Smith, O. R., Yagenska, O., Shpak, V., & Denollet, J. (2009). Preliminary evidence for the cross-cultural utility of the type D personality construct in the Ukraine. *International journal of behavioral medicine, 16*(2), 108. https://doi.org/10.1007/s12529-008-9022-4

Pelle, A. J., Erdman, R. A., van Domburg, R. T., Spiering, M., Kazemier, M., & Pedersen, S. S. (2008). Type D patients report poorer health status prior to and after cardiac rehabilitation compared to non-type D patients. *Annals of Behavioral Medicine, 36*(2), 167–175. https://doi.org/10.1007/s12160-008-9057-4

Pentz, M. A., & Chou, C. P. (1994). Measurement invariance in longitudinal clinical research assuming change from development and intervention. *Journal of Consulting and Clinical Psychology, 62*(3), 450. https://doi.org/10.1037/0022-006X.62.3.450

Piepoli, M. F., Hoes, A. W., Agewall, S., Albus, C., Brotons, C., Catapano, A. L., … Verschuren, W. M. (2016). Guidelines: Editor's choice: 2016 European Guidelines on cardiovascular disease prevention in clinical practice: The Sixth Joint Task Force of the European Society of Cardiology and Other Societies on Cardiovascular Disease Prevention in Clinical Practice (constituted by representatives of 10 societies and by invited experts) Developed with the special contribution of the European Association for Cardiovascular Prevention & Rehabilitation (EACPR). *European heart journal, 37*(29), 2315. https://doi.org/10.1093/eurheartj/ehw106

Prenoveau, J. M., Craske, M. G., Zinbarg, R. E., Mineka, S., Rose, R. D., & Griffith, J. W. (2011). Are anxiety and depression just as stable as personality during late adolescence? Results from a three-year longitudinal latent variable study. *Journal of Abnormal Psychology, 120*(4), 832. https://doi.org/10.1037/a0023939

Ramsey, I., de Rooij, B. H., Mols, F., Corsini, N., Horevoorts, N. J., Eckert, M., & van de Poll-Franse, L. V. (2019). Cancer survivors who fully participate in the PROFILES registry have better health-related quality of life than those who drop out. *Journal of Cancer Survivorship, 13*(6), 829–839. https://doi.org/10.1007/s11764-019-00813-6

Rhemtulla, M., Brosseau-Liard, P.É., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological methods, 17*(3), 354. https://doi.org/10.1037/a0029315

Rijsdijk, F. V., & Sham, P. C. (2002). Analytic approaches to twin data using structural equation models. *Briefings in bioinformatics, 3*(2), 119–133. https://doi.org/10.1093/bib/3.2.119

Roberts, B. W., & DelVecchio, W. F. (2000). The rank-order consistency of personality traits from childhood to old age: A quantitative review of longitudinal studies. *Psychological bulletin, 126*(1), 3. https://doi.org/10.1037/0033-2909.126.1.3

Roberts, B. W., & Mroczek, D. (2008). Personality trait change in adulthood. *Current directions in psychological science, 17*(1), 31–35. https://doi.org/10.1111/j.1467-8721.2008.00543.x

Roberts, B. W., & Nickel, L. B. (2017). A critical evaluation of the Neo-Socioanalytic Model of personality. In J. Specht (Ed.), *Personality Development across the Lifespan* (pp. 157–177). Cambridge, MA: Academic Press.

Roberts, B. W., Walton, K. E., & Viechtbauer, W. (2006). Patterns of mean-level change in personality traits across the life course: A meta-analysis of longitudinal studies. *Psychological bulletin, 132*(1), 1. https://doi.org/10.1037/0033-2909.132.1.1

Roberts, B. W., & Wood, D. (2006). Personality Development in the Context of the Neo-Socioanalytic Model of Personality. In D. K. Mroczek, & T. D. Little (Eds.), *Handbook of personality development* (pp. 11–39). Lawrence Erlbaum Associates Publishers.

Roberts, B. W., Wood, D., & Caspi, A. (2008). The development of personality traits in adulthood. In O. P. John, R. W. Robins, & L. A. Pervin (Eds.), *Handbook of personality: Theory and research* (p. 375–398).

Romppel, M., Herrmann-Lingen, C., Vesper, J. M., & Grande, G. (2012). Six year stability of Type-D personality in a German cohort of cardiac patients. *Journal of Psychosomatic Research, 72*(2), 136–141. https://doi.org/10.1016/j.jpsychores.2011.11.009

Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling and more. Version 0.6–4. *Journal of statistical software, 48*(2), 1–36. https://doi.org/10.18637/jss.v048.i02.

Satorra, A. (2000). *Scaled and adjusted restricted tests in multi-sample analysis of moment structures. In Innovations in multivariate statistical analysis* (pp. 233–247). Boston, MA: Springer.

Schwabe, I., Gu, Z., Tijmstra, J., Hatemi, P., & Pohl, S. (2019). Psychometric modelling of longitudinal genetically-informative twin data. *Frontiers in genetics, 10*, 837. https://doi.org/10.3389/fgene.2019.00837

Shrout, P. E. (1998). Measurement reliability and agreement in psychiatry. *Statistical methods in medical research, 7*(3), 301–317. https://doi.org/10.1177/096228029800700306

Spearman, C. (1904). The proof and measurement of association between two things. *The American journal of psychology, 15*(1), 72–101. https://doi.org/10.2307/1412159

Spindler, H., Kruse, C., Zwisler, A. D., & Pedersen, S. S. (2009). Increased anxiety and depression in Danish cardiac patients with a type D personality: Cross-validation of the Type D Scale (DS14). *International journal of behavioral medicine, 16*(2), 98–107. https://doi.org/10.1007/s12529-009-9037-5

Spinhoven, P. H., Ormel, J., Sloekers, P. P. A., Kempen, G. I. J. M., Speckens, A. E. M., & Van Hemert, A. M. (1997). A validation study of the Hospital Anxiety and Depression Scale (HADS) in different groups of Dutch subjects. *Psychological medicine, 27*(2), 363–370. https://doi.org/10.1017/S0033291796004382

van de Poll-Franse, L. V., Horevoorts, N., van Eenbergen, M., Denollet, J., Roukema, J. A., Aaronson, N. K., … Mols, F. (2011). The patient reported outcomes following initial treatment and long term evaluation of survivorship registry: Scope, rationale and design of an infrastructure for the study of physical and psychosocial outcomes in cancer survivorship cohorts. *European Journal of Cancer, 47*(14), 2188–2194. https://doi.org/10.1016/j.ejca.2011.04.034

van den Berg, S. M., Glas, C. A., & Boomsma, D. I. (2007). Variance decomposition using an IRT measurement model. *Behavior genetics, 37*(4), 604–616. https://doi.org/10.1007/s10519-007-9156-1

Vogelsmeier, L. V., Vermunt, J. K., van Roekel, E., & De Roover, K. (2019). Latent Markov factor analysis for exploring measurement model changes in time-intensive longitudinal studies. *Structural Equation Modeling: A Multidisciplinary Journal, 26*(4), 557–575. https://doi.org/10.1080/10705511.2018.1554445

Weir, J. P. (2005). Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. *The Journal of Strength & Conditioning Research, 19*(1), 231–240. https://doi.org/10.1519/15184.1

Yu, D. S. F., Thompson, D. R., Yu, C. M., Pedersen, S. S., & Denollet, J. (2010). Validating the Type D personality construct in Chinese patients with coronary heart disease. *Journal of psychosomatic research, 69*(2), 111–118. https://doi.org/10.1016/j.jpsychores.2010.01.014

Zheng, X., Yang, J. S., & Harring, J. R. (2022). Latent Growth Modeling with Categorical Response Data: A Methodological Investigation of Model Parameterization, Estimation, and Missing Data. *Structural Equation Modeling: A Multidisciplinary Journal, 29*(2), 182–206. https://doi.org/10.1080/10705511.2021.1930543

Zigmond, A. S., & Snaith, R. P. (1983). The hospital anxiety and depression scale. *Acta psychiatrica scandinavica, 67*(6), 361–370. https://doi.org/10.1111/j.1600-0447.1983.tb09716.x

Zohar, A. H. (2016). Is type-D personality trait (s) or state? An examination of type-D temporal stability in older Israeli adults in the community. *PeerJ, 4*, Article e1690. https://doi.org/10.7717/peerj.1690